PROGRAMME AND ABSTRACTS

16th International Conference on

Computational and Financial Econometrics (CFE 2022)

http://www.cfenetwork.org/CFE2022

and

15th International Conference of the
ERCIM (European Research Consortium for Informatics and Mathematics) Working Group on

Computational and Methodological Statistics (CMStatistics 2022)

http://www.cmstatistics.org/CMStatistics2022

King's College London, UK

17 – 19 December 2022

ISBN 978-9925-7812-6-3

International Organizing Committee:
Ana Colubi, Erricos Kontoghiorghes and Michael Pitt.

CFE 2022 Co-chairs:
Serge Darolles, Carlos Lamarche, Indeewara Perera and Anna Simoni.

CFE 2022 Programme Committee:

Niklas Ahlgren, Alessandra Amendola, David Ardia, Josu Arteche, Monica Billio, Kris Boudt, Raffaella Calabrese, Vincenzo Candila, Massimiliano Caporin, Roberto Casarin, Leopoldo Catania, Joshua Chan, Julien Chevallier, Jonathan Crook, Gianluca Cubadda, Pierangelo De Pace, Eric Eisenstat, Christian Francq, Genevieve Gauthier, Matthew Harding, Alain Hecq, Douglas Hodgson, Benjamin Holcblat, Mohammad Jahan-Parvar, Menelaos Karanasos, Onno Kleen, Edward Knotek, Gary Koop, Nathan Lassance, Leone Leonida, Degui Li, Gael Martin, Claudio Morana, Kanchana Nadarajah, Ingmar Nolte, David Nott, Jose Olmo, Yasuhiro Omori, Michael Owyang, Alessia Paccagnini, Sandra Paterlini, Peter Pedroni, Bin Peng, Mikkel Plagborg-Moller, Tommaso Proietti, Christos Savva, Willi Semmler, Leopold Soegner, Robert Taylor, Gabriele Torri, Martin Wagner, Shixuan Wang, Tomasz Wozniak and Jean-Michel Zakoian.
.

CMStatistics 2022 Co-chairs:
Alexander Aue, John Kornak, Claudia Neves and David Rosell.

CMStatistics 2022 Programme Committee:

Toshihiro Abe, Andreas Alfons, Andreas Anastasiou, Eleni-Rosalina Andrinopoulou, Yuko Araki, Moulinath Banerjee, Andriette Bekker, Boris Beranger, Brenda Betancourt, Stefanie Biedermann, Pier Giovanni Bissiri, Natalia Bochkina, Graciela Boente, Enea Bongiorno, Michelle Carey, Victor Casero-Alonso, Shubhadeep Chakraborty, Radu Craiu, Maria Michela Dickson, Yuexiao Dong, Jochen Einbeck, Ani Eloyan, Sabrina Giordano, Stephane Girard, Jeff Goldsmith, Sonja Greven, Bettina Gruen, Nilabja Guha, Rajarshi Guhaniyogi, Montserrat Guillen, Michele Guindani, Zijian Guo, Christopher Hans, Piotr Jaworski, Sayar Karmakar, Maria Kateri, Tatyana Krivobokova, Andrew Lawson, Christophe Ley, Tianxi Li, Zeng Li, Zhaoyuan Li, Nicola Loperfido, Sara Lopez-Pintado, Saumen Mandal, Andreas Mayr, Bojana Milosevic, Michelle Miranda, Cristina Mollica, Erica Moodie, Alejandro Murua, Kalliopi Mylona, Jens Perch Nielsen, Bernardo Nipoti, Philipp Otto, Ioannis Papastathoupoulos, Arthur Pewsey, Yumou Qiu, Shahina Rahman, Zhao Ren, Aaron Scheffler, Shaun Seaman, Saunak Sen, Russell Shinohara, Sanjoy Sinha, Andrew Spieker, Baoluo Sun, Asaf Weinstein, Roy Welsch and Xavier de Luna.

Local Organizer:
King's Business School and King's Department of Mathematics.
CFEnetwork and CMStatistics.

Dear Friends and Colleagues,

We warmly welcome you to London for the 16th International Conference on Computational and Financial Econometrics (CFE 2022) and the 15th International Conference of the ERCIM Working Group on Computational and Methodological Statistics (CMStatistics 2022). After two years of mainly virtual meetings, we are delighted to have the opportunity to have a mostly in-person conference again. In light of the ongoing uncertainty caused by the pandemic, we have opted for the conference to have a hybrid format, so that the participants can select to participate in person or virtually according to their circumstances; however, this year, most of the sessions are planned to be fully in person or hybrid, with the majority of the participants onsite.

The conference aims to bring together researchers and practitioners to discuss recent developments in computational methods for economics, finance, and statistics. The CFE-CMStatistics 2022 programme consists of about 450 sessions, four plenary talks, and more than 1750 presentations. With around 1930 participants registered, this conference is once again the biggest meeting of the conference series in terms of the number of participants and presentations. The growth of the conference in terms of size and quality makes it undoubtedly one of the most important international scientific events in the field.

The co-chairs have endeavoured to provide a balanced and stimulating programme that will appeal to the diverse interests of the participants. The international organizing committee hopes that the hybrid conference will provide an ideal environment to communicate effectively with colleagues, in many cases, for the first time in months. The conference is the collective effort of many individuals and organizations. The Scientific Programme Committee, the Session Organizers, the supporting universities, and many agents have contributed substantially to the organization of the conference. We acknowledge their work and the support of our networks.

The Kings College London (KCL) provides excellent facilities and a fantastic environment in central London. Through their efforts the local host and sponsoring organizations have substantially contributed to the successful organization of the conference. We thank them all for their support. In particular we express our sincere appreciation to the hosts, the Department of Mathematics at KCL and the Data Analytics for Finance and Macro (DAFM) Research Centre at the Kings Business School.

The Elsevier journal Econometrics and Statistics (EcoSta), inaugurated in 2017, will have its first impact factor next year. The EcoSta is an official journal of the networks of Computational and Financial Econometrics (CFEnetwork) and of Computational and Methodological Statistics (CMStatistics). It publishes research papers in all aspects of econometrics and statistics, and it comprises two sections, namely, Part A: Econometrics and Part B: Statistics. The participants are encouraged to submit their papers to special or regular peer-reviewed issues of EcoSta and its supplement Annals of Computational and Financial Econometrics.

The CMStatistics has also commenced The Annals of Statistical Data Science (SDS), which will be published as a supplement to the Elsevier journal Computational Statistics & Data Analysis (CSDA). The CSDA is also the official journal of CMStatistics. You are encouraged to submit your papers to the Annals of Statistical Data Science or regular peer-reviewed issues of CSDA.

Looking ahead, the CFE-CMStatistics 2023 will be held at HTW Berlin - University of Applied Sciences, from Saturday the 16th of December 2023 to Monday the 18th of December 2023. Tutorials will take place on Friday, the 15th of December 2023. You are invited and encouraged to participate in these events actively.

We wish you a productive and stimulating conference.

Ana Colubi, Erricos J. Kontoghiorghes and Manfred Deistler

Coordinators of CMStatistics & CFEnetwork and EcoSta.

CMStatistics: ERCIM Working Group on COMPUTATIONAL AND METHODOLOGICAL STATISTICS

http://www.cmstatistics.org

The working group (WG) CMStatistics comprises a number of specialized teams in various research areas of computational and methodological statistics. The teams act autonomously within the framework of the WG in order to promote their own research agenda. Their activities are endorsed by the WG. They submit research proposals, organize sessions, tracks and tutorials during the annual WG meetings and edit journal special issues. The Econometrics and Statistics (EcoSta) and Computational Statistics & Data Analysis (CSDA) are the official journals of the CMStatistics.

Specialized teams

Currently, the ERCIM WG has over 1950 members and the following specialized teams

BIO: Biostatistics
BS: Bayesian Statistics
DMC: Dependence Models and Copulas
DOE: Design Of Experiments
FDA: Functional Data Analysis
HDS: High-Dimensional Statistics
IS: Imprecision in Statistics
LVSEM: Latent Variable and Structural Equation Models
MM: Mixture Models
NPS: Non-Parametric Statistics
RS: Robust Statistics
SA: Survival Analysis
SAE: Small Area Estimation
SDS: Statistical Data Science: Methods and Computations
SEA: Statistics of Extremes and Applications
SL: Statistical Learning
TSMC: Times Series

You are encouraged to become a member of the WG. For further information, please contact the Chairs of the specialized groups (see the WG's website) or email at info@cmstatistics.org.

CFEnetwork COMPUTATIONAL AND FINANCIAL ECONOMETRICS

http://www.CFEnetwork.org

The Computational and Financial Econometrics (CFEnetwork) comprises a number of specialized teams in various research areas of theoretical and applied econometrics, financial econometrics and computation, and empirical finance. The teams contribute to the network's activities by organizing sessions, tracks and tutorials during the annual CFEnetwork meetings, and by submitting research proposals. Furthermore, the teams edit special issues currently published under the Annals of CFE. The Econometrics and Statistics (EcoSta) is the official journal of the CFEnetwork. Currently, the CFEnetwork has over 1100 members.

You are encouraged to become a member of the CFEnetwork. For further information, please see the website or contact by email at info@cfenetwork.org.

SCHEDULE (GMT)

2022-12-17

A - Keynote
CFE - CMStatistics
08:30 - 09:30

Coffee Break
09:30 - 10:00

C
CFE - CMStatistics
10:00 - 12:05

Lunch Break
12:05 - 13:35

D
CFE - CMStatistics
13:35 - 15:40

Coffee Break
15:40 - 16:10

E
CFE - CMStatistics
16:10 - 17:50

F
CFE - CMStatistics
18:05 - 19:20

Welcome reception
19:30 - 21:00

2022-12-18

G
CFE - CMStatistics
08:15 - 09:55

Coffee Break
09:55 - 10:25

H
CFE - CMStatistics
10:25 - 12:05

Lunch Break
12:05 - 13:35

I
CFE - CMStatistics
13:35 - 15:15

Coffee Break
15:15 - 15:45

J
CFE - CMStatistics
15:45 - 17:00

K
CFE - CMStatistics
17:15 - 19:20

Christmas Conference Dinner
20:00 - 22:30

2022-12-19

L
CFE - CMStatistics
08:40 - 09:55

Coffee Break
09:55 - 10:25

M
CFE - CMStatistics
10:25 - 12:05

Lunch Break
12:05 - 13:35

N - Keynote
CFE - CMStatistics
13:35 - 14:25

P
CFE - CMStatistics
14:40 - 16:20

Coffee Break
16:20 - 16:50

Q
CFE - CMStatistics
16:50 - 18:30

R - Keynote
CFE - CMStatistics
18:45 - 19:40

TUTORIALS, MEETINGS AND CONFERENCE DETAILS (see maps)

TUTORIALS

Tutorials will take place on Friday the 16th of December 2022 , at the Council Room, Strand building. The first tutorial ("Extreme Value Analysis") will be delivered by Prof. Armelle Guillou, University of Strasbourg, France, 9:00-13:30 (GMT). The second tutorial ('Latent variable dynamic model") will be delivered by Prof. Michael Pitt, King's College London, UK, 15:00 to 19:30 (GMT). Only participants who had subscribed for the tutorial can attend, either in person or virtually through the website.

SPECIAL MEETINGS

The *Econometrics and Statistics (EcoSta) Editorial Board* and the *CSDA and Annals of Statistical Data Science Editorial Board* meetings will take place on Friday the 16th of December 2022, 17:45-18:30 (GMT). Indications to attend the Editorial Board meetings will be sent to the AEs attending the conference in due course.

CONFERENCE DETAILS

Access

- All the participants can attend virtually or in person. However, the in-person access to King's College London is restricted to those who had confirmed their in-person participation during the registration.
- The in-person venue is King's College London, Strand campus (Strand, London WC2R 2LS, United Kingdom).
- Indication to access the virtual part of the conference can be found on the webpage.
- The registration will be open on Friday afternoon, from 14:00 to 19:00, during the weekend from 7:30 to 18:00 and on Monday from 8:00 to 16:30 in the Arcade of the Bush House - Central Block (ground floor).

Scientific programme and social events

- The conference is live streaming, and it will not be recorded. The virtual oral presentations will take place through Zoom, while the social events and poster presentations will run in Gather Town.
- **Scientific programme:** The virtual and hybrid sessions are accessible from the interactive schedule. The conference programme time is set in GMT. Indications to access the in-person and virtual rooms can be found on the website. The in-person participants can use BH (S) 2.01, BH (SE) 2.01, S-2.08 and S-1.08 as quiet rooms and to participate in virtual sessions with their laptops and headphones.
- **Coffee breaks:** The coffee breaks will last 40 minutes each (beginning 10 minutes before the times indicated in the program). These will take place at the Great Hall of the King's Building (ground floor) and the Arcade of the Bush House - Central Block (ground floor). You must have your conference badge in order to attend the coffee breaks.
- **Lunches:** Sandwich box lunches have been organized for the three days of the conference. The lunches will take place at the Great Hall of the King's Building (Ground Floor). The lunches are optional, and registration is required. Participants must bring their conference badge to attend the lunches. Information about the purchased lunches is embedded in the QR code on the conference badge. During lunchtime each day, the conference participants are invited to interact in the conference virtual networking space. Indications to access the networking space can be found on the website.
- **Welcome reception:** The in-person welcome reception for registered participants will take place at the Great Hall of the King's building (Ground floor) and the Arcade of Bush House (Ground floor) on Saturday the 17th of December 2022 from 19:30 to 21:00 (GMT). Participants must bring their conference badge to attend the reception. Information about the welcome reception booking is embedded in the QR code on the conference badge. Simultaneously, a virtual welcome reception will take place in Gather Town. Indications to access the can be found on the website.
- **Christmas Conference Dinner:** The Christmas Conference Dinner will take place on Sunday the 18th of December 2022, at 20:00 at the Ambassadors Bloomsbury Hotel (12 Upper Woburn Pl, London WC1H 0HX). The conference dinner is optional, and registration is required. Participants must bring their conference badge to attend the conference dinner. Information about the purchased conference dinner ticket is embedded in the QR code on the conference badge.

Presentation instructions

The virtual presentations will take place through Zoom. Speakers should install the application, have a stable internet connection, and ensure their video and audio work. They will share their slides when the chair requires it, present their talk, and answer the question after the presentation. The in-person speakers must copy their presentations on the desktop on the conference room PCs and then share them on Zoom. The PCs have a touch screen with a webcam, a mobile support and an omnidirectional desk microphone that collects the sound around the PC desk to make the live streaming easy. Detailed indications for speakers in either virtual or hybrid sessions can be found on the website. As a general rule, each speaker has 20 minutes for the talk and 3-4 minutes for discussion. Strict timing must be observed.

Posters

The poster sessions will take place through Gather Town. The posters should be sent in **png format** to info@CMStatistics.org by the 14th of December. Landscape orientation is advisable. Detailed indications for the poster presentations can be found on the website.

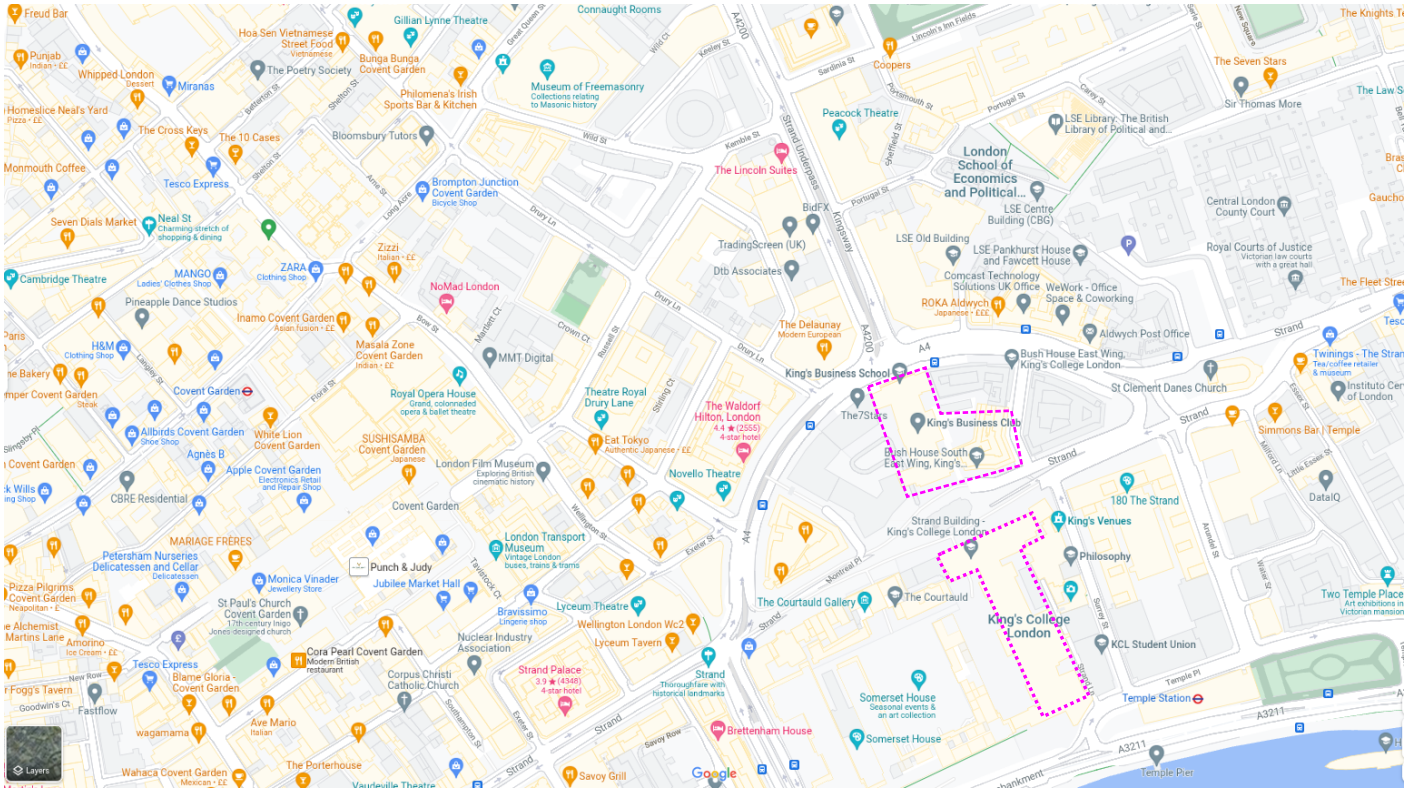
Session chairs

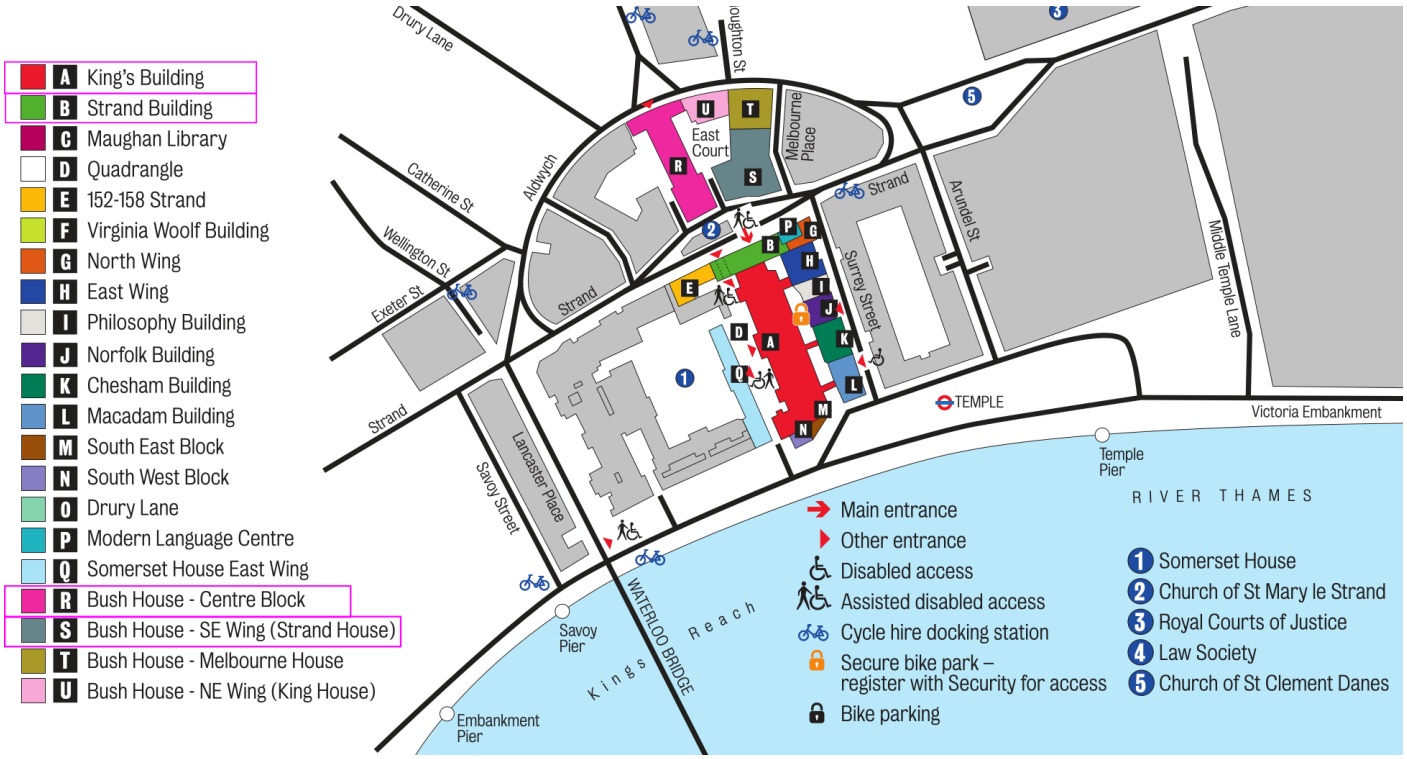
The session chairs will be responsible for introducing the session, the speakers and coordinating the discussion time. A member of the conference staff, identified on Zoom by the name Angel, will assist online. If any speaker is missing or has a technical problem, the chair can pass to the next speaker and come back later to resume if possible. Detailed indications for the session chairs of both virtual and hybrid sessions can be found on the website.

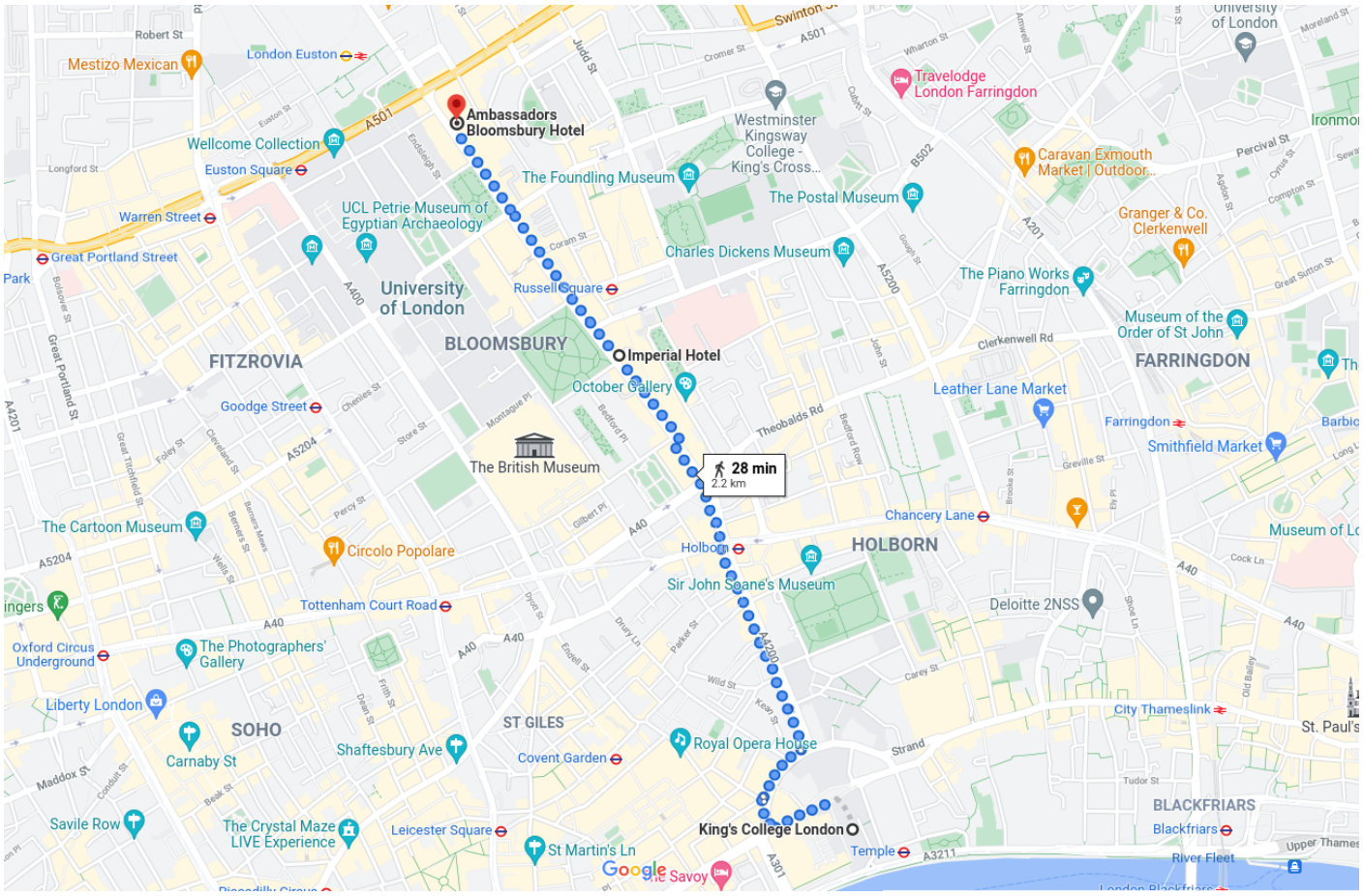
Test session

A test session will be set up for Saturday the 10th of December 2022 from 14:00 to 14:30 GMT. The participants will be able to enter virtually the hybrid room BH (S) 1.01 Lecture Theatre from the interactive programme to test their presentations, video, micro and audio (e.g., through the Slot C). Detailed indications for the test sessions can be found on the website.

Maps of the venue and nearby area

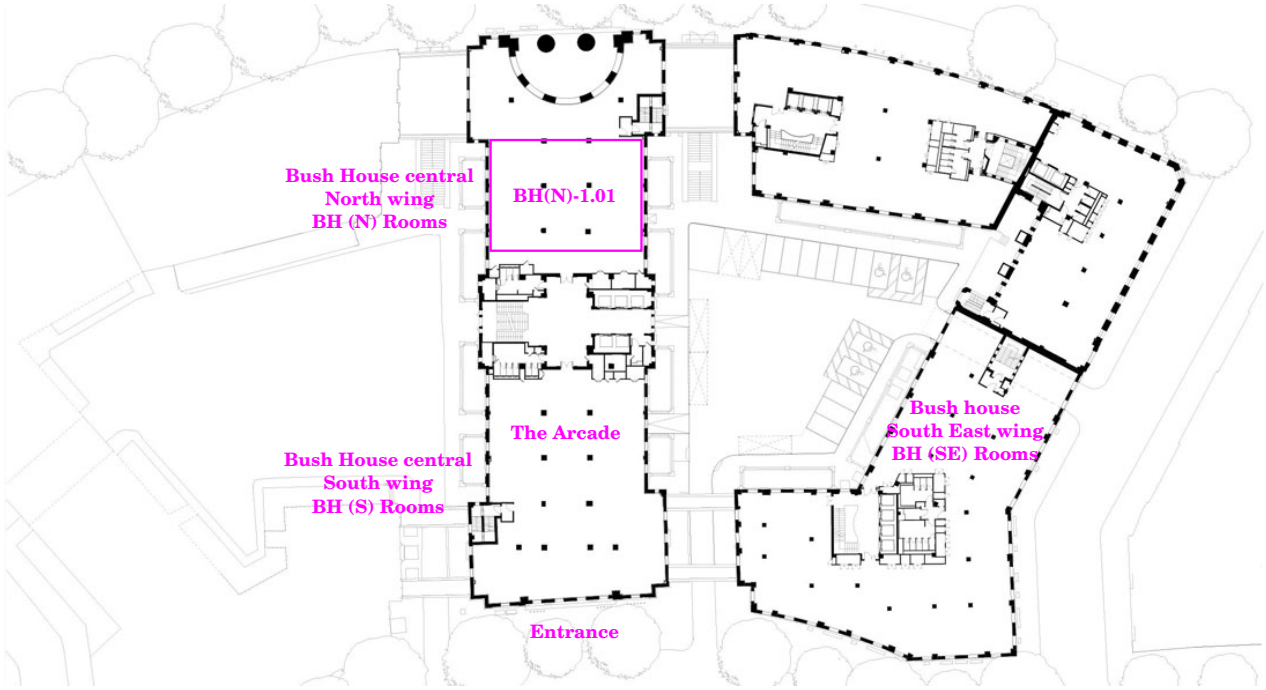




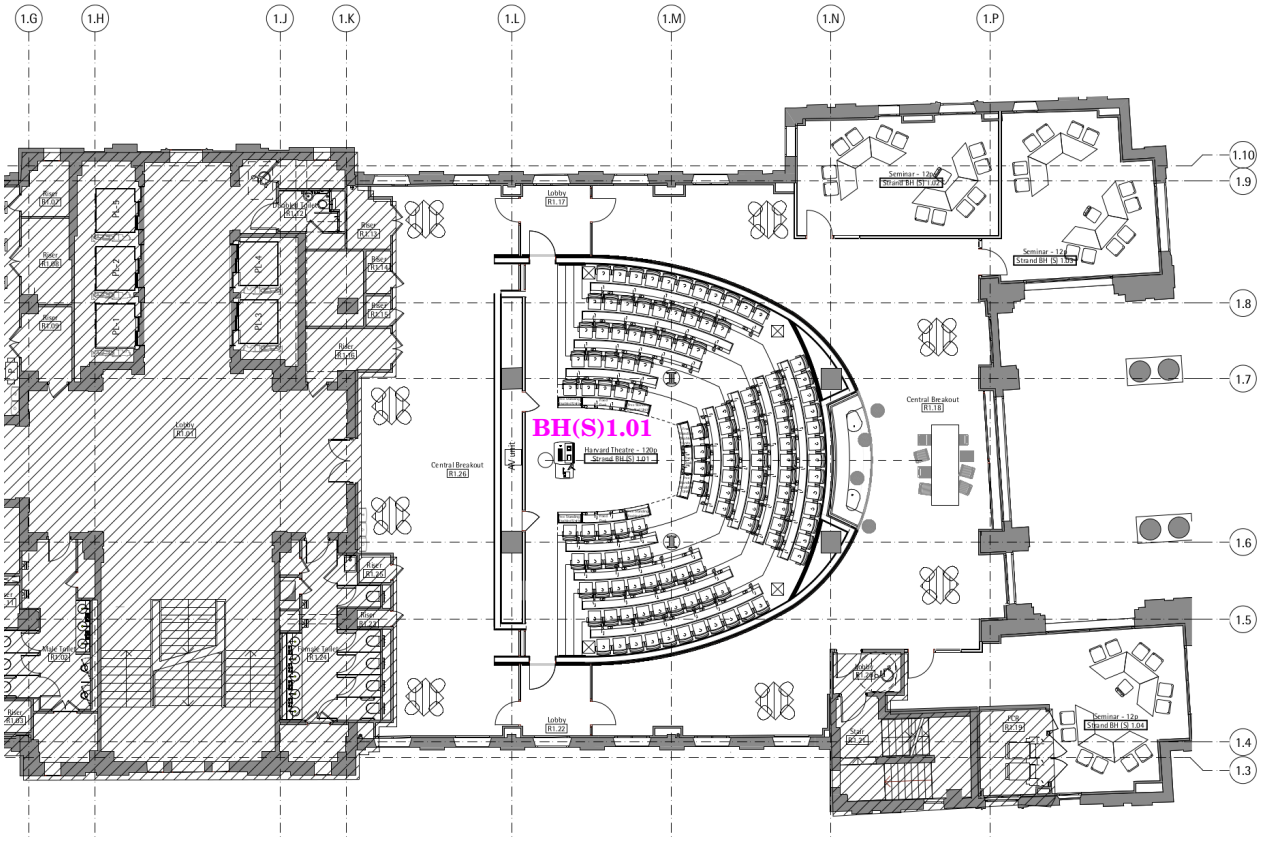


Floor maps

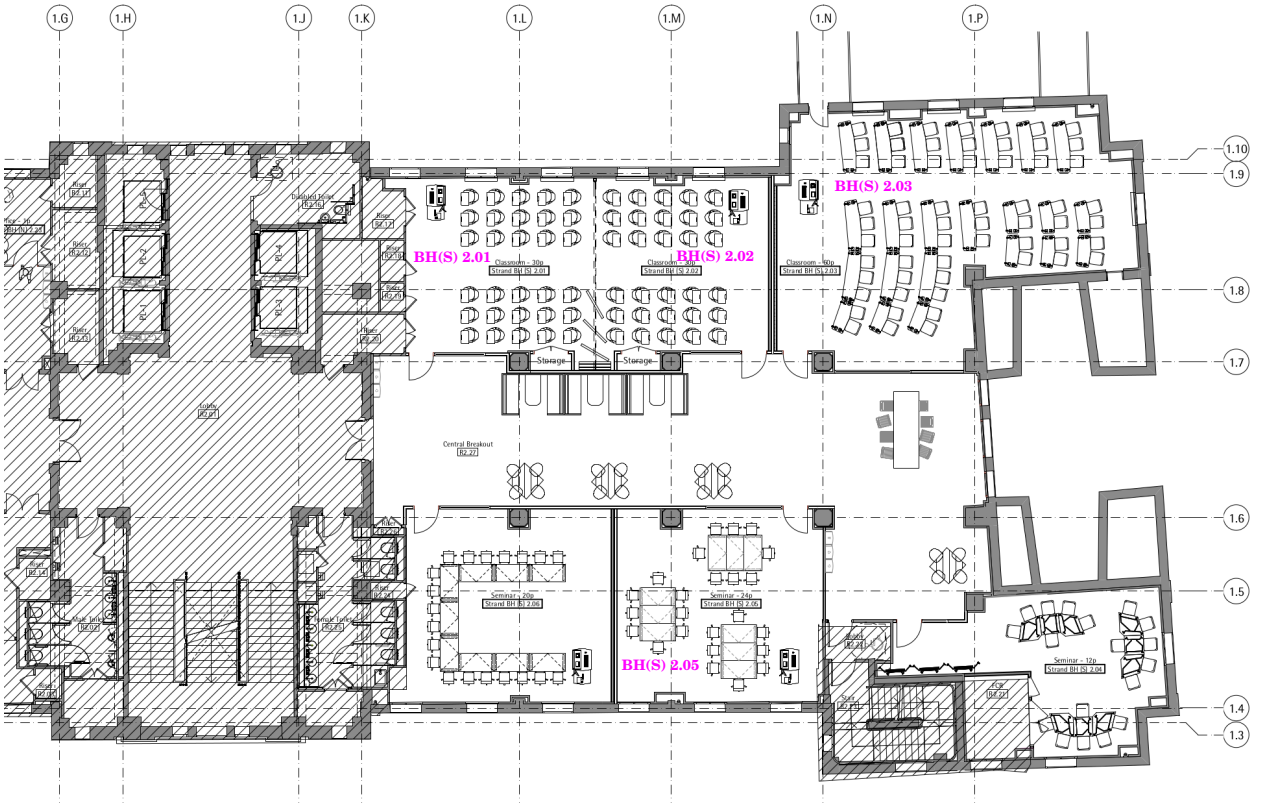
Bush House - Ground Floor



Bush House South Wing - First Floor



Bush House South Wing - Second Floor

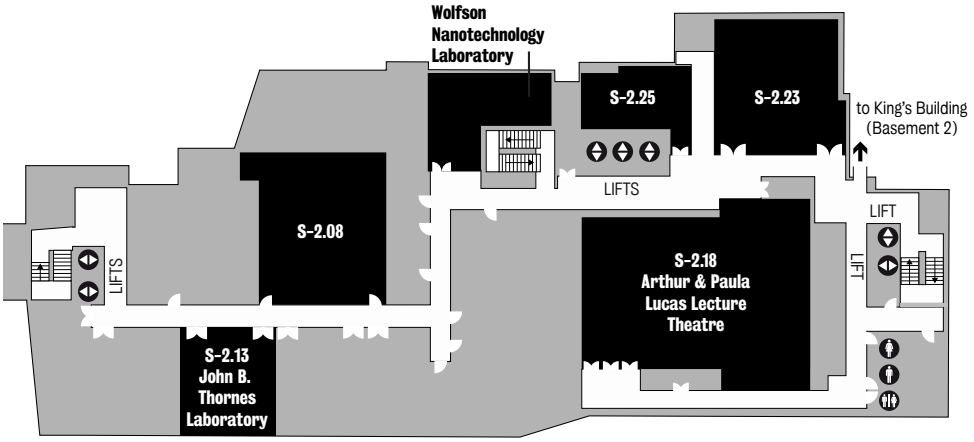


Bush House South East Wing - First Floor

Bush House South East Wing - Second Floor

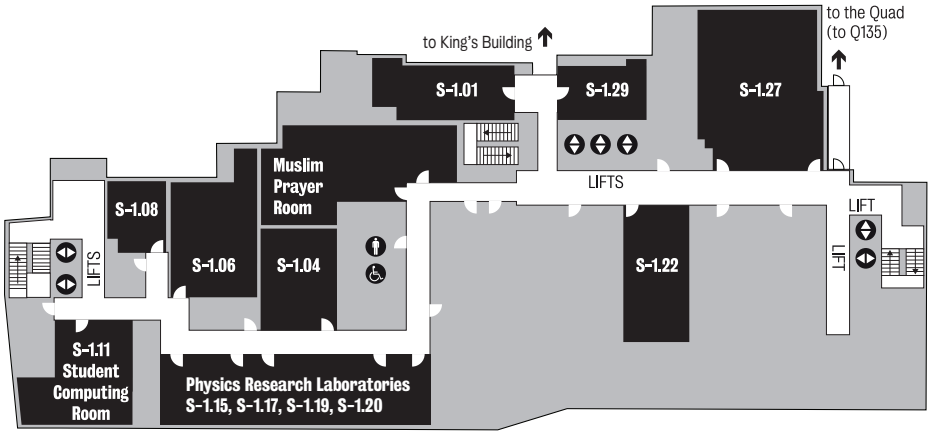
Strand Campus

Strand Building – Basement 2



Strand Campus

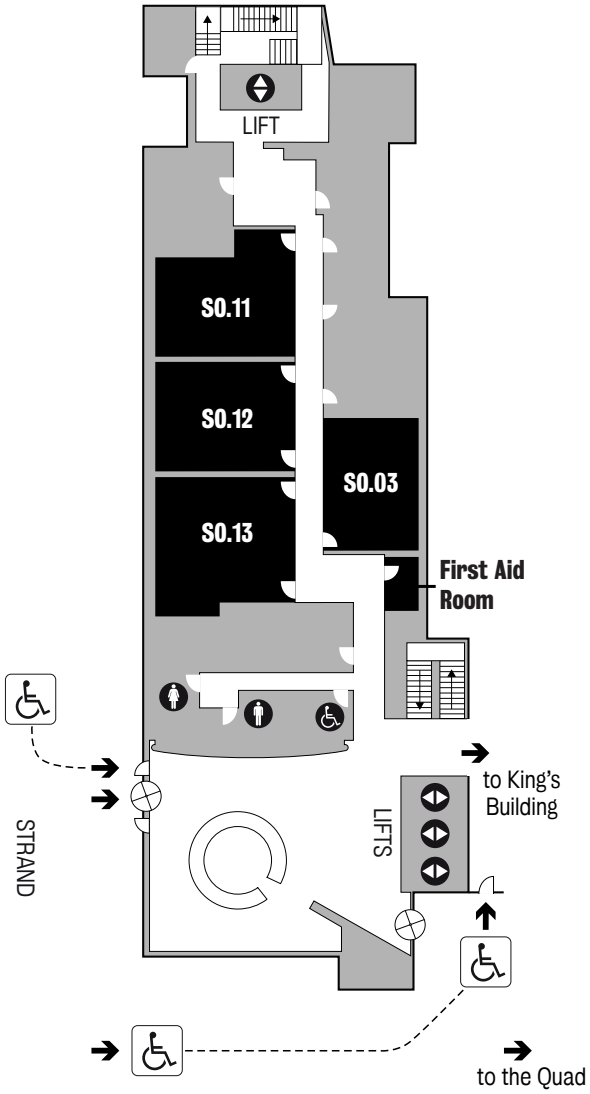
Strand Building – Basement 1



Strand Campus

Strand Building – Ground floor

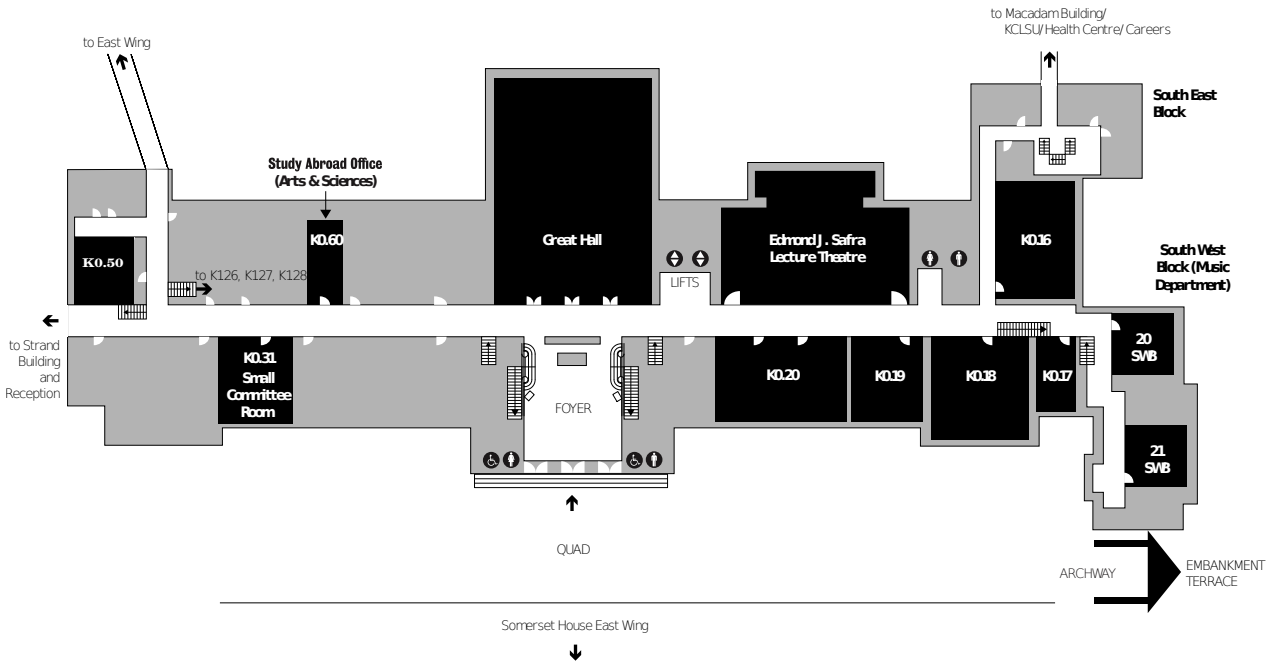
KING'S College LONDON



The non-stepped accessible route in to the building is to the rear of the main reception area. This route is via the black gated entrance and turn left.
There is also a button-controlled self-opening door at the front of the main reception but this requires reception staff to activate it.

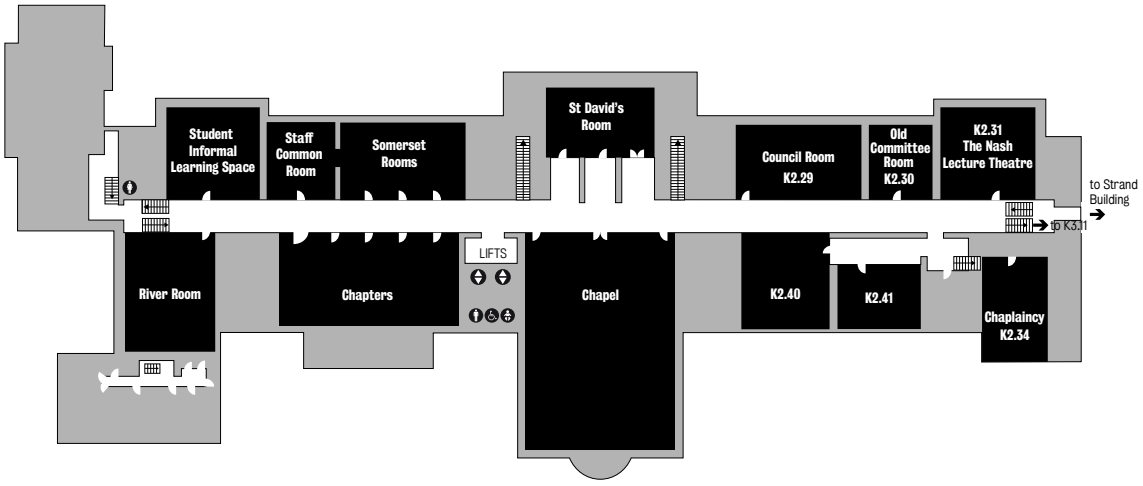
Strand Campus

King's Building – Level 0



Strand Campus

King's Building – Level 2



PUBLICATION OUTLETS

Econometrics and Statistics (EcoSta)

http://www.elsevier.com/locate/ecosta

Econometrics and Statistics (EcoSta), published by Elsevier, is the official journal of the networks Computational and Financial Econometrics and Computational and Methodological Statistics. It publishes research papers in all aspects of econometrics and statistics and comprises two sections:
Part A: Econometrics. Emphasis is given to methodological and theoretical papers containing substantial econometrics derivations or showing a potential of a significant impact in the broad area of econometrics. Topics of interest include the estimation of econometric models and associated inference, model selection, panel data, measurement error, Bayesian methods, and time series analyses. Simulations are considered when they involve an original methodology. Innovative papers in financial econometrics and its applications are considered. The covered topics include portfolio allocation, option pricing, quantitative risk management, systemic risk and market microstructure. Interest is focused as well on well-founded applied econometric studies that demonstrate the practicality of new procedures and models. Such studies should involve the rigorous application of statistical techniques, including estimation, inference and forecasting. Topics include volatility and risk, credit risk, pricing models, portfolio management, and emerging markets. Innovative contributions in empirical finance and financial data analysis that use advanced statistical methods are encouraged. The results of the submissions should be replicable. Applications consisting only of routine calculations are not of interest to the journal.
Part B: Statistics. Papers providing important original contributions to methodological statistics inspired in applications are considered for this section. Papers dealing, directly or indirectly, with computational and technical elements are particularly encouraged. These cover developments concerning issues of high-dimensionality, re-sampling, dependence, robustness, filtering, and, in general, the interaction of mathematical methods, numerical implementations and the extra burden of analysing large and/or complex datasets with such methods in different areas such as medicine, epidemiology, biology, psychology, climatology and communication. Innovative algorithmic developments are also of interest, as are the computer programs and the computational environments that implement them as a complement.
The journal consists, preponderantly, of original research. Occasionally, review and short papers from experts are published, which may be accompanied by discussions. Special issues and sections within important areas of research are occasionally published. The journal publishes as a supplement the Annals of Computational and Financial Econometrics.

Call For Papers Econometrics and Statistics (EcoSta)

http://www.elsevier.com/locate/ecosta

Papers presented at the conference and containing novel components in econometrics or statistics are encouraged to be submitted for publication in special peer-reviewed or regular issues of the Elsevier journal Econometrics and Statistics (EcoSta) and its supplement Annals of Computational and Financial Econometrics. Papers should be submitted using the EM Submission tool. In the EM please select as type of article the CFE conference, CMStatistics Conference or Annals of Computational and Financial Econometrics. Any questions may be directed via email to editor@econometricsandstatistics.org

Call For Papers CSDA Annals of Statistical Data Science (SDS)

http://www.elsevier.com/locate/csda

We are inviting submissions for the 1st issue of the CSDA Annals of Statistical Data Science. The Annals of Statistical Data Science is published as a supplement to the journal of Computational Statistics & Data Analysis. It will serve as an outlet for distinguished research papers using advanced computational and/or statistical methods for tackling challenging data analytic problems. The Annals will become a valuable resource for well-founded theoretical and applied data-driven research. Authors submitting a paper to CSDA may request that it be considered for inclusion in the Annals. Each issue will be assigned to several Guest Associate Editors who will be responsible, together with the CSDA Co-Editors, for the selection of papers.

Submissions for the Annals should contain a significant computational or statistical methodological component for data analytics. In particular, the Annals welcomes contributions at the interface of computing, statistics addressing problems involving large and/or complex data. Emphasis will be given to comprehensive and reproducible research, including data-driven methodology, algorithms and software. There is no deadline for submissions. Papers can be submitted at any time. When they have been received, they will enter the editorial system immediately. All submissions must contain original unpublished work not being considered for publication elsewhere. Please submit your paper electronically using the Elsevier Editorial System: http://ees.elsevier.com/csda (Choose Article Type: Research paper, and then Select "Section IV. Annals of Statistical Data Science").

Editors: Erricos Kontoghiorghes and Ana Colubi (CMStatistics)

Guest Associate Editors: Julyan Arbel, Peter Buhlmann, Stefano Castruccio, Bertrand Clarke, Christophe Croux, Maria Brigida Ferraro, Yulia Gel Michele Guindani, Xuming He, Sangwook Kang, Ivan Kojadinovic, Chenlei Leng, Taps Maiti, Geoffrey McLachlan, Hans-Georg Mueller, Igor Pruenster, Juan Romo, Elvezio Ronchetti, Anne Ruiz-Gazen, Sylvain Sardi, Xinyuan Song, Cheng Yong Tang, Roy Welsch and Peter Winker.

Contents

Saturday 17.12.2022 08:30 - 09:30 Room: BH (N) -1.01 Chair: Tommaso Proietti Keynote talk 1

Dynamic models using score copula innovations

Speaker: **Michael Pitt, Kings College London, United Kingdom**

A new class of observation-driven dynamic models is introduced. The time-evolving parameters are driven by innovations of copula form. The resulting models can be made strictly stationary, and the innovation term is typically chosen to be Gaussian. The innovations are formed by applying a copula approach for the conditional score function, which has close connections to the existing literature on GAS models. This new method provides a unified framework for observation-driven models allowing the likelihood to be explicitly computed using the prediction decomposition. The approach may be used for multiple lag structures and for multivariate models. Strict stationarity can be easily imposed upon the models making the invariant properties simple to ascertain. This property also has advantages for specifying the initial conditions needed for maximum likelihood estimation. One-step and multi-period forecasting is straightforward, and the forecasting density is either in closed form or a simple mixture over a univariate component. The approach is very general, and the illustrations focus on volatility models and duration models. We illustrate the performance of the modelling approach for both univariate and multivariate volatility models.

Saturday 17.12.2022 08:30 - 09:30 Room: Safra Lecture Theatre Chair: Kalliopi Mylona Keynote talk 2

New graphical displays for classification

Speaker: **Peter Rousseeuw, KU Leuven, Belgium** Jakob Raymaekers, Mia Hubert

Classification is a major tool of statistics and machine learning. Several classifiers have interesting visualizations of their inner workings. We pursue a different goal, which is to visualize the cases being classified, either in training data or in test data. An important aspect is whether a case has been classified to its given class (label) or whether the classifier wants to assign it to a different class. This is reflected in the probability of the alternative class (PAC). A high PAC indicates label bias, i.e. the possibility that the case was mislabeled. The PAC is used to construct a silhouette plot which is similar in spirit to the silhouette plot for cluster analysis. The average silhouette width can be used to compare different classifications of the same dataset. We will also draw quasi-residual plots of the PAC versus a data feature, which may lead to more insight into the data. One of these data features is how far each case lies from its given class, yielding so-called class maps. The proposed displays are constructed for discriminant analysis, k-nearest neighbors, support vector machines, CART, random forests, and neural networks. The graphical displays are illustrated and interpreted on data sets containing images, mixed features, and texts.

Monday 19.12.2022 13:35 - 14:25 Room: Safra Lecture Theatre Chair: Benjamin Holcblat Keynote talk 4

The role of energy in UK inflation and productivity

Speaker: **Jennifer L Castle, Oxford University, United Kingdom**

The recent rise in UK price inflation was unanticipated, leading to a flurry of activity rethinking inflation models. We model UK price and wage inflation, productivity and unemployment over a century and a half of data, selecting dynamics, relevant variables, non-linearities and location and trend shifts using indicator saturation estimation. The four congruent econometric equations highlight complex interacting empirical relations. The production function reveals a major role for energy inputs additional to capital and labour, and although the price inflation equation shows a small direct impact of energy prices, the substantial rise in oil and gas prices seen by mid-2022 contributes half of the increase in price inflation. We find empirical evidence for non-linear adjustments of real wages to inflation: a wage-price spiral kicks in when inflation exceeds about 68% p.a. We also find an additional non-linear reaction to unemployment, consistent with involuntary unemployment. A reduction in energy availability simultaneously reduces output and exacerbates inflation.

Monday 19.12.2022 18:45 - 19:40 Room: Safra Lecture Theatre Chair: Juan Romo Keynote talk 3

Conditional tail moment and reinsurance premium estimation under random right censoring

Speaker: **Armelle Guillou , Strasbourg university, France** Yuri Goegebeur, Jing Qin

After introducing extreme value theory, in particular, in the censorship framework, the estimation of the conditional tail moment (CTM) will be discussed when the data are subject to random censorship. The variable of main interest and the censoring variable both follow a Pareto-type distribution. We establish the asymptotic properties of our estimator and discuss bias reduction. Then, the CTM is used to estimate, in case of censorship, the premium principle for excess-of-loss reinsurance. The finite sample properties of the proposed estimators are investigated with a simulation study, and we illustrate their practical applicability on a dataset of motor third-party liability insurance.

Saturday 17.12.2022 10:00 - 12:05 Parallel Session C – CFE-CMStatistics

EO655 Room S-2.23 RECENT ADVANCES IN TREE ENSEMBLE METHODS Chair: Roman Hornung

E0199: **Random forests: Why they work and why that is a problem**
Presenter: **Lucas Mentch**, University of Pittsburgh, United States

Random forests remain among the most popular off-the-shelf supervised machine learning tools with a well-established track record of predictive accuracy in both regression and classification settings. Despite their empirical success, a full and satisfying explanation for their success has yet to be put forth. We will show that the additional randomness injected into individual trees serves as a form of implicit regularization, making random forests an ideal model in low signal-to-noise ratio (SNR) settings. From a model-complexity perspective, this means that the mtry parameter in random forests serves much the same purpose as the shrinkage penalty in explicit regularization procedures like the lasso. Realizing this, we demonstrate that alternative forms of randomness can provide similarly beneficial stabilization. In particular, we show that augmenting the feature space with additional features consisting of only random noise can substantially improve the predictive accuracy of the model. This surprising fact has been largely overlooked within the statistics community, but has crucial implications for thinking about how best to define and measure variable importance. Numerous demonstrations on both real and synthetic data are provided.

E0230: **Sequential permutation testing of random forest variable importance measures**
Presenter: **Alexander Hapfelmeier**, Technical University of Munich, Germany
Co-authors: Roman Hornung, Bernhard Haller

Hypothesis testing of random forest (RF) variable importance measures (VIMP) remains the subject of ongoing research. Among recent developments, heuristic approaches to parametric testing have been proposed whose distributional assumptions are based on empirical evidence. Other formal tests under regularity conditions were derived analytically. However, these approaches can be computationally expensive or even infeasible. This problem also occurs with non-parametric permutation tests, which are, however, distribution-free and can generically be applied to any type of RF and VIMP. Embracing this advantage, it is proposed here to use sequential permutation tests and sequential p-value estimation to reduce the high computational costs of conventional permutation tests. The popular and widely used permutation accuracy VIMP serves as a practical and relevant application example. The results of simulation studies confirm that the theoretical properties of the sequential tests apply; that is, the type-I error probability is controlled at a nominal level, and high power is maintained with considerably fewer permutations needed. The numerical stability of the methods is investigated in two additional application studies. Recommendations for application are given. A respective implementation is provided through the accompanying R package rfvimptest. The approach can easily be applied to any kind of prediction model.

E0783: **Estimating heterogeneous treatment effects with right-censored data via causal survival forests**
Presenter: **Yifan Cui**, Zhejiang University, China
Co-authors: Michael Kosorok, Erik Sverdrup, Stefan Wager, Ruoqing Zhu

Forest-based methods have recently gained popularity for non-parametric treatment effect estimation. We introduce causal survival forests, which can be used to estimate heterogeneous treatment effects in survival and observational settings where outcomes may be right-censored. Our approach relies on orthogonal estimating equations to robustly adjust for both censoring and selection effects under unconfoundedness. In our experiments, we find our approach to perform well relative to a number of baselines.

E1040: **Variable importance for random forests: MDA and Shapley effects**
Presenter: **Clement Benard**, Safran Tech, France

Variable importance measures are the main tools to analyze the black-box mechanisms of random forests. Although the mean decrease accuracy (MDA) is widely accepted as the most efficient variable importance measure for random forests, little is known about its statistical properties. In fact, the exact MDA definition varies across the main random forest software. The objective is to rigorously analyze the behavior of the main MDA implementations. Consequently, we establish their limits when the sample size increases. In particular, we break down these limits into three components: the first two terms are related to Sobol indices, which are well-defined measures of a covariate contribution to the response variance, as opposed to the third term, whose value increases with dependence within covariates. Thus, we theoretically demonstrate that the MDA does not target the right quantity when covariates are dependent, a fact that has already been noticed experimentally. To address this issue, we define new important measures for random forests: the Sobol-MDA and SHAFF. The Sobol-MDA fixes the flaws of the original MDA, and is appropriate for variable selection. On the other hand, SHAFF is a fast and accurate estimate of Shapley's effects, even when input variables are dependent. SHAFF is appropriate to rank all variables for interpretation purposes. We prove the consistency of both the Sobol-MDA and SHAFF, and show that they empirically outperform their competitors.

E1028: **Interpolation and random forests**
Presenter: **Ludovic Arnould**, Sorbonne Universite, France
Co-authors: Erwan Scornet, Claire Boyer

Statistical wisdom suggests that very complex models, interpolating training data, will be poor at predicting unseen examples. Yet, this aphorism has been recently challenged by the identification of benign overfitting regimes, specially studied in the case of parametric models: generalization capabilities may be preserved despite model high complexity. While it is widely known that fully-grown decision trees interpolate and, in turn, have bad predictive performances, the same behavior is yet to be analyzed for random forests. We study the trade-off between interpolation and consistency for several types of random forest algorithms. Theoretically, we prove that interpolation regimes and consistency cannot be achieved simultaneously for non-adaptive random forests. Since adaptivity seems to be the cornerstone to bring together interpolation and consistency, we study interpolating Median RF, which is proven to be consistent in a noiseless scenario. Numerical experiments show that Breiman random forests are consistent while exactly interpolating, when no bootstrap step is involved. We theoretically control the size of the interpolation area, which converges fast enough to zero, so that exact interpolation and consistency occur in conjunction.

EO585 Room S-1.04 ADVANCES IN MULTIVARIATE STATISTICS Chair: Cinzia Franceschini

E0911: **How to measure changing patterns of taste and food consumption in transitional times**
Presenter: **Maria Giovanna Onorati**, Pollenzo University of Gastronomic Sciences, Italy
Co-authors: Cinzia Franceschini, Francesco Domenico dOvidio

Taste is a compound sociological category that draws on both routinized practices and values and aspirational motives that people associate with their consumption choices. How can we explain consumers' "doing" and "saying" in transitional times, when routines are disrupted, and human agency becomes crucial in reshaping them? How can this be done in a way that neither neglects consumer agency nor leads to individualistic explanations informed by social desirability? Categorical principal component analysis (CATPCA), a nonparametric method based on alternative least squares optimal scaling (ALSOS) supported by stepwise removal, is useful in linking practices and values including individual personal characteristics. Alternatively, we can use a tandem combination of factor analysis and cluster analysis, to uncover hidden patterns of values and actions and their distribution in the population under study.

E1239: **Evaluating the individual variability across consumers in texture perception through different classification approaches**
Presenter: **Maria Piochi**, University of Gastronomic Sciences, Italy

Co-authors: Cinzia Franceschini, Luisa Torri

In sensory and consumer science, there is often the need to classify/segment consumers according to different criteria to explain how consumers behave towards food and beverages. Subjects differ in their oral responsiveness, and this could affect their food preferences. Clustering approaches were used in a sample of 151 consumers to study how assessors perceived and liked different food recipes prepared with the same ingredient (carrot) processed to obtain different textures (solid, creamy, crispy). Projection pursuit, linear discriminant analysis and k-means cluster analysis were adopted to identify groups of subjects with different oral behavior in texture perception.

E1165: **An honeypot-semantic approach of data acquisition and data mining for corporate compliance analysis**

Presenter: **Vito Santarcangelo**, iInformatica Srl, Italy

Co-authors: Massimiliano Giacalone, Diego Carmine Sinito

A new interesting approach is shown for corporate distributed data acquisition thanks to the use of an innovative approach based on smart honeypot painting installed in corporate environments, smart honeypot plants installed in outdoor or manufacturing environments, with the support of an application survey. All the data acquired are certified thanks to the use of a blockchain. The concept of a honeypot is the best way to install the device for parameters monitoring in corporate environments without altering the characteristics of the work environment following the empathic design approach. Honeypots are designed in full compliance with confidentiality and GDPR compliance. Data acquired from honeypots and surveys are then processed by a data mining approach considering a specific semantic knowledge base related to the corporate context. This multivariate analysis is very useful for mapping the process gap of compliance, considering significant relations between variables. By following this approach, it is then possible to discover the KPI set of variables that have to be monitored for corporate compliance improvement. Some solutions are proposed for acquisition that can be useful to monitor the compliance of organizational models, and relative examples of multivariate analysis conducted by the use of a platform based on Weka environment.

E1288: **Advances on permutation tests for non-monotonic alternatives**

Presenter: **Massimiliano Giacalone**, University of Campania Luigi Vanvitelli, Italy

Co-authors: Stefano Bonnini, Michela Borghesi

Sample surveys and related inferential statistical methods are suitable tools to support decision-making practices for policymakers. For instance, to achieve the environmental sustainability goals connected to Circular Economy and product and process innovation, Governments and Parliaments should evaluate and stimulate the propensity of companies towards such virtuous behaviors. This practice becomes fundamental in a country like Italy, with an economic system mainly based on small and medium enterprises. The focus is on complex tests of hypotheses for investigating the effects of important factors, such as company size, on the propensity of enterprises towards a Circular Economy. Frequently, the complexity is given by the V-shaped or U-shaped effects of ordinal factors in multisample problems. In other words, in the tested hypothesis, the propensity towards Circular Economy is decreasing up to a pick point and then increasing with respect to the ordinal factor under study. The proposed solution is based on a nonparametric method that consists in a combined permutation test. The good performance of the method and its usefulness in real data problems are proved through a Monte Carlo simulation study and the application to data concerning a survey on the Circular Economy, respectively.

EO264 Room S-1.06 ADVANCES IN MIXTURE MODELLING AND MODEL-BASED CLUSTERING Chair: Bettina Gruen

E0513: **Model-based clustering of multinomial count data under the presence of covariates**

Presenter: **Panagiotis Papastamoulis**, Athens University of Economics and Business, Greece

The problem of inferring an unknown number of clusters in multinomial count data is considered by estimating finite mixtures of multinomial distributions with or without covariates. Both Maximum Likelihood (ML), as well as Bayesian estimation, are taken into account. Under a Maximum Likelihood approach, we provide an Expectation-Maximization (EM) algorithm which exploits a careful initialization procedure combined with a ridge-stabilized implementation of the Newton-Raphson method in the M-step. Under a Bayesian setup, a stochastic gradient Markov chain Monte Carlo (MCMC) algorithm embedded within a prior parallel tempering scheme is devised. The number of clusters is selected according to the Integrated Completed Likelihood criterion in the ML approach and estimating the number of non-empty components in overfitting mixture models in the Bayesian case.

E0630: **Unveiling patterns in spectroscopy data via a Bayesian latent variables approach**

Presenter: **Alessandro Casa**, Free University of Bozen-Bolzano, Italy

Co-authors: Tom O Callaghan, Thomas Brendan Murphy

Infrared spectroscopy techniques represent a convenient and non-disruptive way to collect vast amounts of data rapidly. Nowadays, these data are effectively used in a plethora of different fields, such as medicine, astronomy and food science. Nonetheless, from a statistical viewpoint, they introduce some relevant challenges mainly concerning their high dimensionality and the complex relationships among spectral variables (wavelengths), often due to convoluted chemical processes. In this framework, factor analysis represents a sensible strategy, as it aims to produce parsimonious representations of the data while focusing on the correlation structures. Nonetheless, its standard application does not account for redundancies in the features. Therefore, a modification of factor analysis is proposed, which maps the data into a lower dimensional latent space while simultaneously clustering the variables. A flexible Bayesian estimation procedure is then considered to fit the model. On the one hand, this approach results in an even more parsimonious summary of the data, highlighting which wavelengths carry similar information. On the other hand, from an interpretative point of view, the obtained partition produces useful insights from a chemical standpoint. The method is applied to milk mid-infrared spectroscopy data from cows on different feeding regimens, providing a useful tool to guarantee milk authenticity.

E0697: **Reliable variance matrix priors for Bayesian mixture models with Gaussian kernels**

Presenter: **Michail Papathomas**, University of St Andrews, United Kingdom

Co-authors: Wei Jing, Silvia Liverani

Bayesian mixture modelling is an increasingly popular approach for clustering and density estimation. We study the choice of prior for the variance or precision matrix when Gaussian kernels are adopted. Typically, mixture models are assessed by considering observations in the space of only a handful of dimensions. Instead, we are concerned with higher dimensionality problems, in the space of up to 20 dimensions, observing that the choice of prior becomes increasingly important as the dimensionality increases. After identifying certain undesirable properties of standard priors, we review and implement possible alternative priors. The most promising priors are identified, as well as factors that affect MCMC convergence. Results, using simulated and real data, show that the choice of prior and its implementation are critical for deriving reliable inferences. The focus is on the Dirichlet Process Mixture Model, but we also discuss its relevance to Bayesian Mixtures of Finite Mixture models.

E1546: **Accelerating Bayesian estimation for network Poisson models using frequentist variational estimates**

Presenter: **Stephane Robin**, Sorbonne universita, France

Co-authors: Sophie Donnet

The analysis of ecological interaction networks is the motivation. Poisson stochastic block models are widely used in this field to decipher the structure that underlies a weighted network, while accounting for covariate effects. Efficient algorithms based on variational approximations exist for frequentist inference, but without statistical guarantees as for the resulting estimates. In the absence of variational Bayes estimates, we show that a good proxy of the posterior distribution can be straightforwardly derived from the frequentist variational estimation procedure, using a

Laplace approximation. We use this proxy to sample from the true posterior distribution via a sequential Monte Carlo algorithm. As shown in the simulation study, the efficiency of the posterior sampling is greatly improved by the accuracy of the approximate posterior distribution. The proposed procedure can be easily extended to other latent variable models. We use this methodology to assess the influence of available covariates on the organization of several ecological networks, as well as the existence of a residual interaction structure.

E1981: **Clusterwise multivariate regression of mixed-type panel data**
Presenter: **Jan Vavra**, Charles University, Czech Republic
Co-authors: Arnost Komarek, Bettina Gruen, Gertraud Malsiner-Walli

Multivariate panel data of mixed type are routinely collected in many different areas of application, often jointly with additional covariates, which complicate the statistical analysis. Moreover, it is often of interest to identify unknown groups of units in a study population using such a data structure, i.e., to perform clustering. In the Bayesian framework, we propose a finite mixture of multivariate generalised linear mixed effects regression models to cluster numeric, binary, ordinal and categorical panel outcomes jointly. The specification of suitable priors on the model parameters allows for convenient posterior inference based on Markov chain Monte Carlo (MCMC) sampling with data augmentation. The Bayesian approach allows us to obtain both a classification of the subjects in the data and new subjects as well as cluster-specific parameter estimates. Finally, model estimation and selection of the number of data clusters are simultaneously performed when approximating the posterior for a single model using MCMC sampling without resorting to multiple model estimations. Its application is illustrated in a data set from the Czech part of the EU-SILC survey, where households are annually interviewed to obtain insights into changes in their financial capability.

EO622 Room S-1.27 ADVANCES IN LONGITUDINAL DATA MODELING Chair: Maria Francesca Marino

E1102: **Measurement invariance and latent Markov models: A model selection problem**
Presenter: **Francesco Dotto**, University of Roma Tre, Italy
Co-authors: Roberto Di Mari, Alessio Farcomeni, Antonio Punzo

A general approach is proposed to detect measurement non-invariance in latent Markov models for longitudinal data. We define different notions of differential item functioning in the context of panel data. We then present a model selection approach based on the Bayesian information criterion (BIC) to choose both the number of latent states and the measurement structure. We show the practical relevance by means of an extensive simulation study, and illustrate its use on two real-data examples from the social sciences. Our results indicate that BIC is able to select the correct measurement equivalence structure more than 95% of the time.

E0558: **Quantile mixed hidden Markov models for multivariate longitudinal data: An application to children's SDQ scores**
Presenter: **Luca Merlo**, European University of Rome, Italy
Co-authors: Lea Petrella, Nikos Tzavidis

The identification of factors associated with mental and behavioural disorders in early childhood is critical both for psychopathology research and the support of primary health care practices. Motivated by the Millennium Cohort Study, we study the effect of a comprehensive set of covariates on children's emotional and behavioural trajectories in England. To this end, we develop a quantile mixed hidden Markov model for joint estimation of multiple quantiles in a linear regression setting for multivariate longitudinal data. The novelty of the proposed approach is based on the multivariate asymmetric Laplace distribution, which allows us to jointly estimate the quantiles of the univariate conditional distributions of a multivariate response, accounting for possible correlation between the outcomes. Sources of unobserved heterogeneity and serial dependency due to repeated measures are modelled through the introduction of individual-specific, time-constant random coefficients and time-varying parameters evolving over time with a Markovian structure, respectively. The inferential approach is carried out through the construction of a suitable expectation-maximization algorithm without parametric assumptions on the random effects distribution.

E1240: **Including attributes in dynamic stochastic blockmodels: An application to international trade data**
Presenter: **Silvia Pandolfi**, University of Perugia, Italy
Co-authors: Francesco Bartolucci, Maria Francesca Marino

Dynamic Stochastic Block Models represent a flexible tool for analysis in the presence of longitudinal network data. These models provide a dynamic clustering of network nodes by exploiting a latent variable approach based on a hidden Markov specification. The aim is to simplify the complex data structure typical of networks and extract the relevant information from the data. Motivated by the analysis of network data entailing trade flows among countries recorded in the period 2012-2019, the dyad-independent model is extended by allowing the inclusion of attributes on either the measurement and/or the latent model. Together with information on import-export exchanges, a number of nodal (e.g., the total population and the GDP of the country) and edge (e.g., the existence of a commercial agreement between countries) attributes are indeed available. Including this information in the model allows us a deeper understanding of the determinants of international trade. Due to the intractability of the likelihood function, the estimation of model parameters is performed by relying on a variational approximation, as frequently done in the SBMs framework.

E1331: **Markov-switching conditional logistic regression with application to animal movement data of interacting individuals**
Presenter: **Jennifer Pohle**, University of Potsdam, Germany
Co-authors: Johannes Signer, Ulrike Schlaegel

Integrated step selection analysis is a popular statistical tool in ecology to study animal's movement and habitat selection based on conditional logistic regression. It has also successfully been applied to detect interactions (such as avoidance or attraction) between simultaneously tracked individuals. However, animals usually switch between different behavioural modes (such as resting and foraging), which might influence their preferences, habitat selection, and movement patterns. Ignoring such behavioural states in the analyses might lead to biased results and possible erroneous conclusions. To account for the usually unobserved behaviour, we present an approach where an underlying latent Markov chain is introduced into the framework, which allows preferences and movement patterns to vary over time. A simulation study is used to investigate the performance of the resulting Markov-switching integrated step selection analysis and to compare it to alternative candidate models. Furthermore, the approach is illustrated in a case study on location data from simultaneously tracked bank voles. Besides animal movement data, the inherent Markov-switching conditional logistic regression is also applicable to longitudinal discrete choice and case-control studies.

E1166: **On path-specific natural effects with a binary outcome and two binary mediators**
Presenter: **Marco Doretti**, University of Perugia, Italy
Co-authors: Elena Stanghellini, Paolo Berta, Martina Raggi

Causal mediation analysis with multiple ordered mediators has generated a growing interest in the last few years. From a longitudinal perspective, it also applies to settings where a single mediator is repeatedly measured over time. In the presence of two mediators, univocal definitions of path-specific natural effects have been recently introduced. Under the usual assumptions about consistency and absence of unmeasured confounding, in linear systems, many of these effects can be identified up to a sensitivity parameter, that is, the conditional density of a certain potential outcome given another one. We aim to extend this framework to the case where the outcome and all mediators are binary and modeled via logistic regression. Our purpose is to achieve parametric identification of generic counterfactual probabilities, so that casual contrasts of any nature (and on any scale) can be recovered from the regression coefficients. To this end, we adapt to our context the sensitivity analysis approach performed in the linear case and introduce an alternative solution, which is based on the continuous latent trait usually assumed to underlie binary variables. Our proposal

is illustrated via an application to administrative data from birth assistance certificates, collected in Lombardy (Italy) in the years 2010-2012.

EO624 Room K0.19 ADVANCED METHODS FOR BAYESIAN MODELING Chair: Florian Frommlet

E0651: **Bayesian modeling and clustering for spatio-temporal areal data**
Presenter: **Alexander Mozdzen**, University of Klagenfurt, Austria
Co-authors: Gregor Kastner, Annalisa Cadonna, Andrea Cremaschi, Alessandra Guglielmi

Spatio-temporal areal data can be seen as a collection of time series which are spatially correlated according to a specific neighboring structure. Incorporating the temporal and spatial dimensions into a statistical model poses challenges regarding the underlying theoretical framework as well as the implementation of efficient computational methods. We propose to include spatio-temporal random effects using a conditional autoregressive prior, where the temporal correlation is modeled through an autoregressive mean decomposition and the spatial correlation by the precision matrix inheriting the neighboring structure. Their joint distribution constitutes a Gaussian Markov Random Field, whose sparse precision matrix enables the usage of efficient sampling algorithms. We cluster the areal units using a nonparametric prior, thereby learning latent partitions of the areal units. The performance of the model is assessed via an application to study regional unemployment patterns in Italy. When compared to other spatial and spatio-temporal competitors, our model shows more precise estimates, and the additional information obtained from the clustering allows for an extended economic interpretation of the unemployment rates of the Italian provinces.

E0589: **Forecasting macroeconomic data with Bayesian VARs: Sparse or dense**
Presenter: **Luis Gruber**, University of Klagenfurt, Austria
Co-authors: Gregor Kastner

Vectorautogressions (VARs) are widely applied when it comes to modeling and forecasting macroeconomic variables. In high dimensions, however, they are prone to overfitting. Bayesian methods, more concretely shrinking priors, have shown to be successful in improving prediction performance. The contribution is threefold: (1) We introduce the recently developed R^2-induced Dirichlet decomposition prior to the VAR framework and compare it to refinements of well-known priors in the VAR literature. (2) We develop a semi-global framework, in which we replace the traditional global shrinkage parameter with group-specific shrinkage parameters. We demonstrate the virtues of the proposed framework in an extensive simulation study and an empirical application forecasting data of the US economy. (3) We shed more light on the ongoing "Illusion of Sparsity" debate. We find that forecasting performances under sparse/dense priors vary across evaluated economic variables and across time frames; dynamic model averaging, however, can combine the merits of both worlds.

E0603: **Sample free inference for Bayesian inverse problems, a local approximation**
Presenter: **Odd Kolbjornsen**, University of Oslo, Norway
Co-authors: Charlotte Semin-Sanchis

Many problems of indirect measurements fit into a framework of inverse problems where the data are linked to a target property through an intermediate property. The intermediate property is essential to explain the physics of the problem, and is dependent on the target variable but can be considered a nuisance parameter in the inference. In x-ray tomography, the data are line integrals of the absorption, whereas the target property is tissue type or density. For seismic data, the interest might be the rock type or the porosity, but the physics is related to intermediate properties such as sound velocity and density. The hierarchical structure of the problem and the multiple levels of uncertainty makes the problem well-suited for a Bayesian formulation. However, the general solution to Bayesian inference through McMC sampling is, in general, too time-consuming for large-scale problems. We present an approximate computation which provides a sampling-free Bayesian inversion based on the principles of expectation propagation. The approach is valid for a large class of inverse problems. Going from a global problem, we build on the likelihood principle to provide an approximate likelihood which is suited for local inference. We show examples from CT images of rock and seismic amplitude inversion.

E0785: **Flexible Bayesian nonlinear model configuration**
Presenter: **Florian Frommlet**, Medical University Vienna, Austria
Co-authors: Geir Olve Storvik, Aliaksandr Hubin

Regression models are used in a wide range of applications, but simple nonlinear models are often not sufficient to describe complex relationships. For large data sets, neural networks have become increasingly popular for prediction tasks, but they provide less interpretable models and suffer from potential overfitting. Alternatively, nonlinear regression might be used, but the correct specification of such models is, in general, difficult. We introduce a method to construct nonlinear parametric regression models. Nonlinear features are generated hierarchically, similarly to deep learning, but even slightly more general. This amount of flexibility is combined with Bayesian variable selection, where model priors are chosen to penalize the complexity of nonlinear features. As a consequence, we end up with highly interpretable non-linear models selected from an extremely flexible model space. A genetically modified mode jumping Markov chain Monte Carlo algorithm is adopted to perform Bayesian inference and estimate model posterior probabilities. We illustrate in various applications that our algorithm is capable of delivering meaningful nonlinear models. Additionally, we compare its predictive performance with several machine learning algorithms. Finally, we hint at possible extensions for future work.

E0632: **Genetically modified mode jumping MCMC approach for Bayesian multivariate fractional polynomials**
Presenter: **Aliaksandr Hubin**, NMBU, Norway
Co-authors: Riccardo De Bin, Georg Heinze

A framework is suggested to fit fractional polynomials based on the Bayesian Generalized Nonlinear Models. A version of the Genetically Modified Mode Jumping Markov Chain Monte Carlo (GMJMCMC) algorithm is adopted. Preliminary simulation runs show promising results in terms of identifying the data generation mechanism: The suggested approach uniformly outperforms the existing Bayesian fractional polynomial framework oth in terms of Power and false discovery rate (FDR). Also, the performance is on par (somewhat better) with that of frequentist fractional polynomials. Still, the results indicate that work on the priors is likely to improve the performance even further.

EO629 Room K0.20 EXTREME VALUE STATISTICS Chair: Antoine Usseglio-Carleve

E0961: **Extreme value modelling of SARS-CoV-2 community transmission using discrete Generalised Pareto distributions**
Presenter: **Abdelaati Daouia**, Fondation Jean-Jacques Laffont, France
Co-authors: Gilles Stupfler, Antoine Usseglio-Carleve

Superspreading has been suggested to be a major driver of overall transmission in the case of SARS-CoV-2. It is, therefore, important to statistically investigate the tail features of superspreading events (SSEs) to understand virus propagation and control better. Our extreme value analysis of different sources of secondary case data indicates that SSEs associated with SARS-CoV-2 may be fat-tailed, although substantially less so than predicted recently in the literature, but also less important relative to SSEs associated with SARS-CoV. The results caution against pooling data from both coronaviruses. This could provide policy- and decision-makers with a more reliable assessment of the tail exposure to SARS-CoV-2 contamination. Going further, we consider the broader problem of large community transmission. We study the tail behaviour of SARS-CoV-2 cluster cases documented both in official reports and in the media. Our results suggest that the observed cluster sizes have been fat-tailed in the vast majority of surveyed countries. We also give estimates and confidence intervals of the extreme potential risk for those countries. A key component

of our methodology is up-to-date discrete Generalised Pareto models, which allow for maximum-likelihood-based inference of data with a high degree of discreteness.

E1603: **Regression-type analysis for multivariate extreme values**
Presenter: **Miguel de Carvalho**, FCiencias.ID - Associacao para a Investigacao e Desenvolvimento de Ciencias, Portugal
Co-authors: Alina Kumukova, Goncalo dos Reis

A regression-type model is presented for the situation where both the response and covariates are extreme. The proposed approach is designed for the setting where the response and covariates are modeled as multivariate extreme values, and thus contrary to standard regression methods, it takes into account the key fact that the limiting distribution of suitably standardized componentwise maxima is an extreme value copula. An important target in the proposed framework is the regression manifold, which consists of a family of regression lines obeying the latter asymptotic result. To learn about the proposed model from data, we employ a Bernstein polynomial prior on the space of angular densities, which leads to an induced prior on the space of regression manifolds. Numerical studies suggest a good performance of the proposed methods, and a finance real-data illustration reveals interesting aspects of the conditional risk of extreme losses in two leading international stock markets.

E1640: **Accurate ways to measure risks of extreme events**
Presenter: **Ivette Gomes**, FCiencias.ID, Universidade de Lisboa and CEAUL, Portugal
Co-authors: Frederico Caeiro, Fernanda Otilia Figueiredo, Ligia Henriques-Rodrigues

Among the great variety of alternative methodologies available to deal with the management of risks of extreme events, and for stationary sequences from a model $F(\cdot)$, with a heavy right tail function, i.e. a positive extreme value index (EVI), the value at risk (VaR) and the conditional tail expectation (CTE) will be under discussion. For these Pareto-type models, the classical EVI-estimators are the Hill (H) estimators, and hence the possibility of considering associated H VaR and CTE-estimators. Since H can be replaced by any consistent EVI-estimator, improvements in the performance of the H CTE-estimators, through the use of reliable EVI-estimators based on different generalised means, are now suggested and studied, both asymptotically and for finite samples.

E1727: **Modelling of discrete extremes through extended versions of discrete generalized Pareto distribution**
Presenter: **Touqeer Ahmad**, University of Padova, Italy, Italy
Co-authors: Carlo Gaetan, Philippe Naveau

The statistical modelling of integer-valued extremes, such as large counting of fires or avalanches, has received less attention than their continuous counterpart in the extreme value theory (EVT) literature. One approach to go from continuous to discrete extremes is to model threshold exceedances of integer random variables by the discrete version of the generalized Pareto distribution. Still, the optimal threshold selection that defines exceedances remains a problematic issue. In addition, within a regression framework, the treatment of the many data points (those below the chosen threshold) is either ignored or decoupled from extremes. Considering these issues, we extend the idea of using a smooth transition between the two tails (lower and upper) to force large and small discrete extreme values to be in compliance with EVT. In the case of zero inflation, we also develop models with an additional parameter. To incorporate covariates, we extend the Generalized Additive Models (GAM) framework to discrete extreme responses. In the GAM forms, the parameters of our proposed models are quantified as a function of covariates. The maximum likelihood estimation procedure is implemented for estimation purposes. With the advantage of bypassing the threshold selection step, our proposed models appear to be more flexible and robust than competing models (i.e. discrete generalized Pareto distribution and Poisson distribution).

E1964: **Principal component analysis of max-stable distributions**
Presenter: **Felix Reinbott**, Otto von Guericke University Magdeburg, Germany
Co-authors: Martin Schlather, Anja Janssen

Multivariate extreme value distributions in practice are often driven by few underlying physical or economic phenomena; thus, many applications are interested in recovering a latent structure in the data. We propose a procedure similar to PCA that yields a transformation to a lower dimensional space by minimizing an appropriate distance measure from the reconstruction to the data. This approach to PCA for multivariate extremes allows us to identify possible driving factors behind extreme events and has good statistical properties that preserve the structure of the extreme value distribution for the latent state. Finally, we demonstrate that the procedure is applicable to real datasets up to moderately high dimensions.

EO274 Room K0.50 DESIGN OF EXPERIMENTS AND DATA ANALYSIS Chair: Kalliopi Mylona

E1081: **Robust design for mixture experiments: An efficient class of exchangeable designs for Scheff polynomials**
Presenter: **Irene Garcia-Camacha Gutierrez**, University of Castilla-La Mancha, Spain
Co-authors: Raul Martin-Martin, Jose Luis Polo Sanz, Angela Sebastia Bargues

Modern industry, engineering, and science are interested in exploring new methods to determine the composition which optimally describes certain features of their products. The aim of mixture design is to identify the proportions of different blends over a simplex-shape experimental region which suitably describes the property under study. Designs obtained using Optimal Experimental Design (OED) theory are extremely model-dependent, and practitioners vaguely know the model form prior to run their experiments. Theory and methods are provided to address the mixture design problem of model robustness for Scheffe polynomials. Asymptotic designs are obtained by optimizing a function that assumes a potential model subject to a class of unknown contaminating functions. Theoretical results are proven for binary blends, whereas two strategies are provided for ternary blends: (i) an analytical solution for the continuous problem under strict assumptions on both the design region and the designs, and (ii) a numerical alternative for the general case, which involves the discretization of the problem. The geometrical properties observed in the obtained designs encouraged the investigation of a class of restricted designs named exchangeable designs. They keep the natural structure of symmetry in the simplex through medians. Results reveal that this class of restricted designs may be widely recommended for its simplicity, well performance and computational saving.

E1092: **Design of experiments for networks**
Presenter: **Vasiliki Koutra**, King's College London, United Kingdom

A methodology is proposed for constructing efficient designs which control for variation among the experimental units arising from network interference, so that the direct treatment effects can be precisely estimated. We provide evidence that our approach can lead to efficiency gains over conventional designs, such as randomised designs that ignore the network structure. We illustrate its usefulness for a variety of example experiments.

E1585: **D-optimal two-level designs for main-effects models: Some new results**
Presenter: **Peter Goos**, KU Leuven, Belgium
Co-authors: Jose Nunez Ares, Mohammed Saif Ismail Hameed

Two-level orthogonal arrays are known to be D-optimal for main-effects models in the event the number of runs is a multiple of 4. Complete catalogs of non-isomorphic orthogonal arrays have been enumerated and investigated to identify those orthogonal arrays that minimize the aliasing between main effects and two-factor interactions and the aliasing among two-factor interactions. It turns out that there exist many non-isomorphic D-optimal designs for main-effects models in the event the number of runs is not a multiple of 4 too, and that some of these designs involve substantially less aliasing between the main effects and the two-factor interactions as well as among the two-factor interactions. We will discuss how we identify non-isomorphic D-optimal designs for main-effects models and investigate the differences between these designs.

E1601: **MultiDOE: A multi-criteria design of experiments R package**
Presenter: **Matteo Borrotti**, University of Milan-Bicocca, Italy
Co-authors: Francesca Cucchi, Andrea Melloncelli, Francesco Sambo, Kalliopi Mylona

Many real experiments involve some factors whose levels are more difficult to set than others, most times due to high cost and/or limited time. A possible solution is to design the experiment according to a multi-stratum structure, where restrictions on the complete randomization of the experiment limit the total number of hard-to-set factor level changes. Considering optimal experimental designs, multi-criteria optimization searches for the best trade-off between competing research objectives. The MultiDoE R package can be used to construct multi-stratum experimental designs (for any number of strata) that optimize up to six statistical criteria simultaneously. In the first place, the algorithm relies on a local search procedure to find a locally optimal experimental design. More precisely, it is an extension of the Coordinate-Exchange (CE) algorithm that allows both the search of experimental designs for any type of nested multi-stratum experiment and the optimization of multiple criteria simultaneously. In the second place, the final solution implements a Two-Phase Local Search framework, that can generate a good Pareto front approximation for the optimization problem under study. The package provides different ways to choose the final optimal experimental design among those belonging to the Pareto front.

E1649: **Sequential multi-objective planning of factorial experiments with restricted randomisation**
Presenter: **Olga Egorova**, King's College London, United Kingdom
Co-authors: Steven Gilmour, Kalliopi Mylona

Implementing a series of controlled experiments in order to study a relationship between a response and a set of parameters of the process of interest is a common approach in various applications. It might be organised as a sequence of a pre-determined number of batches, or with potential additional experiments on the agenda, and it is sensible to devise a strategy of approaching the optimal design search for such frameworks sequentially as well. We focus on sequential planning of factorial experiments, including the cases of restricted randomisation. The design search methodology is constructed with multiple objectives, corresponding to (1) the quality of inference and prediction, and (2) ensuring robustness of the fitted model against potential model misspecification and managing the lack-of-fit. Gathered data and interim inference are utilised for planning the next stage through shaping the primary model and amending optimality criteria functions, including Bayesian updating of the parameters. We consider an example of a split-plot experimental setting and construct a set of Pareto-optimal designs; some particularities of the sequential planning and the choices that are to be made by the experimenters are discussed.

EO601 Room S0.11 COMPLEXITY AND COMPUTATIONAL ASPECTS OF MCMC METHODS Chair: Florian Maire

E0590: **Stereographic Markov Chain Monte Carlo**
Presenter: **Jun Yang**, University of Oxford, United Kingdom

High dimensional distributions, especially those with heavy tails, are notoriously difficult for off-the-shelf MCMC samplers: the combination of unbounded state spaces, diminishing gradient information, and local moves, results in empirically observed "stickiness" and poor theoretical mixing properties – lack of geometric ergodicity. We introduce a new class of MCMC samplers that map the original high-dimensional problem in Euclidean space onto a sphere and remedy these notorious mixing problems. In particular, we develop random-walk Metropolis type algorithms as well as versions of Bouncy Particle Sampler that are uniformly ergodic for a large class of light and heavy-tailed distributions and also empirically exhibit rapid convergence in high dimensions. In the best scenario, the proposed samplers can enjoy the "blessings of dimensionality" that the mixing time decreases with dimension.

E1617: **Adaptive MCMC methods for Bayesian variable selection**
Presenter: **Samuel Livingstone**, University College London, United Kingdom

Choosing which variables to include in a probabilistic model is a classical problem. The Bayesian solution is to place a prior distribution on a model with each possible combination of the p variables under consideration. This leads to a posterior over 2^p possible models. In order to either decide on the best model or average over them, for e.g. prediction, this model space must be explored, and when p is large, this can be challenging. Recently sophisticated Markov chain Monte Carlo (MCMC) algorithms have been proposed for this purpose. Some rely on intelligent global approximations of the posterior distribution, while others consider sophisticated moves within a neighbourhood of the current model. We will argue that in the latter case, the choice of neighbourhood is crucial to performance and scalability, and that ideas from the former case can help design such a neighbourhood, leading to algorithms that combine many recently proposed approaches to produce a practical and scalable methodology for variable selection.

E1681: **Pseudo-marginal piecewise deterministic Monte Carlo**
Presenter: **Giorgos Vasdekis**, University College London, United Kingdom

Piecewise Deterministic Markov Processes (PDMPs) have recently caught the attention of the MCMC community for having a non-diffusive behavior, potentially allowing them to explore the state space efficiently. This makes them good candidates for generating MCMC algorithms. One important problem in Bayesian computation is inference for models where the pointwise evaluation of the posterior is not available, but one has access to an unbiased estimator of the posterior. A technique to deal with this problem is the Pseudo-marginal Metropolis-Hastings algorithm. We describe a PDMP algorithm that can be used in the same posterior free setting and can be seen as the analogue of Pseudo-marginal for Piecewise Deterministic Monte Carlo. We show that the algorithm targets the posterior of interest. We also provide some numerical examples, focusing on the case of Approximate Bayesian Computation (ABC), a popular method to deal with problems in the setting of likelihood-free inference.

E1823: **Parallel tempering with a variational reference**
Presenter: **Saifuddin Syed**, University of Oxford, United Kingdom
Co-authors: Trevor Campbell, Nikola Surjanovic, Alexandre Bouchard

Sampling from complex target distributions is a challenging task fundamental to Bayesian inference. Parallel tempering (PT) addresses this problem by constructing a Markov chain on the expanded state space of a sequence of distributions interpolating between the posterior distribution and a fixed reference distribution, which is typically chosen to be the prior. However, in the typical case where the prior and posterior are nearly mutually singular, PT methods are computationally prohibitive. We address this challenge by constructing a generalized annealing path connecting the posterior to an adaptively tuned variational reference. The reference distribution is tuned to minimize the forward (inclusive) KL divergence to the posterior distribution using a simple, gradient-free moment-matching procedure. We show that our adaptive procedure converges to the forward KL minimizer, and that the forward KL divergence serves as a good proxy for a previously developed measure of PT performance. We also show that in the large-data limit in typical Bayesian models, the proposed method improves in performance, while traditional PT deteriorates arbitrarily. Finally, we introduce PT with two references one fixed, one variational, with a novel split annealing path that ensures stable variational reference adaptation. Finally, experiments that demonstrate the large empirical gains achieved by our method in a wide range of realistic Bayesian inference scenarios are discussed.

E0286: **Manifold Markov chain Monte Carlo methods for Bayesian inference in diffusion models**
Presenter: **Alexandros Beskos**, University College London, United Kingdom
Co-authors: Matthew Graham, Alexandre Thiery

Bayesian inference for nonlinear diffusions, observed at discrete times, is a challenging task that has prompted the development of a number of

algorithms, mainly within the computational statistics community. We propose a new direction, and accompanying methodology borrowing ideas from statistical physics and computational chemistry for inferring the posterior distribution of latent diffusion paths and model parameters, given observations of the process. Joint configurations of the underlying process noise and of parameters, mapping onto diffusion paths consistent with observations, form an implicitly defined manifold. Then, by making use of a constrained Hamiltonian Monte Carlo algorithm on the embedded manifold, we are able to perform computationally efficient inference for a class of discretely observed diffusion models. Critically, in contrast with other approaches proposed in the literature, our methodology is highly automated, requiring minimal user intervention and applying alike in a range of settings, including: elliptic or hypo-elliptic systems; observations with or without noise; linear or non-linear observation operators. Exploiting Markovianity, we propose a variant of the method with complexity that scales linearly in the resolution of path discretisation and the number of observation times.

EO112 Room S0.13 PROJECTION PURSUIT: THEORY Chair: Nicola Loperfido

E1144: **Partial least squares and interesting directions in data**
Presenter: **John Kent**, University of Leeds, United Kingdom

Consider the usual multiple linear regression of a response random variable y on a p-dimensional vector of explanatory random variables x. Ordinary least squares estimation looks for the linear combination of x that has the highest correlation with y. In contrast, partial least squares (PLS) is an iterative method; the first iteration looks for the standardized linear combination of x that has the highest covariance with y. The focus on standardized linear combinations makes PLS a "regularized" method of regression analysis. Higher-order iterations yield linear combinations of x that have no "direct" correlation with y, but instead have an "indirect" correlation through their correlations with earlier linear combinations of x. Partial least squares have close links to envelope models and Krylov matrix decompositions. The sequence of optimal linear combinations identified by PLS can be viewed as a sequence of random variables that is dual to a one-dimensional Gaussian Markov chain; indirect correlation in the regression setting corresponds to conditional dependence in the Markov chain setting. These connections provide some novel insights into the behavior of PLS.

E0763: **The location of a minimum variance squared distance functional**
Presenter: **Zinoviy Landsman**, University of Haifa, Israel
Co-authors: Tomer Shushi

A novel multivariate functional is introduced that represents a position where the intrinsic uncertainty of a system of mutually dependent risks is maximally reduced. The proposed multivariate functional defines the location of the minimum variance of squared distance (LVS) for some n-variate vector of risks X. We compute the analytical representation of $\mathrm{LVS}(X)$, which consists of the location of the minimum expected squared distance, $\mathrm{LES}(X)$, covariance matrix A, and a matrix B of the multivariate central moments of the third order of X. From this representation, it follows that $\mathrm{LVS}(X)$ coincides with $\mathrm{LES}(X)$ when X has a multivariate symmetric distribution, but differs from it in the non-symmetric case. As $\mathrm{LES}(X)$ is often considered a neutral multivariate risk measure, we show that $\mathrm{LVS}(X)$ also possesses the important properties of multivariate risk measures: translation invariance, positive homogeneity, and partial monotonicity. We also study the mean-variance approach based on the balanced sum of an expectation and a variance of the square of the aforementioned Euclidean distance and control for the closeness of $\mathrm{LES}(X)$ and $\mathrm{LVS}(X)$. The proposed theory and the results are distribution-free, meaning that we do not assume any particular distribution for the random vector X. The results are demonstrated with real data on Danish fire losses.

E0788: **Optimal portfolio projections and their applications to skew-elliptically distributed portfolio returns**
Presenter: **Tomer Shushi**, Ben Gurion University of the Negev, Israel
Co-authors: Nicola Loperfido

The concept of optimal portfolio projection is defined, as a procedure that projects the vector of weights of the portfolio return to a lower dimension such that one can explicitly solve the problem of optimal portfolio selection for any given risk measure. We study the class of skew-elliptically distributed risks. We show that following the proposed procedure, we are able to obtain explicit optimal weights for such risks, with a dramatic reduction of the complexity of such an optimization problem.

E0341: **A computational perspective on projection pursuit in high dimensions: Feasible or infeasible feature extraction**
Presenter: **Chunming Zhang**, University of Wisconsin-Madison, United States

Finding a suitable representation of multivariate data is fundamental in many scientific disciplines. Projection pursuit (PP) aims to extract interesting "non-Gaussian" features from multivariate data and tends to be computationally intensive even when applied to data of low dimension. In high-dimensional settings, recent work on PP addresses asymptotic characterization and conjectures of the feasible projections as the dimension grows with sample size. To gain practical utility and learn theoretical insights into PP in an integral way, data analytic tools needed to evaluate the behaviour of PP in high dimensions become increasingly desirable but are less explored in the literature. The focus is on developing computationally fast and effective approaches central to finite sample studies for (i) visualizing the feasibility of PP in extracting features from high-dimensional data, as compared with alternative methods like PCA and ICA, and (ii) assessing the plausibility of PP in cases where asymptotic studies are lacking or unavailable, with the goal of better understanding the practicality, limitation and challenge of PP in the analysis of large data sets.

E0907: **Projection pursuit in high dimensions**
Presenter: **Nicola Loperfido**, University of Urbino, Italy

Projection pursuit is a multivariate statistical technique aimed at finding interesting data projections. It suffers from several problems when applied to high-dimensional datasets. These problems are investigated within the framework of skewness-based projection pursuit, when the interesting projections are the maximally skewed ones. We address the problems by means of generalized tensor eigenvectors and symmetrizing linear projections. We illustrate the problems and the proposed solutions with a simple dataset with more variables than units.

EO060 Room Safra Lecture Theatre METHODS AND APPLICATIONS FOR FUNCTIONAL DATA ANALYSIS Chair: Enea Bongiorno

E1084: **Mapping Brexit debate on twitter via functional graphical models**
Presenter: **Nicola Pronello**, University of Chieti-Pescara, Italy
Co-authors: Emiliano del Gobbo, Lara Fontanella, Rosaria Ignaccolo, Luigi Ippoliti, Sara Fontanella

In recent years a literature on multivariate functional graph models has been developed. The graphical representation of the conditional dependence among a finite number of random variables is indeed appealing in different applications, such as for example the analysis of the brain connectivity. We want to investigate a novel extension of this methodology, considering random functions spatially and temporally correlated. A motivating example is the analysis of the semantic network emerging from twitter users. In particular the main goal of our analysis is to track the change of the Brexit debate on Twitter across UK during a particular time frame. By considering the change in time of a word usage as a functional realization, the semantic network regarding the topic of interest is then defined as a graphical representation of the conditional dependence among functional variables. Since each tweet considered is localized in both time and space we shall take into accounts such features to properly define the functional semantic network.

E1112: **Robust control charts for multivariate functional data**
Presenter: **Christian Capezza**, University of Naples Federico II, Italy

Co-authors: Fabio Centofanti, Antonio Lepore, Biagio Palumbo

Profile monitoring evaluates the stability of a functional quality characteristic over time in order to identify special causes of variation that affect a process. Modern manufacturing processes in Industry 4.0 applications allow for acquiring large amounts of profile data, which, however, are frequently contaminated by anomalous observations in the form of both casewise and cellwise outliers. Then, profile monitoring techniques need to cope with outliers since they can significantly influence the monitoring performance. To achieve this, we propose a novel framework, referred to as a robust multivariate functional control chart (RoMFCC), which can monitor multivariate functional data while being robust to both functional casewise and cellwise outliers. The RoMFCC relies on four key components: (I) a univariate filter to find functional cellwise outliers that are replaced by missing components; (II) a robust functional data imputation method for missing values; (III) a casewise robust dimensionality reduction; and (IV) a monitoring strategy for the multivariate functional quality characteristic. To compare the RoMFCC with other competing methods in the literature, a thorough Monte Carlo simulation analysis is conducted to assess the monitoring performance of the RoMFCC. The proposed framework is then applied to monitor a resistance spot welding process in the automotive industry in a motivating real-case study.

E1381: **Analysis of variability in three-dimensional curves: Development of a generative model**

Presenter: **Perrine Chassat**, LaMME, Universita Paris-Saclay, CNRS, France

Co-authors: Juhyun Park, Nicolas Brunel

Current advanced signal recording techniques offer new data as multidimensional functional observations, such as human motion trajectories. Developing generative models for multidimensional curves is an important task in statistical functional analysis. These models must consider the geometric characteristics inherent to these multidimensional curves and thus differ from classical approaches developed for scalar functional data. Two types of variation play a key role in characterizing multidimensional curves: parametrisation (or time) variability and shape variability. We identify these two variations using a method based on the geometric representation of curves by the Frenet-Serret equations, defining the mean and variations within a set of curves through the mean of functional parameters characterising geometry, velocity, and their non-linear transformations. Inspired by generative models of scalar functional data considering phase and amplitude variations, we analyse the non-linear variations in time and geometry by functional PCA. Completing with classical PCA of the curves' initial positions in Euclidean space and an extended PCA on manifolds of the initial orientations, we develop a generative model by imposing a joint probability model on the principal coefficients of these PCA components. We apply this method to develop a realistic generative model of three-dimensional wrist movement trajectories in sign language from a real data set.

E1685: **A specification test for the single functional index model**

Presenter: **Kwo Lik Lax Chan**, Universita degli Studi del Piemonte Orientale, Italy

The problem of specifying the link function that models the dependence structure between two random elements is a very important task in regression analysis. In the framework of scalar on-function regression, the link function is described through a real-valued operator acting on a functional space, and it is difficult to visualise and hence select a coherent specification. A test for specification in such a framework that exploits semi-parametric principles is illustrated, in particular by exploiting the Single Functional Index Model. The test statistic is a special form of U-statistic, and once it is defined, its asymptotic null distribution is derived under suitable conditions. The finite sample performances of the test are analyzed through a simulation study by using both the asymptotic p-value and some bootstrap approaches. To demonstrate the potentialities of the method, an application to a spectrometric real dataset is performed.

E0797: **Flexible Hilbertian additive regression with small errors-in-variables**

Presenter: **Germain Van Bever**, Universite de Namur, Belgium

Co-authors: Jeong Min Jeon

A new framework is presented for additive regression modelling for data in very generic settings. More precisely, we tackle the problem of estimating component functions of additive models where the regressors and/or response variable belong to general Hilbert spaces and can be imperfectly observed. By this, we mean that some variables can be either measured incompletely or with errors. Smooth backfitting methods are used to estimate the component functions consistently and we provide explicit rates of convergence. We amply illustrate our methodology in various settings, including the functional, Riemannian and Hilbertian settings.

EO614 Room Virtual R01 NON-REGULAR TECHNIQUES FOR STATISTICAL MODELING AND COMPUTING Chair: Tsung-I Lin

E0206: **Multivariate linear mixed models with censored and nonignorable missing outcomes**

Presenter: **Wan-Lun Wang**, National Cheng Kung University, Taiwan

Co-authors: Tsung-I Lin

The analysis of multivariate longitudinal data could encounter some complications due to censorship induced by detection limits of the assay and non-response occurring when participants missed scheduled visits intermittently or discontinued participation. A generalization of the multivariate linear mixed model is established that can accommodate censored responses and nonignorable missing outcomes simultaneously. To account for the nonignorable missingness, the selection approach which decomposes the joint distribution as a marginal distribution for the primary outcome variables and a model describing the missing process conditional on the hypothetical complete data is used. A computationally feasible Monte Carlo expectation conditional maximization (MCECM) algorithm is developed for parameter estimation with the maximum likelihood (ML) method. Furthermore, a general information-based approach is presented to assess the variability of ML estimators. The techniques for the prediction of censored responses and the imputation of missing outcomes are also discussed. The methodology is motivated and exemplified by a real dataset concerning HIV-AIDS clinical trials. A simulation study is conducted to examine the performance of the proposed method compared with other traditional approaches.

E0398: **Robust class of non-normal cluster-wise regression models**

Presenter: **Elham Mirfarah**, National Cheng Kung University, Taiwan

Co-authors: Mehrdad Naderi, Wan-Lun Wang, Tsung-I Lin

One of the widely used statistical frameworks for modelling, classification, and clustering of data is to adopt a mixture regression model (MRM), in which it is assumed that data coming from several hidden clusters have different regression functions. Since the conventional MRMs are sensitive to departures from normality, caused by extra skewness and possible heavy tails, various extensions built on flexible distributions have been put forward in the last decade. The class of normal mean-variance mixture (NMVM) distributions that arise from scaling both the mean and variance of a normal random variable with a common mixing distribution encompasses many prominent (symmetric or asymmetrical) distributions as special cases. We aim to introduce a unified approach to robustifying MRMs by considering the class of NMVM distributions for component errors. An analytical expectation-maximization (EM) type algorithm is developed to obtain the maximum likelihood parameter estimates. The finite-sample performance, effectiveness, and robustness of the proposed model against outliers for contaminated and noisy data are illustrated by conducting four simulation studies and analyzing two real-world datasets.

E0402: **Joint random partition models for multivariate change point analysis**

Presenter: **Mauricio Castro**, Pontificia Universidad Catolica de Chile, Chile

Co-authors: Jose Javier Quinlan Binelli, Garritt Page

Change point analyses are concerned with identifying positions of an ordered stochastic process that undergo abrupt local changes of some un-

derlying distribution. When multiple processes are observed, it is often the case that information regarding the change point positions is shared across the different processes. A method is described that takes advantage of this type of information. Since the number and position of change points can be described through a partition with contiguous clusters, our approach develops a joint model for these types of partitions. We describe computational strategies associated with our approach and illustrate improved performance in detecting change points through a small simulation study. We then apply our method to a financial data set of emerging markets in Latin America and highlight interesting insights discovered due to the correlation between change point locations among these economies.

E0418: **Robust clusterwise regression analysis for the censored data**
Presenter: **Mehrdad Naderi**, National Chung Hsing University, Taiwan
Co-authors: Elham Mirfarah

In clusterwise regression modelling, it is assumed that data coming from several hidden clusters have different regression links. Various methods are used to identify observations' membership and subsequently to estimate each regression parameter. One of the widely used statistical frameworks for analyzing, clustering and classification purposes is a Mixture of linear Expert (MoE) models. Compared to the finite mixture regression models, the MoE models exploit the logistic function to allocate each observation to a specific group. This advantage of the MoE models enables us to use more information from the data and obtain an improvement in the data clustering. We introduced a robust MoE model for model-based clustering of the censored data with the scale-mixture of normal (SMN)distributional assumption on the unobserved error terms. An analytical expectation-maximization (EM) type algorithm is developed to obtain the maximum likelihood parameter estimates. The performance, effectiveness, and robustness of the proposed methodology are illustrated by conducting various simulation studies and analyzing a real-world dataset.

EO392 Room Virtual R02 ADVANCES IN NON- AND SEMI-PARAMETRIC INFERENCE FOR COMPLEX DATA Chair: Catia Scricciolo

E0458: **Drift burst test statistic in a pure jump semimartingale model**
Presenter: **Cecilia Mancini**, University of Verona, Italy

The focus is on a recent test statistic devised to obtain insight into the causes of *flash crashes* occurring at particular moments in time in the price of a financial asset. Under an Ito semimartingale model containing a Brownian component and finite variation jumps, it is possible to distinguish when the cause is a drift burst (the statistic explodes) or not (it is asymptotically Gaussian). We complete the investigation showing how infinite variation jumps contribute asymptotically. The result is that, when there are no bursts, explosion only can occur in the absence of the Brownian part and when the jumps have finite variation. In that case, the explosion is due to the compensator of the small jumps. We also find that the statistic could be adopted for a variety of tests useful for investigating the nature of the data-generating process, given discrete observations.

E1245: **Bayesian sensitivity analysis for a missing data model**
Presenter: **Bart Eggen**, Delft University of Technology, Netherlands
Co-authors: Aad van der Vaart, Stephanie van der Pas

In many fields, sensitivity analysis is very important to assess the robustness of study conclusions to key assumptions. We consider the missing outcomes model and perform sensitivity analysis under the assumption that missing outcomes are missing completely at random. We provide theoretical guarantees for a Bayesian approach to estimating the mean outcome, conditional on the sensitivity parameter. We show two Bernstein-von Mises theorems for different parametrisations of the model. The results are obtained using Dirichlet process priors on the distribution of the outcome and on the distribution of the outcome conditional on being observed. We also provide a simulation study, showing the performance of the methods in finite sample scenarios.

E0455: **Adaptive deep learning for nonparametric time series regression**
Presenter: **Daisuke Kurisu**, Yokohama National University, Japan
Co-authors: Riku Fukami, Yuta Koike

A general theory is developed for adaptive nonparametric estimation of mean functions of nonstationary and nonlinear time series using deep neural networks (DNNs). We first consider two types of DNN estimators, non-penalized and sparse-penalized DNN estimators, and establish their generalization error bounds for general nonstationary time series. We then derive minimax lower bounds for estimating mean functions belonging to a wide class of nonlinear autoregressive (AR) models that include nonlinear generalized additive AR, single index, and threshold AR models. Building upon the results, we show that the sparse-penalized DNN estimator is adaptive and attains the minimax optimal rates up to a poly-logarithmic factor for many nonlinear AR models. Through numerical simulations, we demonstrate the usefulness of the DNN methods for estimating nonlinear AR models with intrinsic low-dimensional structures and discontinuous or rough mean functions, which is consistent with our theory.

E1736: **Semiparametric Bayesian two-stage meta-analysis for association between ambient temperature and new cases of COVID-19**
Presenter: **Dongu Han**, Korea University, Korea, South
Co-authors: Kiljae Lee, Yeonseung Chung, Taeryon Choi

In environmental epidemiological studies, two-stage meta-analysis has been a popular tool to investigate a short-term association between environmental exposure and a health response by analyzing daily time-series data collected from multiple locations. We propose a novel Bayesian approach for a two-stage meta-analysis by innovating each of the existing first and second-stage models in Bayesian frameworks. Specifically, for the first stage model, we propose a new Bayesian distributed lag nonlinear model which accommodates three kinds of nonlinearities. For the second stage model, we propose a matrix-variate Dirichlet process mixture multivariate meta-regression that is robust when the assumptions of existing linear models are violated. The proposed second stage model also allows for identifying subgroups of locations through model-based clustering. We validate the methodologies through simulation studies and apply them to study a short-term association between ambient temperature and new cases of COVID-19 in South Korea and the United States.

E1752: **A general framework for constructing locally self-normalized multiple-change-point tests**
Presenter: **Kin Wai Chan**, The Chinese University of Hong Kong, Hong Kong
Co-authors: Cheuk Hin Cheng

A general framework is proposed to construct self-normalized multiple-change- point tests with time series data. The only building block is a user-specified one-change-point detecting statistic, which covers a wide class of popular methods, including cumulative sum process, outlier-robust rank statistics and order statistics. Neither robust and consistent estimation of nuisance parameters, selection of bandwidth parameters, nor pre-specification of the number of change points is required. The finite-sample performance shows that our proposal is size-accurate, robust against misspecification of the alternative hypothesis, and more powerful than existing methods. Case studies of NASDAQ option volume and Shanghai-Hong Kong Stock Connect turnover are provided.

EO218 Room Virtual R03 RECENT ADVANCES IN DIRECTIONAL STATISTICS Chair: Arthur Pewsey

E0572: **Multimodal regression with circular data**
Presenter: **Rosa Crujeiras**, University of Santiago de Compostela, Spain
Co-authors: Maria Alonso-Pena

There is a diverse range of practical situations where one may encounter random variables which are not defined in Euclidean spaces, as is the

case for circular data. Circular measurements may be accompanied by other observations, either defined on the unit circumference or the real line and, in such cases, it may be of interest to model the relationship between the variables from a regression perspective. It is not infrequent that parametric models fail to capture the underlying model, given their lack of flexibility. Still, it may also happen that the usual paradigm of (classical) mean regression. Some recent advances in nonparametric multimodal regression, showing an adaptation of the mean-shift algorithm for regression scenarios involving circular response and/or covariate. Real data illustrations will also be presented.

E0443: **Geodesic projection of directional distributions for projection pursuit**
Presenter: **Sungkyu Jung**, Seoul National University, Korea, South

Geodesic projections of directional data are investigated, following either von Mises-Fisher (vMF) distribution or the angular Gaussian (AG) distribution. The vMF distribution for random directions on the $(p-1)$-dimensional unit hypersphere $\mathbb{S}^{p-1} \subset \mathbb{R}^p$ plays the role of multivariate normal distribution in directional statistics, and the rotationally symmetric AG distribution is very similar to the vMF. Projections onto geodesics are one of the main ingredients of modeling and exploring directional data. We show that the projection of vMF-distributed random directions onto any geodesic is approximately vM-distributed, albeit not exactly the same, while the projection of any AG distribution onto subspheres, including geodesics, is AG-distributed. As one of the potential applications of the result, we contemplate a projection pursuit exploration of high-dimensional directional data. We show that in a high dimensional model almost all geodesic projections of directional data are nearly vM, and sometimes exactly AG, thus measures of non-vM-ness are a viable candidates for projection index.

E0347: **The role of finite sample smeariness for directional statistics**
Presenter: **Stephan Huckemann**, University of Goettingen, Germany
Co-authors: Benjamin Eltzner, Shayan Hundrieser

A central tool in nonparametric directional statistics is the Fr'echet mean, and, more generally, so-called generalized Fr'echet means, such as, for instance, principal nested spheres. Even if they are unique - which is a very nontrivial issue that is still to date not satisfactorily solved on spheres of dimension higher than 1 - and even if they exhibit asymptotic normality scaled with the root of sample size, just as their Euclidean kin, this approximation turns out to be invalid in practice in many situations. This phenomenon is called finite sample smeariness (FSS). We give a complete picture of FSS on circles and tori and discuss open questions on spheres. It turns out that the bootstrap can satisfactorily deal with FSS, allowing for statistically controlled inference under FSS. This is not the case in general, however, for classical quantile-based inference. In view of assessing the impact of climate change, we illustrate the role of FSS and inference under the presence of FSS in directional wind data of European cities over the last 20 years.

E0939: **A Cauchy-type model for cylindrical data**
Presenter: **Shogo Kato**, Institute of Statistical Mathematics, Japan
Co-authors: Arthur Pewsey

Cylindrical data consist of bivariate observations on a linear variable and circular variable pairing and arise in numerous scientific contexts. A family of five-parameter distributions for cylindrical data is proposed. Its density can be expressed in a simple closed form involving no integrals, infinite sums or special functions. Moreover, the proposed model is unimodal, its five parameters have clear interpretations, and all of its marginal and conditional distributions are either Cauchy or wrapped Cauchy. Related regression models follow from an alternative complex representation of the density, and their regression curves can be expressed using fractional linear transformations. When they exist, the method of moments estimators of the parameters of the proposed cylindrical distribution has closed-form expressions. The Fisher information matrix and the asymptotic covariance matrix for the maximum likelihood estimator have simple closed forms that involve no integrals.

E0804: **Inference for a Cauchy-type model for cylindrical data**
Presenter: **Arthur Pewsey**, University of Extremadura, Spain
Co-authors: Shogo Kato

Cylindrical data arise when the values taken by a linear variable and a circular variable are jointly observed, and consequently abound in numerous scientific disciplines. We consider inference for an appealing unimodal model for such data whose: density can be expressed in a simple closed form involving no integrals, infinite sums or special functions; parameters have clear interpretations; marginal and conditional distributions are all either Cauchy or wrapped Cauchy. For parameter estimation, we suggest a combined method of moments and robust approach and numerically-based maximum likelihood estimation, both of which are found to be computationally fast. We also propose a successful model validation tool, permutation tests for independence between the two variables, and a parametric bootstrap goodness-of-fit test. Results from Monte Carlo experiments designed to explore the finite sample characteristics of some of the inferential methods are presented, and the application of the proposed model and methods is illustrated in an analysis of data recorded during a three-day period up to and including the Great East Japan Earthquake which took place on 11th March 2011.

EO648 Room Virtual R04 ADVANCES IN ROBUST FUNCTIONAL AND HIGH DIMENSIONAL DATA ANALYSIS Chair: Ioannis Kalogridis

E0475: **Robust adaptive variable selection in ultra-high dimensional linear regression models**
Presenter: **Abhik Ghosh**, Indian Statistical Institute, India
Co-authors: Maria Jaenada, Leandro Pardo

The focus is on the problem of simultaneous variable selection and estimation of the corresponding regression coefficients in an ultra-high dimensional linear regression model. The adaptive penalty functions are used in this regard to achieve the oracle variable selection property along with an easier computational burden. However, the usual adaptive procedures (e.g., adaptive LASSO) based on the squared error loss function are extremely non-robust in the presence of data contamination. We present a regularization procedure for the ultra-high dimensional data using a robust loss function based on the popular density power divergence (DPD) measure along with the adaptive LASSO penalty. We theoretically study the robustness and the large-sample properties of the proposed adaptive robust estimators for a general class of error distributions; in particular, we show that the proposed adaptive DPD-LASSO estimator is highly robust, satisfies the oracle variable selection property, and the corresponding estimators of the regression coefficients are consistent and asymptotically normal under easily verifiable set of assumptions. Numerical illustrations are provided for the mostly used normal error density. Finally, the proposal is applied to analyze an interesting spectral dataset in the field of chemometrics.

E1781: **AMP meets expectiles: Testing heteroscedasticity and asymmetry in high dimensions**
Presenter: **Jing Zhou**, KU Leuven, Belgium
Co-authors: Hui Zou

Heteroscedasticity is commonly observed in high-dimensional data and has been understudied. Specifically, testing heteroscedasticity and asymmetry of the error distribution in high dimensions remains underexplored. We introduce an expectile-based test that exploits the conditional distribution of the response variable at different expectile levels. We propose to use the approximate message-passing algorithm to perform an asymptotic analysis for the proposed test, assuming the sample size and the number of predictive variables follow a linear growth rate. The numerical performance of the proposed test will be validated by using simulated and real data.

E0810: **Doubly robust feature selection with mean and variance outlier detection and oracle properties**
Presenter: **Luca Insolia**, University of Geneva, Switzerland

Co-authors: Francesca Chiaromonte, Runze Li, Marco Riani

High-dimensional linear regression models are nowadays pervasive in most research domains. We propose a general approach to handle data contaminations that might disrupt the performance of feature selection and estimation procedures. Specifically, we consider the co-occurrence of mean-shift and variance-inflation outliers, which can be modeled as additional fixed and random components, respectively, and evaluated independently. Our proposal performs feature selection while detecting and down-weighting variance-inflation outliers, detecting and excluding mean-shift outliers, and retaining non-outlying cases with full weights. Feature selection and mean-shift outlier detection are performed through a robust class of nonconcave penalization methods. Variance-inflation outlier detection is based on the penalization of the restricted posterior mode. The resulting approach satisfies a robust oracle property for feature selection in the presence of data contamination – which allows the number of features to exponentially increase with the sample size – and detects truly outlying cases of each type with asymptotic probability one. This provides an optimal trade-off between a high breakdown point and efficiency. Effective and computationally efficient heuristic procedures are also presented. We illustrate the finite-sample performance of our proposal through an extensive simulation study and real-world applications.

E0583: **Robust penalized estimators for functional linear regression**

Presenter: **Stefan Van Aelst**, University of Leuven, Belgium

Co-authors: Ioannis Kalogridis

Functional data analysis is a fast-evolving branch of statistics, but estimation procedures for the popular functional linear model either suffer from a lack of robustness or are computationally burdensome. To address these shortcomings, a flexible family of penalized lower-rank estimators based on a bounded loss function is proposed. The proposed class of estimators is shown to be consistent and can attain high rates of convergence with respect to prediction error under weak regularity conditions. These results can be generalized to higher dimensions under similar assumptions. The finite-sample performance of the proposed family of estimators is investigated by a Monte-Carlo study which shows that these estimators reach high efficiency while offering protection against outliers. The proposed estimators compare favourably to existing robust as well as non-robust approaches. The good performance of our method is also illustrated on a real complex dataset.

EO156 Room K2.31 (Nash Lec. Theatre) MODEL ASSESSMENT Chair: Maria Dolores Jimenez-Gamero

E0379: **Conditional distribution function estimation and bandwidth selection under censoring**

Presenter: **Dimitrios Bagkavos**, University of Ioannina, Greece

Co-authors: Montserrat Guillen, Jens Perch Nielsen

A smooth and continuous nonparametric estimate of the conditional cumulative distribution function for any arbitrary number of covariates in the right censored data setting is proposed. The estimate is obtained as the combination of the multivariate local linear smoothing of the data across all covariate dimensions and the kernel smoothing of the Kaplan-Meier estimate in the response direction. Implementation of the estimate in practice is facilitated by the development of a plug-in type bandwidth selector. The rule yields different amounts of smoothing for each coordinate direction, thus optimizing the estimate's performance across the full spectrum of its support. The asymptotic properties of all methodological contributions are quantified analytically and discussed in detail. Finally, numerical evidence is provided on the finite sample performance of the proposed methodological advances, and a real data analysis illustrates their benefits in practice.

E0496: **A low computational cost approximation for a independence test in non-parametric regression**

Presenter: **Gustavo Rivas**, National University of Asuncion, Paraguay

Co-authors: Maria Dolores Jimenez-Gamero

A common assumption in nonparametric regression models is the independence of the covariate and the error. Some procedures have been suggested for testing that hypothesis. A test statistic is considered, which compares estimators of the joint and the product of the marginal characteristic functions of the covariate and the error. It is proposed to approximate the null distribution of such a statistic by means of a weighted bootstrap estimator. The resulting test is able to detect any fixed alternative as well as local alternatives converging to the null at the rate $n^{1/2}$, n denoting the sample size. The finite sample performance of this approximation is assessed by means of a simulation study, where it is also compared with other estimators. From a computational point of view, the proposed approximation is very efficient. Two real data set applications are also included.

E0531: **Model checking for logistic models when the number of parameters tends to infinity**

Presenter: **Feifei Chen**, Beijing Normal University, China

Co-authors: Xiumin Li, Hua Liang, David Ruppert

A projection-based test is proposed to check logistic regression models when the dimension of the covariate vector may be divergent. The proposed test achieves a reduction in dimension, and the proposed method behaves as if only a single covariate is present. The test is shown to be consistent and can detect root-n local alternatives. We derive the asymptotic distribution of the proposed test under the null hypothesis and establish the test's asymptotic behavior under the local and global alternatives. The numerical performance is remarkably attractive compared to the existing methods. Real examples are presented for illustration.

E0916: **A goodness-of-fit test for the compound Poisson exponential model**

Presenter: **Daniel Gaigall**, FH Aachen University of Applied Sciences, Germany

Co-authors: Ludwig Baringhaus

On the basis of bivariate data, assumed to be observations of independent copies of a random vector (S,N), we consider testing the hypothesis that the distribution of (S,N) belongs to the parametric class of distributions that arise with the compound Poisson exponential model. Typically, this model is used in stochastic hydrology, with N as the number of raindays, and S as total rainfall amount during a certain time period, or in actuarial science, with N as the number of losses, and S as total loss expenditure during a certain time period. The compound Poisson exponential model is characterized in the way that a specific transform associated with the distribution of (S,N) satisfies a certain differential equation. Mimicking the function part of this equation by substituting the empirical counterparts of the transform we obtain an expression the weighted integral of the square of which is used as test statistic. We deal with two variants of the latter, one of which being invariant under scale transformations of the S-part by fixed positive constants. Critical values are obtained by using a parametric bootstrap procedure. The asymptotic behavior of the tests is discussed. A simulation study demonstrates the performance of the tests in the finite sample case. The procedure is applied to rainfall data and to an actuarial dataset. A multivariate extension is also discussed.

E1150: **Homogeneity of marginal distributions for a large number of populations**

Presenter: **Virtudes Alba-Fernandez**, University of Jaen, Spain

Co-authors: Maria Dolores Jimenez-Gamero

Assume that a random vector (X,Y) is observed in k populations, and independent samples of that random vector are available at each population. Assume that X and Y have the same dimension. The aim is to test the equality of the marginal distributions of X and Y in the k populations when k is large in comparison with the sample sizes. We consider a test statistic that compares the empirical characteristic functions in each population. Under the null, the test statistic is asymptotically normally distributed. A simulation study investigates the finite sample properties of the proposed test.

E1237: **New tests for the Weibull distribution using Stein's method in the presence of random right censoring**

Presenter: **Jaco Visagie**, North-West University, South Africa

Co-authors: James Allison, Elzanie Bothma

Two new classes of tests are developed for the Weibull distribution based on Stein's method. The proposed tests are applied in the full sample case as well as in the presence of random right censoring. We investigate the finite sample performance of the new tests using a Monte Carlo study. In both the absence and presence of censoring, it is found that the newly proposed classes of tests outperform competing tests against the majority of the alternative distributions considered. In the cases where censoring is present, we consider various censoring distributions.

EC825 Room S-2.25 APPLIED MACHINE LEARNING Chair: Alejandro Murua

E1775: **Emotions affect investors' willingness to take risks during trading sessions: A machine learning experimental research**

Presenter: **Sofia Poggi**, Sapienza University, Italy

Analyzing human emotions is fundamental in decision-making, considering that they determine more than 90% of our behaviors. The aim is to capture emotions during trading decision-making leveraging cutting-edge technology: artificial intelligence can transform micro-facial expressions into emotions. To better understand whether fear could be responsible for the change in risk aversion and to better identify the emotional channel, we rely on treatment and control in field experiments on an actual trade market. Half of the participants will watch a short horror video before a trading session. Since the subject will be randomly assigned to watch the video, the idea is that this difference in treatment should entirely drive the difference in risk aversion and emotion intensity detected between the groups. For example, watching a horror movie triggers an emotional and physical response similar to those produced by a severe financial loss. In a few words, the purpose is to detect the impact of emotion on investors' willingness to take risks during trading sessions by applying a Machine learning algorithm able to see facial micro-expression and transform it into data.

E1494: **Feature ranking in the diagnosis of cardiovascular diseases**

Presenter: **Arash Negahdari Kia**, University of Limerick, Ireland

Co-authors: Paria Sarzaeim, Kevin Burke, Parisa Shamsi

Since cardiovascular diseases such as heart attacks are one of the leading causes of death annually, it is essential to diagnose them correctly or more accurately. Many reasons may cause cardiovascular diseases, like high blood pressure or high cholesterol. Such symptoms can help diagnose cardiovascular diseases. Developing prediction models requires an understanding of which predictors are most relevant in diagnosis. There are many machine learning algorithms for feature selection, each with advantages and disadvantages. The first step is to do an exhaustive search on a range of machine learning predictors. This will enable us to find the models with the highest performance using different evaluation criteria. The effect of each feature elimination on various models and evaluation criteria is calculated in best-fit models. These reductions are then used in a novel weighted average score to calculate the importance score of each feature. Cleveland and Kaggle datasets for Heart Disease Diagnosis are employed. Finally, features are ranked based on their score and the results regarding the best-fit models and feature ranks are discussed.

E1943: **Top-down classifiers with hierarchy based on taxonomic resolution used by human experts**

Presenter: **Salme Karkkainen**, University of Jyvaskyla, Finland

Co-authors: Johanna Arje, Jenni Raitoharju, Alexandros Iosifidis, Ville Tirronen, Kristian Meissner, Moncef Gabbouj, Serkan Kiranyaz

Our problem arises from bioassessments based on taxa recognition manually performed by experts. The practical target was to speed up taxa recognition by the methods of machine learning. For that purpose, we developed a software and technical prototype that allows for multiple images per specimen. Besides multiple images, the objects have hierarchical levels specifically created mimicking the classification strategy performed by human taxonomic experts. In that case, we constructed a top-down approach, that is, hierarchical classifiers built on support vector machines (SVM) and deep convolutional neural networks (CNN). The results were compared with typical flat classifiers and human experts using actual specimens. The lowest classification error 6.1% was obtained by human experts, the second lowest 11.4% and the third lowest 13.8% by a flat CNN and a hierarchical CNN, respectively. Contrary to previous findings in the literature, we found that the flat classification approach commonly used in machine learning performs better than the hierarchical approach also called as a local per parent node approach. Moreover, we shared our unique dataset to serve as a public benchmark dataset in this field.

E0544: **Statistical post-processing of visibility ensemble forecasts**

Presenter: **Maria Nagy-Lakatos**, University of Debrecen, Hungary

Co-authors: Sandor Baran

Accurate and reliable forecasting of visibility is vital in aviation meteorology and has great importance in shipping and road transportation as well. Recently, major meteorological services issue ensemble forecasts of visibility; however, these forecasts are often uncalibrated and have far lower forecast skill than ensemble forecasts of other weather quantities. Hence, some form of statistical post-processing is required to improve predictive performance. According to the suggestions of the World Meteorological Organization, visibility observations are reported in discrete values. Thus, calibration can be considered as a classification problem. Based on visibility ensemble forecasts of the European Centre for Medium-Range Weather Forecasts for Central-Europe for calendar years 2020-2021 and corresponding observations, we investigate the predictive performance of proportional odds logistic regression and multilayer perceptron neural network, which approaches provide the best forecast skill in a similar problem of calibrating total cloud cover ensemble forecasts. We show that compared with the raw ensemble forecasts, post-processing results in more than 20% improvement in forecast skill, and clustering the observation stations based on station climatology is superior to regional modelling.

E1884: **Autoregressive neural networks for predicting temperature, relative humidity, and weight of beehives**

Presenter: **Maria del Carmen Robustillo Carmona**, Universidad de Extremadura, Spain

Co-authors: Carlos Javier Perez Sanchez, M Isabel Parra Arevalo, Lizbeth Naranjo Albarran

Bee populations have been declining worldwide in recent decades, which is a serious problem for the environment since they help to maintain biodiversity through pollination. In this context, precision beekeeping emerges as a tool that allows greater control over the status of bees, helping the beekeeper to improve their care and maintenance. The objective is to predict internal temperature, relative humidity, and weight using sensor data of internal conditions and meteorological information. For this purpose, data obtained by the we4bee project in the Vohburg and Markt Indersdorf hives have been used. Two different scenarios have been considered: firstly, neural networks that use only climatological variables and, secondly, neural networks considering both climatological and internal variables. Predictions were made for one, three and seven days ahead. To validate this model, a rolling window 100 fold cross validation was performed, obtaining mean absolute errors lower than 0.750 Celsius degrees, 3.1%, and 210 g in the one-day predictions of temperature, humidity, and weight, respectively. Improvements ranging from 2.28% to 45.66% have been observed when comparing both scenarios, concluding that the incorporation of internal hive information is important. These results show the power of autoregressive neural networks to make predictions about hive conditions.

EC821 Room S-1.01 STATISTICAL MODELLING I Chair: Andriette Bekker

E0245: **Modeling multivariate circular-linear data in a biomechanical study**

Presenter: **Priyanka Nagar**, University of Pretoria, South Africa

Co-authors: Andriette Bekker, Mohammad Arashi, Cor-Jacques Kat, Annette-Christi Barnard

High-dimensional data containing circular and linear variables is common in biomechanical and orthopedic data. In most cases, the circular and linear variables are considered in isolation. The joint distribution modelling based on high-dimensional data containing circular and linear data is vital given the large amounts of directional data and the vast applications thereof. We propose a modelling framework applicable to the 6D

joint distribution of circular-linear data based on vine copulas. The pair-copula decomposition concept of vine copulas represents the dependence structure as a combination of circular-linear, circular-circular and linear-linear pairs modelled by their respective copulas. This allows us to assess the dependencies in the joint distribution. The motivation comes from the modelling of biomechanical data, i.e. the fracture displacements, that are used as a measure in external fixator comparisons. A case study based on the rotational and translational variables from an external fixator experiment illustrates the distribution's application.

E0377: **An extension of the quantile splicing technique for piece-wise distributions**
Presenter: **Brenda Macoduol**, University of Pretoria, South Africa

Quantile splicing was developed as a skewing mechanism for developing two-piece families of distributions with quantile functions of half distributions as the building block. The technique involves splicing quantile functions at the median point and introducing an asymmetry parameter to the half of the distribution whose domain is below the median point, while maintaining the kurtosis level of the parent distribution. This mechanism can be implemented for distributions defined primarily through their CDF, PDF, or quantile function. An extension of quantile splicing is proposed for generating piecewise distributions, where asymmetry is introduced to univariate distributions through splicing the quantile functions at location points other than the median ($0 < k < 1$). A general formula for the rth order L-moments is derived, which can be expressed in terms of the L-moments of the parent distribution and the expectations of the rth largest observation in a sample of size r from the kth piece distribution. The extension will consider the case where the quantile functions are spliced at the 25th percentile. The closed-form expressions for the families of distributions will be derived, as well as an application to data in which the method of L-moments estimation will be used.

E0549: **Distribution-free location-scale regression**
Presenter: **Sandra Siegfried**, University of Zurich, Switzerland
Co-authors: Lucas Kook, Torsten Hothorn

A generalized additive model for location, scale, and shape (GAMLSS) next of kin is introduced, aiming at distribution-free and parsimonious regression modelling for arbitrary outcomes. We replace the strict parametric distribution formulating such a model with a transformation function, which in turn is estimated from data. Doing so not only makes the model distribution-free but also allows limiting the number of linear or smooth model terms to a pair of location-scale predictor functions. The likelihood for continuous, discrete, and randomly censored observations, along with corresponding score functions, can be derived for these models. A plethora of existing algorithms is leveraged for model estimation, including constraint maximum-likelihood, the original GAMLSS algorithm, and transformation trees. Parameter interpretability in the resulting models is closely connected to model selection. We propose the application of a novel best subset selection procedure to achieve especially simple ways of interpretation. All techniques are motivated and illustrated by a collection of applications from different domains, including crossing and partial proportional hazards, complex count regression, nonlinear ordinal regression, growth curves, and receiver operating characteristics. The models can be estimated using the "tram" add-on package to the R system for statistical computing and graphics.

E1578: **A novel differentiable unification of least absolute deviations and least squares**
Presenter: **Kevin Burke**, University of Limerick, Ireland

Arguably, two of the most important error distributions are the normal and the Laplace distributions, respectively, being equivalent to classical least squares and least absolute deviations estimation. The key difference between these procedures is the use of a square function or an absolute value function within the objective function. Although least absolute deviations produce regression coefficients that are less impacted by outliers, their usage in applications has historically been much less widespread than least squares. This is likely due to the non-differentiable nature of the objective function, which requires more specialized treatment. However, we demonstrate that standard gradient-based optimization procedures can be applied if the absolute value function is replaced with a commonly used smooth approximation. Perhaps unexpectedly, we also show that this same procedure (designed for least absolute deviations approximation) can yield least squares estimates for particular values of its smoothing parameter. Ultimately, we develop a unified likelihood-based estimation procedure that can produce both the least absolute deviations and least squares estimates, as well as solutions between the two. Moreover, due to the equivalence with Laplace and normal distributions, we derive a new Laplace-normal-type distribution whose density function is differentiable.

E1793: **Modelling cooperation in a dynamic common pool resource game with deep reinforcement learning**
Presenter: **Simon Gero Haastert**, University of Munster, Germany
Co-authors: Matthias Hettich

The rapid degradation of natural common pool resources such as common fisheries, rain forests, or the global greenhouse gas budget is one of the greatest threats to human well-being. Such non-excludable, finite resources are exposed to socially inadequate appropriation strategies - a phenomenon often termed the tragedy of commons. However, game theory, as well as many lab experiments and real-life case studies, assert that cooperation towards the sustainable use of a common-pool resource is feasible. Certain conditions like reciprocal punishment or reward mechanisms, preferences such as inequity-aversion, or ways of communication can facilitate cooperation. We propose voting on an ad valorem tax rate as another instrument to enable multiple agents to coordinate their appropriation efforts. We model the common pool resource game with an agent-based model. Agents act individually in a partially observable Markov game in a trial-and-error fashion. They influence each other by reducing the common resource stock and voting on a tax rate. We train the agents via deep reinforcement learning with a state-of-the-art algorithm, which allows a continuous state and action space. In this setting, agents learn to indirectly punish agents for above-average use and over-exploitation of the common-pool resource by voting for a non-zero tax rate. We find that introducing the tax voting mechanism facilitates cooperative behavior and improves both resource sustainability and social welfare.

EC807 Room S-1.22 SURVIVAL ANALYSIS I Chair: Shaun Seaman

E0569: **Mixed-effects additive transformation models**
Presenter: **Balint Tamasi**, University of Zurich, Switzerland
Co-authors: Torsten Hothorn

Statistical models that accommodate non-normal, potentially correlated data and allow for nonlinear predictor-outcome relationships are crucial in applied regression settings. Traditional approaches typically rely on the idea that conditional response distribution can be fully captured with a few parameters of a predefined distribution type. Picking the correct distribution is often difficult in practice, and misspecifications can lead to incorrect inference. Transformation models provide a general and flexible approach to modeling the whole conditional distribution that forgoes the a priori specification of the response distribution by estimating its shape from the data. Mixed-effects additive transformation models extend the transformation model framework with random effects and penalized additive terms. The resulting model class can be readily used for modeling complex, dependent data structures (e.g., grouped data, temporal or spatial heterogeneity) in the presence of non-linear effects. Fully parametric likelihood-based estimation and inference of the model are discussed, and a fast and efficient R implementation is presented. The motivating example is an ecological experiment on carrion decomposition times under various environmental settings, where the non-normal, interval-censored time-to-event response, potential nonlinear covariate effects and grouped data structure render traditional regression tools inapplicable.

E0750: **Residual mean survival times in the presence of cure fractions in a two-sample problem**
Presenter: **Dennis Dobler**, Vrije Universiteit Amsterdam, Netherlands
Co-authors: Eni Musta

Many serious diseases do not end fatally for the patient. In such cases, the classical Kaplan-Meier estimator or Cox models might not be the best

approach to assess survival chances. Instead, cure models incorporate the fact that a fraction of the patients will never experience the event of interest. Our research is driven by the question: if the cure fraction is similar for two available treatments, how else can we determine which is preferable? To this end, we estimate the residual mean survival times in the uncured fractions of both treatment groups and develop permutation tests for inference. These are based on non- or semiparametric approaches. The methods are illustrated with medical data of cancer patients.

E0571: **Nonparametric survival estimation with missing not at random censoring indicators**
Presenter: **Mikael Escobar-Bach**, University of Angers, France
Co-authors: Olivier Goudet

In the presence of right-censored data with random covariates, the conditional Kaplan-Meier estimator (also referred to as the Beran estimator) consistently estimates the conditional survival function. However, it relies on the knowledge of each individual censoring status, which might be missing in practice. We thus show a study for the Beran estimator when the censoring indicators are not clearly specified, and next, propose a new method for the conditional survival function estimation with missing not at random (MNAR) censoring indicators. Along with the theoretical results, we illustrate how the estimators work for small samples by means of a simulation study and show their practical applicability with the analysis of synthetic data.

E1797: **Modelling the association structure in clustered right-censored survival data by factor copulas**
Presenter: **Roel Braekers**, Hasselt University, Belgium

In clustered right-censored survival data, frailty and copula models are commonly used to describe the association between different lifetimes within the same cluster. We present a copula model based on factor copula functions to model the structure between the different event times. This new methodology allows the intracluster dependence to be flexibly modeled by any parametric family of bivariate copulas. In this way, we encompass a wide range of dependence structures. We establish three estimation procedures in this model: a one- and two-stage parametric method and a two-stage semiparametric method where marginal survival functions are estimated using a Cox proportional hazards model. For the parameter estimators, we prove that they are consistent and asymptotically normally distributed, and assess their finite sample behavior with simulation studies. Furthermore, we illustrate the proposed methods on a data set containing the time to the first insemination after calving in dairy cattle clustered in herds of different sizes.

E1760: **Variable selection in the power-generalized Weibull distributional regression model**
Presenter: **Laura McQuaid**, University of Limerick, Ireland
Co-authors: Shirin Moghaddam, Kevin Burke

In medical applications, non-proportional hazards are often encountered. In these scenarios, standard survival modeling techniques are not appropriate. Instead, we propose using distributional regression where covariates enter the hazard function via multiple distributional parameters (e.g., scale and shape) simultaneously. This allows for more complex covariate effects, such as time-varying hazard ratios, to be captured. We develop the adapted power-generalized Weibull (APGW) distributional regression model, which, with three parameters (one scale, two shapes), encompasses various common survival models (Weibull, log-logistic, Gompertz) and hazard shapes (constant, increasing, decreasing, up then down, down then up), making it a highly flexible model. Variable selection is challenging in this setting (and distributional regression more generally) since covariates can enter the model in various ways. Thus, we propose the use of a computationally feasible adaptive lasso penalized estimation procedure for variable selection and explore its performance using numerical studies and real-world data application.

EC816 Room K0.16 NETWORK DATA Chair: Yumou Qiu

E0239: **Network regression and supervised centrality estimation**
Presenter: **Junhui Jeffrey Cai**, University of Notre Dame, United States
Co-authors: Dan Yang, Wu Zhu, Haipeng Shen, Linda Zhao

The centrality in a network is a popular metric for agents' network positions and is often used in regression models to model the network effect on an outcome variable of interest. In empirical studies, researchers often adopt a two-stage procedure to estimate the centrality and then infer the network effect using the estimated centrality. Despite its prevalent adoption, this two-stage procedure lacks theoretical backing and can fail in both estimation and inference. We, therefore, propose a unified framework, under which we prove the shortcomings of the two-stage in centrality estimation and the undesirable consequences in the regression. We then propose a novel supervised network centrality estimation (SuperCENT) methodology that simultaneously yields superior estimations of the centrality and the network effect and provides valid and narrower confidence intervals than those from the two-stage. We showcase the superiority of SuperCENT in predicting the currency risk premium based on the global trade network.

E0629: **Linear hotspot detection for a point pattern in the vicinity of a linear network**
Presenter: **Inger Fabris-Rotelli**, University of Pretoria, South Africa
Co-authors: Jacob Modiba, Alfred Stein, Gregory Breetzke

The analysis of point patterns on linear networks is receiving current attention in spatial statistics. This refers to the analysis of points in a spatial domain that coincides with a linear network like a road network. The linear network is modelled as a set of lines that are connected at their ends or are intersecting, that is, modelled as mathematical graphs. Limited research so far has been conducted on spatial points that fall on the Euclidean space containing the linear network. New steps are addressed by exploring points in the vicinity of the network that do not necessarily fall on the linear network. We present a novel method that is motivated by crime locations amongst a road network. The aim is to detect spatial hotspots around a linear network, where crime locations are considered as a point pattern lying in the vicinity of the linear road network. A new connectivity measure is also introduced to define the line segment neighbours of a line segment. The methodology is applied to crime data in Khayelitsha, South Africa. We detect a pattern of crime locations within the network that can be well interpreted. We conclude that our method is well applicable and could potentially help governmental organisations to allocate measures to reduce criminality.

E1541: **Identifying relations between summary statistics and parameters in ABC: A network approach**
Presenter: **Yangqi Zhang**, The University of New South Wales, Australia
Co-authors: Valentyn Panchenko

Likelihood-free methods such as approximate Bayesian computation (ABC) are popular tools to perform inference for complex models by simulation when the likelihood is intractable. Summary statistics are used for parameter inference in ABC. The choice of summary statistics often relies on prior knowledge or intuition. A novel network approach is proposed to identify and visualise the relations between the summary statistics and parameters of interest in ABC. After a pilot ABC process, pairwise partial correlations among summary statistics and parameters are computed and visualised as a weighted network of parameters and summary statistics. The resulting network is then pruned using network filtering and communication detection techniques. Such a network can improve the overall ABC performance, especially in high-dimensional settings. The authors also discuss the relationship between the network approach and other summary statistics selection techniques in ABC literature. Furthermore, the network provides information on decomposing complex models into several coupled modules. The decomposition is useful to the "cutting feedback" method, which provides robust inference when the model is misspecified. The performance of the network method is demonstrated in several simulated and real examples.

E1766: **Local assortativity in weighted and directed complex networks**
Presenter: **Marc Sabek**, University of Wuppertal, Germany
Co-authors: Uta Pigorsch

Assortativity measures the tendency of a vertex to bond with another based on similarity. It is commonly defined as the correlation coefficient between the excess degrees of both ends of an edge and is often associated with the robustness of a network against exogenous shocks. In this context, it is interesting to know which of the vertices or edges of a network are the most endangering on the one hand, and which are the most protective on the other hand. The assortativity coefficient, being a global measure, however, cannot provide answers to those kinds of questions. There is a need for a local assortativity measure, which can be either vertex or edge-based, in order to identify those vertices or edges that contribute most to the global assortativity structure of a network, respectively. Many real-world networks are weighted networks; however, local assortativity has been exclusively considered for unweighted networks, so far. By generalizing this concept to weighted and (un)directed networks, we unify two approaches used in the literature, and derive distinct measures that allow us to determine the assortativeness of individual edges and vertices as well as of entire components of a weighted network. We demonstrate the usefulness of our measures by applying them to theoretical and real-world networks. Along this way, we also explain how to compute local assortativity profiles, which are informative about the pattern of local assortativity with respect to edge weight.

E1018: **Supervariants identification for brain connectivity**
Presenter: **Ting Li**, Hong Kong Polytechnic University, Hong Kong

Identifying genetic biomarkers for brain connectivity helps us understand genetic effects on brain function. The unique and important challenge in detecting associations between brain connectivity and genetic variants is that the phenotype is a matrix rather than a scalar. We study a new concept of super-variant for genetic association detection. Similar to but different from the classic concept of gene, a super-variant is a combination of alleles in multiple loci, but contributing loci can be anywhere in the genome. We hypothesize that the super-variants are easier to detect and more reliable to reproduce in their associations with brain connectivity. By applying a novel ranking and aggregation method to the UK Biobank databases, we discovered and verified several replicable super-variants. Specifically, we investigate a discovery set with 16,421 subjects and a verification set with 2,882 subjects, where they are formed according to to release date, and the verification set is used to validate the genetic associations from the discovery phase. We identified 12 replicable super-variants on Chromosomes 1, 3, 7, 8, 9, 10, 12, 15, 16, 18, and 19. These verified super-variants contain single nucleotide polymorphisms that locate in 14 genes which have been reported to have an association with brain structure and function, and/or neurodevelopmental and neurodegenerative disorders in the literature.

EC754 Room K0.18 TIME SERIES Chair: Tommaso Proietti

E0340: **Statistical inference in binomial time series when the number of trials is small**
Presenter: **Takis Besbeas**, Athens University of Economics and Business, Greece
Co-authors: Fiori Labrinakou

Many authors have studied the problem of estimating the parameters in Bernoulli trials with Markov dependence. We consider the problem of analysing dependent count data arising from a series of binomial experiments, which includes dependent Bernoulli trials as a special case. We adopt a parameter-driven approach to modelling the data, involving a generalized linear model (GLM) for the binomial response based on an autoregressive AR(1) latent process to introduce autocorrelation. We focus on the case where the number of trials at time t, n_t, is small, and consider model-fitting by maximum likelihood using different estimation methods. Using theory and simulation, we show that the likelihood contains a ridge when $n_t = 1$ but the MLE is typically not located on the ridge. However, the estimation can be highly uncertain, even for large sample sizes. Further, we illustrate that estimation improves dramatically even for n_t as low as 2, which may provide a more pragmatic sampling alternative to obtaining a large binary sample in practice. An application to a rainfall example is given, allowing for different hypotheses on the probability of rain across years to be tested.

E0682: **Semiparametric estimation for time series: A frequency domain approach based on optimal transportation theory**
Presenter: **Manon Felix**, University of Geneva, Switzerland
Co-authors: Davide La Vecchia

A novel approach is proposed for estimation in stationary linear processes. Our estimation is semi-parametric: we have a Euclidean parameter, but we do not assume any distribution for the innovation term. Working with the frequency domain approach, we use the Wasserstein distance (1-Wasserstein distance and 2-Wasserstein distance) to derive minimum distance estimators. To do this, we rely on the fact that the standardized periodogram ordinates are asymptotically independent and have an exponential distribution with rate one. We give heuristic arguments for their asymptotics and provide algorithms for their implementation. Monte-Carlo simulations illustrate the performance of our estimators, under different data-generating mechanisms (e.g. leptokurtic underlying distributions, time domain additive outliers and frequency domain outliers). The numerical exercises highlight the improvements of our new estimators on the routinely-applied Whittles estimator.

E0335: **A Bayesian hierarchical time series model for estimating sex ratios in youth mortality**
Presenter: **Fengqing Chao**, King Abdullah University of Science and Technology, Saudi Arabia
Co-authors: Bruno Masquelier, Haavard Rue, Hernando Ombao, Leontine Alkema

Producing accurate estimates of sex ratios of mortality rates is essential in understanding population structure and dynamics, and in revealing sex discrimination. We introduce a Bayesian hierarchical model to estimate the disparity in age-specific mortality by sex for all countries over time, focusing on youth (15-24) mortality. The Bayesian model synthesizes data with varying levels, trends and associated uncertainties. The hierarchical modeling structure allows information exchange between data-saturated country periods and data-poor ones to assist estimation in country-periods lacking observations. Within the age groups 15-19 and 20-24, we model the global expected sex ratio using all the observations with a random walk model of order 2 (RW2). The RW2 is flexible to capture the non-linear global trend and is computationally efficient relative to splines model. We demonstrate the model-building process and motivate the choices of functions for the model elements. The model can be used to estimate sex disparity in mortality for other age groups, and we provide some illustrative results for modeling sex differences in 15-19 mortality. We conclude that the Bayesian model is an efficient and robust approach for estimating sex ratios of mortality.

E1043: **Semi-parametric estimation for the quantile coherence in multivariate time series**
Presenter: **Cristian Felipe Jimenez Varon**, King Abdullah University of Science and Technology, Saudi Arabia
Co-authors: Ta-Hsin Li, Ying Sun

In multivariate time series analysis, coherence measures the linear dependency between two-time series at different frequencies; however, real data applications often exhibit nonlinear dependency in the frequency domain. Conventional coherence analysis fails to capture such dependency among the time series. Quantile coherence, on the other hand, characterizes the nonlinear dependency by defining the coherence at a set of quantile levels. Although quantile coherence is a more powerful tool, its estimation remains challenging due to the high level of noise. A new estimation technique is proposed for quantile coherence. The proposed method is semiparametric, which uses the parametric form of the spectrum of the vector autoregressive model (VAR) as an approximation to the quantile spectral matrix, along with nonparametric smoothing across both quantile levels and frequencies. First, the method approximates the matrix-valued quantile partial autocorrelation function (QPACF) as well as the quantile autocovariance function (QACF) with the multivariate version of the Durbin-Levinson algorithm. Then, the QPACF is smoothed across quantile

levels by a nonparametric smoother. Finally, quantile coherence is computed from the smoothed QPACF. Numerical results show that the proposed estimation method outperforms other conventional nonparametric methods.

E2031: **Modeling and simulating dependence in networks using topological data analysis**
Presenter: **Anass El Yaagoubi Bourakna**, King Abdullah University of Science and Technology, Saudi Arabia
Co-authors: Moo K Chung, Hernando Ombao

Topological data analysis (TDA) approaches are becoming increasingly popular for studying the dependence patterns in multivariate time series data. In particular, various dependence patterns in brain networks may be linked to specific tasks and cognitive processes, which can be altered by various neurological and cognitive impairments such as Alzheimer's and Parkinson's diseases, as well as attention deficit hyperactivity disorder (ADHD). Because there is no ground-truth with known dependence patterns in real brain signals, testing new TDA methods on multivariate time series is still a challenge. Simulations are crucial for evaluating the performance of proposed TDA methods and testing procedures, as well as for creating computation-based confidence intervals. To our knowledge, there are no methods that simulate multivariate time series data with specific and manually imposed connectivity patterns. We present a novel approach to simulate multivariate time series with a specific number of cycles/holes in its dependence network. Furthermore, we also provide a procedure for generating higher dimensional topological features.

EC809 Room S0.03 SPATIAL STATISTICS Chair: Christopher Hans

E1689: **Spatial clustering of time series using Bayesian quantile regression**
Presenter: **Paolo Girardi**, Ca' Foscari University, Italy
Co-authors: Victor Muthama Musau, Carlo Gaetan

A large research literature has developed methodologies for identifying clusters of units in a spatial setting. When we deal with longitudinal or temporal data, the proposed methods are mainly based on mean regression. However, the resulting classification could not be robust in presence of outliers, skewed distributions and/or heteroscedasticity. Furthermore, the researcher might be interested in classifying units according to whether certain thresholds are exceeded. We propose a model-based approach for clustering spatial units, that is based on median or, more in general, quantile regression. In this way, we want to cope better with the aforementioned issues. The spatial units are supposed to belong to a network. The model specification is hierarchical that allows a Bayesian inference based on Markov chain Monte Carlo methods. As an illustration and motivating example, we consider data on the sea surface temperature (SST) of the Mediterranean Sea. The dataset is a result of a model re-analysis that provides 251 time series of temperature in 1-degree gridded data covering the temporal window from 1982 to 2012. Specifying a quantile of interest (e.g. in ecology, is 0.9), we aim to identify areas with similar trends and cyclic patterns.

E1716: **Spatio-temporal Bayesian methods for historical data**
Presenter: **Tiia-Maria Pasanen**, University of Jyvaskyla, Finland

A Bayesian spatio-temporal model is developed to study pre-industrial grain market integration during the Finnish famine of the 1860s. The data consists of 80 regional time series covering nine years of monthly grain prices. The model takes into account several problematic features often present when analysing such spatially interdependent time series. For example, compared with the error correction methodology commonly applied in econometrics, this approach allows simultaneous modelling of multiple interdependent time series, avoiding cumbersome statistical testing needed to predetermine the (often artificial) market leader as a point of reference. Furthermore, introducing a flexible spatio-temporal structure enables analysing of detailed regional and temporal dynamics of the market mechanisms, for example, the asymmetric neighbour dependencies. The whole process, from deriving the model to interpreting the results, is covered with the famine application from the point of view of a statistician.

E1841: **Robust spatial blind source separation**
Presenter: **Sara Taskinen**, University of Jyvaskyla, Finland
Co-authors: Mika Sipila, Klaus Nordhausen

Assume a spatial blind source separation model in which the observed multivariate spatial data is assumed to be a linear mixture of latent stationary spatially uncorrelated random fields. The goal is then to recover an unknown mixing procedure as well as latent uncorrelated random fields. Recently, spatial blind source separation methods that are based on simultaneous diagonalization of two or more scatter matrices were proposed. In case of uncontaminated data such methods are capable of solving the blind source separation problem, but in presence of outlying observations the methods perform poorly. We propose a robust blind source separation method which uses robust global and local scatter matrices based on generalized spatial signs in simultaneous diagonalization. Simulation studies are used to illustrate robustness and efficiency properties of proposed methods in various scenarios.

E1776: **A multi-resolution approximation by linear projection and covariance tapering for large spatial datasets**
Presenter: **Toshihiro Hirano**, Kanto Gakuin University, Japan

Estimation and prediction for large spatial datasets, such as maximum likelihood estimation and kriging, are impractically time-consuming. A multi-resolution approximation via linear projection has been developed to deal with this computational burden. However, this method can cause the partly mismatched fitting around the origin for the covariance function if the two locations are not in the same subregion at the low resolution. Additionally, in this case, there is a possibility of producing artificiality in the predictive surface. To solve this problem, we propose an algorithm that approximates the covariance function by iteratively applying the covariance tapering instead of dividing the region at each resolution. We also elicit fast computation algorithms for estimation and prediction by using the approximated covariance function. A real data analysis for air dose rates demonstrates that our proposed method works well and avoids artificiality in the predictive surface.

E1996: **Non-normal estimation of multiple spatial data using multivariate skews normal process**
Presenter: **Kassahun Abere Ayalew**, Centers for Disease Control and Preventions, South Africa
Co-authors: Samuel Manda, Bo Cai

A Multivariate Gaussian Intrinsic Conditional Autoregressive (MICAR-normal) model is used in joint spatial modeling. However, the modelled multivariate data could be highly tailed and skewed. We present a multivariate skew-normal Intrinsic Conditional Autoregressive model (MICAR-skew-normal) to capture the non-normal distribution of the spatial data. We show how to obtain estimates of the model parameters using a fully Bayesian analysis using a stochastic approximation of the EM algorithm (SAEM). Using extensive simulation studies, we demonstrate the capabilities of the proposed model and its usefulness with an analysis of HIV data from South Africa.

EC822 Room S0.12 MULTIVARIATE STATISTICS AND COMPLEX DATA Chair: Michelle Carey

E0262: **Geometric goodness of fit measure to detect patterns in data point clouds**
Presenter: **Maikol Solis**, Universidad de Costa Rica, Costa Rica
Co-authors: Alberto Hernandez

A geometric goodness-of-fit index similar to R^2 is derived using geometric data analysis techniques. We build the alpha shape complex from the data-cloud projected onto each variable and estimate the area of the complex and its domain. We create an index that measures the difference in area between the alpha shape and the smallest squared window of observation containing the data. By applying ideas similar to those found in the closest neighbor distribution and empty space distribution functions, we can establish when the characterizing geometric features of the point set

emerge. This allows for a more precise application for our index. We provide some examples with anomalous patterns to show how our algorithm performs. We also present the R package spatgeom, which performs the estimation of the space-filling distribution via alpha-shape reconstructions.

E0442: **Individual claims reserving using activation patterns**
Presenter: **Marie Michaelides**, Universita du Quabec a Montraal, Canada

The occurrence of a claim often impacts not one but multiple insurance coverages provided in the contract. To account for this multivariate feature, we propose a new individual claims reserving model built around the activation of the different coverages to predict the reserve amounts. Using the framework of multinomial logistic regression, we model the activation of the different insurance coverages for each claim and their development in the following years, i.e. the activation of other coverages in the later years and all the possible payments that might result from them. As such, the model allows us to complete the individual development of the open claims in the portfolio. Using a recent automobile dataset from a major Canadian insurance company, we demonstrate that this approach generates accurate predictions of the total reserves as well as of the reserves per insurance coverage. This allows the insurer to get better insights into the dynamics of his claims reserves.

E1740: **High dimensional multivariate Bernoulli distributions with identical margins**
Presenter: **Roberto Fontana**, Politecnico di Torino, Italy
Co-authors: Patrizia Semeraro

The main contributions are algorithms to sample from multivariate Bernoulli distributions and to determine the distributions and the bounds of a wide class of indices and measures of probability mass functions. Probability mass functions of *exchangeable* Bernoulli distributions are points in a convex polytope, and we provide an analytical expression for the extremal points of this polytope. The more general class of multivariate Bernoulli distributions with identical marginal Bernoulli distributions with parameter p is also a convex polytope but finding its extremal points is a more challenging task. Our novel theoretical contribution is to use an algebraic approach to find a set of analytically available generators. We also solve the problem of finding the lower bound in the convex order of multivariate Bernoulli distributions with given margins, but with unspecified dependence structure.

E1852: **Multivariate statistical quality control for autocorrelated data**
Presenter: **Ioulia Papageorgiou**, Athens University of Economics and Business, Greece
Co-authors: Stefanos Voutsinas

With new technologies introduced in monitoring industrial processes, such as automatic sensor technology and sophisticated software, the data collected to be used for statistical analysis and control of a process, are most often multivariate today. Moreover, the time between observations can be very short. These two factors result in multivariate autocorrelated data. Standard methods for statistical quality control can be problematic or inferior with respect to their performance in such applications, and false conclusions may be drawn. One of the leading approaches to treat the existence of autocorrelation in the univariate case is the model-based, where a time series model is fitted to the data first, and all standard techniques are then implemented to the residuals instead of initial data. Attempts to extend these methodologies to the multivariate case have been reported to present both difficulties in practical use and questionable results in efficiency. We present two model-free approaches proposed to cope with autocorrelation and assist in monitoring a process.

E2032: **Decomposing systemic risk and the roles of contagion and common exposures: A structural approach**
Presenter: **Ruben Hipp**, Bank of Canada, Canada
Co-authors: Grzegorz Halaj

A derivative function is used to estimate contagion based on granular and confidential data. This function allows for the endogeneity of the banks capital via value functions of assets and liabilities. With a structural regression similar to that in the structural VAR literature, we infer how the success and risk of one bank have been transmitted to other banks. We accomplish this by allowing banks to be correlated solely via contagion – from direct exposures, fire sales, and market-based sentiment – or via common exposures – from portfolio overlaps. We apply this model to confidential network data of the Canadian banking market, which consists of six large banks and five smaller ones. We discover that contagion has fluctuated over time, with the highest levels around the Great Financial Crisis (GFC) in 2008 and lower levels for the pandemic periods. Finally, we decompose contagion and common exposures by channels. We find that some channels' strength has notably changed since the GFC, hinting that interbank contagion has structurally changed.

EC812 Room K2.40 DIMENSION REDUCTION AND SHRINKAGE METHODS Chair: Onno Kleen

E0332: **A comparative study of linear and nonlinear sufficient dimension reduction**
Presenter: **Jiyeong Kang**, Ewha Womans University, Korea, South
Co-authors: Kyongwon Kim

dr package is a widely used tool to implement linear sufficient dimension reduction, which is useful to extract core information from a high-dimensional dataset. However, because big data can include a complicated nonlinear structure, some features of the dataset cannot be fully explained by a linear sufficient dimension reduction. The nonlinear sufficient dimension reduction can be an alternative to address this issue. However, the theoretical formulation of nonlinear sufficient dimension reduction relies on the linear operators in Hilbert space, and this hampers many users from applying nonlinear sufficient dimension reduction methods in real data analysis. We compare the theoretical background and numerical results between linear and nonlinear sufficient dimension reduction using a widely used dr package and a recently developed nsdr package. We further present nonlinear sufficient dimension reduction methods can be applied to a classification problem by using the wine cultivar dataset.

E0369: **On cross-distance selection algorithm for hybrid sufficient dimension reduction**
Presenter: **Yujin Park**, Ewha Womans University, Korea, South
Co-authors: Kyongwon Kim, Jae Keun Yoo

Given the extensive development of a variety of sufficient dimension reduction (SDR) methodologies, a hybrid SDR method was proposed combining two pre-existing SDR methods. In particular, a bootstrap approach was used to select a proper weight. Since bootstrapping is computationally intensive and time-consuming, the hybrid reduction approach has not been widely used, although it is more accurate than conventional single SDR methods. To overcome these deficits, we propose a novel cross-distance selection algorithm. Similar to the bootstrapping method, the proposed selection algorithm is data-driven and has a strong rationale for its performance. The numerical studies demonstrate that the chosen hybrid method from our proposed algorithm offers a good estimation quality and reduces the computing time dramatically at the same time. Furthermore, our real data analysis confirms that the proposed selection algorithm has potential advantages with its practical usefulness over the existing bootstrapping method.

E0534: **Comparing Grassmann manifold optimization and sequential candidate set algorithm in a principal fitted component model**
Presenter: **Chaeyoung Lee**, Ewha Womans University, Korea, South
Co-authors: Jae Keun Yoo

The parameter estimation by Grassmann manifold optimization and the sequential candidate set algorithm are compared in a structured principal-fitted component model. The PFC model is a model-based dimension reduction method which can be divided into three distinct variations according to the forms of the covariance matrix of a random error. The structured PFC model, which is our main focus, has an extended form of the simplest

covariance matrix. The extension relieves the limit that occurs due to the simple form of the covariance matrix of a random error. However, the structured PFC model does not have a closed form for parameter estimation in dimension reduction. Therefore, the estimation needs to be done numerically. The computation can be done through Grassmann manifold optimization and sequential candidate set algorithm. We conducted several numerical simulations for comparison. First, we compared the determined dimension obtained from the sequential dimension. In addition, we calculated trace correlation values to compare the accuracy of the estimated basis to conduct dimension reduction. From the simulation results, we could conclude that while Grassmann manifold optimization outperforms the sequential candidate set algorithm in dimension determination, the sequential candidate set algorithm is better in basis estimation for dimension reduction. In other words, there is no optimal method that shows great performance in both aspects.

E0322: **Intensive comparison of semi-parametric and non-parametric dimension reduction methods in forward regression**
Presenter: **Minju Shin**, Ewha Womans University, Korea, South
Co-authors: Jae Keun Yoo

Principal Fitted Component (PFC) is a semi-parametric sufficient dimension reduction (SDR) method. The PFC has a connection with other usual non-parametric SDR methods. The connection is limited to sliced inverse regression and ordinary least squares. Since there is no direct comparison between the two approaches in various forward regressions up to date, practical guidance between the two approaches is necessary for usual statistical practitioners. To fill this practical necessity, we newly derive a connection of the PFC to covariance methods (Yin and Cook, 2002), which is one of the most popular SDR methods. Also, intensive numerical studies have been done closely to examine and compare the estimation performances of the semi- and non-parametric SDR methods for various forward regressions. The founding from the numerical studies is confirmed in a real data example.

E2028: **Projection inference for high-dimensional covariance matrices with structured shrinkage targets**
Presenter: **Fabian Mies**, RWTH Aachen University, Germany
Co-authors: Ansgar Steland

Analyzing large samples of high-dimensional data under dependence is a challenging statistical problem as long time series may have change points. Inference for large covariance matrices is especially difficult due to noise accumulation, resulting in singular estimates and poor power of related tests. The singularity of the sample covariance matrix can be overcome by a linear combination with a structured target matrix, typically of diagonal form. We consider covariance shrinkage towards structured nonparametric estimators of the bandable or Toeplitz type, respectively, aiming at improved estimation accuracy and statistical power of tests even under nonstationarity. We derive feasible Gaussian approximation results for bilinear projections of the shrinkage estimators, which are valid under nonstationarity and dependence. These approximations enable us to formulate a statistical test for structural breaks in the marginal covariance structure of high-dimensional time series without restrictions on the dimension, and which is robust against nonstationarity of nuisance parameters. We show via simulations that shrinkage helps to increase the power of the proposed tests. Moreover, we suggest a data-driven choice of the shrinkage weights, and assess its performance by means of a Monte Carlo study. The results indicate that the proposed shrinkage estimator is superior for non-Toeplitz covariance structures close to fractional Gaussian noise.

EC815 Room K2.41 NEUROIMAGING Chair: Russell Shinohara

E1555: **Cross-validation for high-dimensional testing in imaging genetics**
Presenter: **Iris Ivy Gauran**, King Abdullah University of Science and Technology, Saudi Arabia
Co-authors: Hernando Ombao, Zhaoxia Yu

In estimating the accuracy of a predictive rule, it is essential to understand both the quality of predictions and model selection. Cross-validation is an algorithmic technique extensively used for estimating the prediction error, tuning the regularization parameter, and choosing between competing predictive rules. However, its behavior is non-trivial because of various, complex factors at play. We investigate the performance of cross-validation as a test statistic for high-dimensional testing. We propose a nested cross-validation procedure that performs regularization parameter tuning and accurate variance estimation to formulate the statistic and perform the test rigorously. We present our findings on strategies for improving the statistical power in the high-dimensional as well as the low and equidimensional scenarios. The application of the proposed method to an imaging genetics study and biological data is also presented.

E1563: **Hypothesis testing and spatial localization of associations in multi-modal neuroimaging studies**
Presenter: **Sarah Weinstein**, University of Pennsylvania, United States
Co-authors: Russell Shinohara, Jun Young Park

Many neuroimaging-based studies of neurodevelopment and mental health involve collecting and integrating measures of both brain structure and function. In such studies, we are often interested in conducting hypothesis tests of associations between these different modalities and evaluating whether these associations differ across subgroups. Until recently, statistical methods in this area have primarily evaluated global associations–for example, testing whether associations exist throughout the entire brain or within pre-defined anatomical subregions or functional networks. We propose a new method for spatial localization of inter-modal associations. We first adjust for the underlying spatial autocorrelation structure in each imaging modality to address heterogeneity between modalities. Second, we use clusterwise inference to leverage spatial information and construct an interpretable map of spatially enhanced test statistics. Finally, we use permutation for inference to ensure type I error and family-wise error rate control. Through simulation studies using multi-modal neuroimaging data from the Philadelphia Neurodevelopmental Cohort, we illustrate our methods' statistical power, interpretability, and ability to replicate findings even in small-sample settings.

E1608: **Normative brain mapping of 3-dimensional morphometry imaging data using skewed functional data analysis**
Presenter: **Marco Palma**, University of Cambridge, United Kingdom
Co-authors: Shahin Tavakoli, Julia Brettschneider, Thomas Nichols, Ana-Maria Staicu

Tensor-based morphometry (TBM) aims at showing local differences in brain volumes with respect to a common template. TBM images are smooth, but they exhibit (especially in diseased groups) higher values in some brain regions called lateral ventricles. More specifically, a voxelwise analysis shows a mean-variance relationship in these areas and evidence of spatially dependent skewness, which can be missed in the standard functional data analysis (FDA) settings, which focus only on the first two functional moments. We propose a model for 3-dimensional functional data where mean, variance, and shape functions vary smoothly across brain locations. We model the voxelwise distributions as skew-normal. The smooth effects of age and sex are estimated on a reference population of cognitively normal subjects from the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset and mapped across the whole brain. The parameter functions allow the transformation of each TBM image (in the reference population as well as in a test set) into a Gaussian process. These subject-specific normative maps are used to derive indices of deviation from a healthy condition which could help to assess the individual risk of pathological degeneration or to cluster different disease groups.

E1730: **Club Exco: Clustering brain extreme communities from multi-channel EEG data**
Presenter: **Matheus Guerrero**, King Abdullah University of Science and Technology, Saudi Arabia
Co-authors: Hernando Ombao, Raphael Huser

Current methods for clustering nodes in a brain network are determined by cross-dependence measures, which are computed from the entire range of values of the EEG signals. One limitation of these measures is that they do not distinguish whether the signals are dependent (or synchronous)

only at large amplitudes or over the entire range of values. We develop the Club Exco method for clustering brain-extreme communities to overcome these shortcomings. Club Exco uses a spherical k-means procedure applied to the "pseudo-angles", derived from extreme amplitudes of EEGs, to cluster multi-channel EEG data. With this approach, a cluster center is considered an extremal prototype, revealing nodes sharing the same extreme behavior (i.e., large amplitudes of the signal from one node are in tandem with large amplitudes of the others). Non-extreme-value techniques cannot identify this important feature. Club Exco serves as a tool to classify EEG channels into mutually asymptotically (in)dependent groups. It provides insights into how the brain network organizes itself during an extreme event (e.g., epileptic seizure) in contrast to normal states. We apply Club Exco to investigate the differences in EEG brain connectivity networks of a patient diagnosed with epilepsy, a chronic neurological disorder affecting more than 50 million people globally. Our method reveals a substantial difference in the organization of the alpha band in the brain network compared to coherence-based methods.

E1690: **Regression and alignment for functional data and network topology**
Presenter: **Danni Tu**, University of Pennsylvania, United States
Co-authors: Julia Wrobel, Azeez Adebimpe, Theodore Satterthwaite, Jeff Goldsmith, Jan Gertheiss, Danielle Bassett, Russell Shinohara

The human functional brain network dynamically reorganizes during adolescence. Changes in mesoscale topology can be assessed by modularity and participation coefficient, two diagnostics which capture the community structure of the brain network. By proportionally thresholding the network edges, we obtain a sequence of diagnostics for each threshold, resulting in diagnostic curves that describe network structure at multiple scales. Previous methods that evaluate network diagnostic curves have relied on permutation-based or pointwise comparisons, which are less powerful and less informative than comparisons of curves in their entirety. We propose a functional regression framework that addresses biases introduced by systematic differences in the distribution of edge strengths between networks, which we conceptualize as phase variation in diagnostic curves. Our novel method therefore simultaneously performs regression and curve alignment through an iterative, penalized estimation procedure. The illustrated procedure is widely applicable to domains of neuroscience where the goal is to study heterogeneity among a mixture of function- and scalar-valued measures.

CO645 Room Virtual R05 REGULARISATION IN ECONOMETRIC MODELS Chair: Ralf Wilke

C0683: **Testing the identification of causal effects in data**
Presenter: **Jannis Kueck**, University of Hamburg, Germany
Co-authors: Martin Huber

The aim is to demonstrate the existence of a testable condition for the identification of the causal effect of a treatment on an outcome in observational data, which relies on two sets of variables: observed covariates to be controlled for and a suspected instrument. Under a causal structure commonly found in empirical applications, the testable conditional independence of the suspected instrument and the outcome given the treatment and the covariates has two implications. First, the instrument is valid, i.e. it does not directly affect the outcome (other than through the treatment) and is unconfounded conditional on the covariates. Second, the treatment is unconfounded conditional on the covariates such that the treatment effect is identified. We suggest tests of this conditional independence based on machine learning methods that account for covariates in a data-driven way and investigate their asymptotic behavior and finite sample performance in a simulation study. We also apply our testing approach to evaluating the impact of fertility on female labor supply when using the sibling sex ratio of the first two children as supposed instrument, which by and large points to a violation of our testable implication for the moderate set of socio-economic covariates considered.

C0725: **Learning network with focally sparse structure**
Presenter: **Chen Huang**, Aarhus University, Denmark
Co-authors: Victor Chernozhukov, Weining Wang

Network connectedness with a focally sparse structure is studied. We uncover the network effect with a flexible sparse deviation from a predetermined adjacency matrix. More specifically, the sparse deviation structure can be regarded as latent or as misspecified linkages to be estimated. To obtain high-quality estimators for the parameters of interest, we propose using a debiased-regularized, high-dimensional generalized method of moments (GMM) framework. Moreover, this framework also enables us to conduct inference on the parameters. Theoretical results on consistency and asymptotic normality are provided, while accounting for general spatial and temporal dependency of the underlying data-generating processes. Simulations demonstrate a good performance of our proposed procedure. Finally, we apply the methodology to study the spatial network effect of stock returns.

C0962: **Estimation of group structures in panel models with individual fixed effects**
Presenter: **Enno Mammen**, Heidelberg University, Germany
Co-authors: Ralf Wilke, Kristina Maria Zapp

The fixed effects (FE) panel model is one of the main econometric tools for empirical economic research, despite the non-identifiability of the parameters on time-constant covariates. A new approach is presented to grouping FE in the linear panel model to reduce their dimensionality and ensure identifiability, therefore overcoming the main practical limitation of the linear FE model. By using unsupervised nonparametric density-based clustering, cluster patterns, including their location and number, are not restricted. The approach works with large data structures (units and groups) and only clusters units that are sufficiently similar, while leaving others as unclustered atoms. Asymptotic theory and rates of convergence are presented. With the help of simulations and an application to economic data it is shown that the suggested method performs well and gives more insightful and efficient results than conventional panel models.

C1051: **Inference after multiple hypothesis testing**
Presenter: **Andreas Dzemski**, University of Gothenburg, Sweden
Co-authors: Wenjie Wang, Ryo Okui

Some empirical studies estimate multiple treatment effects corresponding to different sub-populations, different treatments or different outcome variables and focus on interpreting the results of the significant specifications. We develop new estimators that are unbiased under this kind of data-driven model selection along with corresponding valid confidence intervals. Our framework admits a large class of rules for determining significance, including many step-down and step-up methods. In an empirical application, we compare our estimator to a naive estimator that does not account for the selection step. For effects that are significant but close to the threshold of insignificance, our approach detects large selection bias and produces estimates that are very different from the estimates obtained by a naive approach.

CO030 Room BH (S) 1.01 Lecture Theatre 1 CURRENT CHALLENGES TO MACRO AND FINANCIAL STABILITY Chair: Claudio Morana

C0360: **Does crime pay? The financial consequences of bank misconduct**
Presenter: **Jose Fernandez de Bilbao**, Universidad Pontificia de Comillas (ICADE), Spain
Co-authors: Isabel Figuerola-Ferretti, Alvaro Santos Moreno, Ioannis Paraskevopoulos

A new approach is proposed to the study of the impact of bank misconduct on bank profitability. Most papers addressing this topic have traditionally used the announcement of fines and penalties received as the key independent variable to determine bank misconduct. On this basis, they have reached contradictory conclusions, including that misconduct has no significant impact on the bank's after-tax profitability. We argue that P&L misconduct costs and provisions are a more appropriate variable for bank misconduct. In order to support our position, we show that bank accounting rules force banks to accrue for misconduct costs several periods before the announcement of a fine or penalty and that fines and

penalties do not fully capture the full breadth of misconduct costs incurred by banks. We apply our proposed variable to estimate a Generalized Method of Moments (GMM) regression model on a dynamic panel data sample of Global Systemically Important Banks. Our analysis shows that bank misconduct does have a significant and strong impact on both pre-tax and after-tax profitability and that misconduct costs do not have any significant effect on effective tax rates. These results have both regulatory and social implications as they provide evidence that banks do suffer negative financial consequences from their misbehaviour.

C0479: **When do investors go green: Evidence from a time-varying asset-pricing model**
Presenter: **Elisa Ossola**, Universita Milano - Bicocca, Italy
Co-authors: Lucia Alessi, Roberto Panzica

The focus is on the evolution of the greenium, i.e. a risk premium linked to firms' greenness and environmental transparency, based on individual stock returns. We estimate an asset pricing model with time-varying risk premia, where the greenium is associated with a priced 'greenness and transparency' factor, which considers both companies' greenhouse gas emissions and the quality of their environmental disclosures. We show that investors in the European equity market tend to accept lower returns, ceteris paribus, to hold greener and more transparent assets when the shift of the economy towards low-carbon becomes more credible. This happened after the Paris Agreement, the first Global Climate Strike and the announcement of the EU Green Deal. Signals going in the opposite direction, such as increasing fossil fuel prices and more bad news about climate change, are associated with increases in the greenium.

C0520: **The systemic risk of US oil and natural gas companies**
Presenter: **Roberto Panzica**, European Commission - Joint Research Centre, Italy
Co-authors: Massimiliano Caporin, Fulvio Fontini

The evolution of the systemic risk impact of oil and natural gas companies since 2000 is analyzed. This period is characterised by several events that affected energy source markets: the real effect of the global financial crisis, the explosion of shale production and the diffusion of the Covid-19 pandemic. The price of oil and natural gas showed extreme swings, impacting companies financial situations, which, accompanied by technological developments in shale production, had an impact on debt issuance and the overall risk level of the oil and natural gas sector. By studying the systemic impact of oil and natural gas companies on risk in the financial market, measured by the CoVaR, we observe that in the most recent decade, their role is sensibly increasing compared to 2000-2010, even accounting for the possible effect associated with the increase in companies sizes. In addition, our results show evidence of a decreasing relevance of traditional drivers of systemic risk, suggesting that additional factors might be present. Finally, when focusing on the impact of Covid-19, we document its relevant role in fueling the increase in the oil and natural gas companies' systemic impact.

C0328: **Getting in all the cracks: Monetary policy, financial vulnerabilities, and macro risk**
Presenter: **Andrea Ajello**, Board of Governors of the Federal Reserve System, United States

The effect of monetary policy on financial vulnerabilities is estimated, and the implications for macroeconomic tail risk are studied. We first extract a small set of common factors from a large dataset of financial vulnerability indicators, estimating a factor-augmented proxy SVAR to study the response of aggregate economic activity, inflation, and financial vulnerabilities to monetary policy shocks. We then estimate the effect of changes in the financial vulnerability factors on macroeconomic tail risk via quantile regressions. We find that an unexpected monetary policy tightening can lower asset valuation vulnerabilities in the short term and slow down credit growth in the medium term. As tighter monetary policy reduces asset valuation pressures, it does so at the cost of a sizable increase in macro tail risk in the short term that is only partially offset by a modest reduction in tail risk in the medium term, induced by a slowdown in credit growth.

C0849: **Euro area inflation and a new core inflation measure**
Presenter: **Claudio Morana**, Universita di Milano Bicocca, Italy

The driver of euro area inflation since its foundation in 1999 is investigated, focusing on current inflationary pressures. The analysis is carried out within an innovative modeling framework, where headline inflation is disentangled into a core or medium to long-term component, a cyclical component associated with business cycle developments, and a residual component related to other, short-lived factors. The proposed decomposition approach is straightforward and requires the sequential implementation of regression and principal components analysis. The proposed core inflation measure bears the interpretation of conditional expectation for headline inflation, where conditioning is made relative to monetary and supply-side drivers of underlying inflation. Also, cyclical inflation carries information on expected headline inflation, conditional to demand-side and (short-term) supply-side developments. Within this context, we are then able to track the evolution of euro area inflation since its inception.

CO358 Room BH (SE) 1.02 RECENT DEVELOPMENTS IN STRUCTURAL VARS Chair: Joshua Chan

C0481: **Advances in using vector autoregressions to estimate structural magnitudes**
Presenter: **Christiane Baumeister**, University of Notre Dame, United States

Recent advances are surveyed for drawing structural conclusions from vector autoregressions, providing a unified perspective on the role of prior knowledge. We describe the traditional approach to identification as a claim to have exact prior information about the structural model and propose Bayesian inference as a way to acknowledge that prior information is imperfect or subject to error. We raise concerns from both a frequentist and a Bayesian perspective about the way that results are typically reported for VARs that are set-identified using signs and other restrictions. We call attention to a common but previously unrecognized error in estimating structural elasticities and show how to correctly estimate elasticities even in the case when one only knows the effects of a single structural shock.

C0647: **Heteroskedastic proxy VARs: Testing for time-varying impulse responses in the presence of multiple proxies**
Presenter: **Martin Bruns**, University of East Anglia, United Kingdom
Co-authors: Helmut Luetkepohl

A test is proposed for time-varying impulse responses in heteroskedastic structural vector autoregressions that can be used when the shocks are identified by external proxy variables as a group. The test can be used even if the shocks are not identified individually. The asymptotic analysis is supported by small sample simulations, which show good properties of the test. An investigation of the impact of productivity shocks in a small macroeconomic model for the U.S. illustrates the importance of the issue for empirical work.

C1468: **Understanding instruments in macroeconomics: A study of high-frequency identification**
Presenter: **Pooyan Amir Ahmadi**, University of Illinois and Amazon, United States
Co-authors: Christian Matthes, Mu-Chun Wang

The effects of monetary policy shocks are regularly estimated using high-frequency surprises in asset prices around central bank meetings as an instrument. These studies assume a constant relationship between the instrument and the monetary policy shock. By allowing for time variation in this relationship, we show that only a few distinct periods are informative about monetary policy shocks. We thus build a narrative for instrument-based identification and sharpen results: for the instrument, the effect of monetary policy shocks on the (log) price level is almost 50 percent larger than the standard specification would suggest. This result can be obtained using only 10 percent of all available observations of the instrument.

C1631: **Learning based uncertainty quantification in set identified SVARs**
Presenter: **Tilmann Haertl**, University of Konstanz, Germany

Inference in set-identified structural vector autoregressions requires capturing both the uncertainty about the structural models in the identified set

as well as the estimation uncertainty regarding the reduced form parameters of the model. Learning the identified set based on a bootstrapped set of reduced form parameters enables uncertainty quantification regarding the structural parameters directly via the learning guarantees of the learning algorithm. This approach is silent with regard to the underlying identification strategy or combinations of identification strategies and is not computationally costly. Hence, it is also applicable for cases in which conventional inference is potentially not worked out yet or tedious and offers an alternative way to quantify the uncertainty inherent in the identification. Simulations for the structural impulse responses show that the learned sets yield informative value for the structural parameters.

C1416: **Large Bayesian VARs with many structural restrictions**
Presenter: **Joshua Chan**, Purdue University, United States
Co-authors: Christian Matthes, Xuewen Yu

Large Bayesian VARs are increasingly used in macroeconomic analysis. But identifying large models using many popular structural restrictions, such as sign restrictions, remains practically infeasible. We develop a new approach to estimate large Bayesian VARs identified using a large number of signs and other structural restrictions. The methodology is illustrated using a 35-variable VAR with sign and ranking restrictions to identify 8 structural shocks, namely, demand, investment, financial, monetary policy, government spending, technology, labor supply and wage bargaining.

CO266 Room BH (SE) 1.05 CREDIT RISK MODELLING Chair: Raffaella Calabrese

C0515: **Quantifying model uncertainty of machine learning methods for loss given default estimation**
Presenter: **Maximilian Nagl**, University of Regensburg, Germany
Co-authors: Matthias Nagl, Daniel Roesch

The use of machine learning methods has increasingly found its way into the credit risk literature. They focus mainly on a sounder prognosis of the main credit risk parameters and are shown to be superior to standard statistical models. However, the quantification of their accompanied (model) uncertainty is neglected so far. This type of uncertainty measures how certain the model is with each prediction. Therefore, it is imminent for risk managers and regulators, and its quantification increases the transparency and stability of machine learning methods to risk management tasks. We fill this gap by using a novel approach called evidential learning. We evaluate the model uncertainty for loss-given default estimation techniques and apply explainable artificial intelligence (XAI) methods to evaluate its drivers. We discover that model uncertainty increases out of time and with extreme realizations of macroeconomic drivers.

C0900: **Detecting local bias and error in credit risk models**
Presenter: **Anthony Bellotti**, University of Nottingham Ningbo China, China

Financial institutions have been using predictive analytics for credit risk estimation for many decades and have been deploying complex machine learning algorithms more recently. These models operate as black boxes, in the sense that although they give an uplift in predictive performance, it is difficult to explain how they make decisions, and they may hide undesirable behaviour. In particular, the models may give rise to undetected bias in certain population subgroups. If such a subgroup is a protected group, such as ethnic or gender, the deployment of the model may infringe equality laws. Although manual checking for bias in prescribed subgroups is possible, this may still mean bias in other subgroups remains undetected and also may miss bias in intersectional subgroups. We use meta-modelling to detect local bias and error in the population. The methodology is tested on both traditional and machine learning credit scoring models, revealing bias in intersectional subgroups in both cases.

C1037: **A multilevel scoring model with pooled cross-sectional and multi-period data including financial agents efficiency**
Presenter: **Vassilis Ioannou**, University of Edinburgh, United Kingdom
Co-authors: Raffaella Calabrese, Finn Lindgren

Researchers and practitioners in credit risk often need to use data pooled from various sources, which are clustered in a number of dimensions, such as different geographies, or contributing lenders. We suggest modelling default risk using a multilevel approach in order to capture the complexity of Industry level data to predict accurately across different subsegments of the population. Default risk is modelled using discrete-time survival analysis, in the presence of the competing risk of prepayment, allowing for time-varying coefficients. Finally, extensive national-level data are used to explore whether the efficiency of financial agents, such as originators and servicer companies, has an impact on the conditional probability of default.

C1946: **How climate change and supply chain affect SMEs financing**
Presenter: **Raffaella Calabrese**, University of Edinburgh, United Kingdom
Co-authors: Marc Cowling, Yijun Wang

As Small and Medium Enterprises (SMEs) represent the large majority of UK companies, they are crucial to achieving the net zero target by 2050. In the climate change literature, transition risk is the uncertainty caused by changing policies, strategies and investments to decrease carbon emissions. Given SME limited availability of internal funds and resources, they usually rely on external credit, customers and suppliers. Brexit and Covid cause major supply chain issues in the UK and worldwide. Therefore, we analyse the effects of supply chain and climate change on SME finance to understand if SMEs operating in sectors largely affected by transition risk are facing greater financial constraints.

C1701: **The generalised extreme value gradient boosting decision tree for credit scoring**
Presenter: **Junfeng Zhang**, University of Edinburgh, United Kingdom
Co-authors: Raffaella Calabrese, Yizhe Dong, Baofeng Shi

The performance of the credit scoring models can be undermined due to the imbalanced datasets, as the number of defaulted borrowers is much lower than that of non-defaulters. We propose a gradient-boosting decision tree with the generalized extreme value distribution (GEV-GBDT) to handle the imbalance learning problem. Four real-life loan portfolio datasets are used, including the LendingClub P2P loan dataset and three small and medium enterprise loan datasets provided by Chinese commercial banks. The empirical result shows our model harvests better classification performance compared with other commonly used imbalance learning methods, including SMOTE and cost-sensitive framework with GBDT, logistic regression and random forest. We also test the performance of GEV-GBDT on a series of designed datasets with different imbalance ratios and find GEV-GBDT performs even better on extremely imbalanced datasets.

C2022: **The fairness of credit scoring models**
Presenter: **Sebastien Saurin**, University of Orlaans, France
Co-authors: Christophe Hurlin, Christophe Perignon

In credit markets, screening algorithms aim to discriminate between good-type and bad-type borrowers. However, when doing so, they also often discriminate between individuals sharing a protected attribute (e.g. gender, age, racial origin) and the rest of the population. We show how (1) to test whether there exists a statistically significant difference between protected and unprotected groups, which is called lack of fairness and (2) to identify the variables responsible for the lack of fairness. We then use these variables to optimize the fairness-performance trade-off. The framework provides guidance on how algorithmic fairness can be monitored by lenders, controlled by their regulators, and improved for the benefit of protected groups.

CO366 Room BH (SE) 2.05 CRYPTOCURRENCY ANALYTICS Chair: Marcell Tamas Kurbucz

C0532: **Contagion from Bitcoin to stock markets: Instability, contemporaneous and long-run volatility transmission**
Presenter: **Walter Bazan-Palomino**, Fordham University, United States

The volatility transmission is investigated from Bitcoin to stock markets in North America, Europe, and Asia-Pacific from 2013 to 2021. After calculating the range volatility time series for all markets, we employ a Heterogeneous Autoregressive Distributed Lag Model of Range Volatility (HARDL-RV) to assess volatility spillovers. We find structural breaks in the parameters of the HARDL-RV model, indicating time-varying volatility spillovers. Moreover, the rolling-window estimation of the HARDL-RV model detects a strong and sudden increase in volatility transmission in all regions from March 2020 onward. During this latter period, all regions experienced a positive and stronger contemporaneous volatility contagion. However, Bitcoin's long-run effect on the North American and three European markets tends to disappear, while it tends to remain in most of the Asia-Pacific markets. Therefore, policymakers might monitor Bitcoin volatility to preserve financial stability, whereas investors should be aware of the difference in contemporaneous and long-run contagion across regions for risk management, and futures and options pricing.

C0542: **Moving-average technical rules with a variable band in cryptocurrency markets**
Presenter: **Daniel Svogun**, The Catholic University of America, United States

There has been a lot of study on medium and long-run moving average trade rules in a variety of asset markets, including the cryptocurrency market, with varying levels of return beyond buy-and-hold. The bulk of those in the literature considers a 0% or 1% band parameter. We present a unique trading algorithm that adjusts yearly to the most profitable rules with all meaningful and positive bands that change the trade pattern considered. We compare its performance to standard trade rules in the literature and buy-and-hold. Initial results vary dramatically by year, and, interestingly, tend towards more stable returns for the algorithm, with a slight average return cost for it.

C0577: **On empirical challenges in forecasting market betas in crypto markets**
Presenter: **Jan Sila**, UTIA AV CR, v.v.i., Czech Republic
Co-authors: Michael Mark, Ladislav Kristoufek

The predictability of market betas for crypto assets is investigated. The market beta is the optimal weight of a short position in a simple two-asset portfolio hedging the market risk. Investors are, therefore, keen to forecast the market beta accurately. Estimating the market beta is a fundamental financial problem, and we document pervasive empirical issues that arise in the emerging market of crypto assets. Although recent empirical results about US stocks suggest predictability of the future realized betas is about 55%, predictability for the universe of crypto assets is at most 20%. Our results suggest that the crypto market betas are highly sensitive not only to the beta estimation method but also to the selection of the market index. Thus, we also contribute to the discussion on the appropriate market representation.

C0604: **Influencer detection between cryptocurrency sectors via sparse network analysis**
Presenter: **Kexin Zhang**, City University of Hong Kong, Hong Kong
Co-authors: Simon Trimborn

The rapidly changing cryptocurrency market evolved from a market focused on payment systems into various sectors operating on blockchains but dedicated towards entirely different goals. The emergence of sectors within the cryptocurrency market raises the question of the interdependence between asset classes belonging to different groups. We introduce a 3-layer sparse network AutoRegressive estimator to identify the influential cryptocurrencies and sectors. We study the asymptotic properties of the estimator and validate its performance in extensive synthetic data experiments. We study 55 cryptocurrency sectors and highly capitalized cryptocurrencies during 2015-2020 for their influence on each other. During the bear market of 2015-2016, the lead-lag effect is not significant throughout the years. However, during the bull market of 2017-2019, including the 2017 market frenzy, Monero and Bitcoin frequently determine the market direction and influence sectors' performance. From 2020 onwards, following a period of strong market growth, various sectors and Bitcoin influenced each other's performance. This suggests that the market entered a new stage since its behavior suggests interdependencies between previously unrelated sectors. Further, it suggests that Bitcoin still plays an important role despite the market entering a new stage.

C0660: **Predicting the price movement of cryptocurrencies using linear law-based feature space transformations**
Presenter: **Marcell Tamas Kurbucz**, Wigner Research Centre for Physics, Corvinus University of Budapest, Hungary
Co-authors: Antal Jakovac, Peter Posfay

The aim is to investigate the effect of a novel method called Linear Law-based feature space Transformations (LLT) on the accuracy of intraday cryptocurrency price movement prediction. To this, first, sample series of the 1-minute interval price data of Bitcoin, Ethereum, and Ripple between 1 January 2021 and 2 August 2022 are obtained from Bitstamp. Then, the training and test sets of samples are separated. After that, LLT identifies the governing patterns (laws) of each input sequence in the training set by applying time-delay embedding and spectral decomposition. Finally, the laws of the training set are used to transform the feature space of the test set. The transformed test set is classified by traditional machine learning algorithms, such as decision tree (DT), support vector machine (SVM), and k-nearest neighbor (KNN) with 10-fold cross-validation. The results emphasize the potential of the LLT method in terms of both accuracy and calculation time in price movement prediction.

CO730 Room BH (SE) 2.10 MEASURING CLIMATE-RELATED FINANCIAL RISKS Chair: Juan-Angel Jimenez-Martin

C0324: **Tail sensitivity of stocks to carbon risk: A sectoral analysis**
Presenter: **Dolores Robles**, Universidad Complutense de Madrid, Spain
Co-authors: Laura Garcia-Jorcano, Juan-Angel Jimenez-Martin

The tail risk dependence of industry returns and carbon-related climate risk as proxied by the CO2 emission allowance returns is analyzed. We estimate the tail carbon-betas in four scenarios in which extreme conditions in both markets occur together and find asymmetric tail dependence that is also different in the last two phases of the EU-ETS implementation. The study of 60 US industries from 2009 to 2020 reveals that tail dependence is stronger when the industry is under upside risk, and the carbon market signals a brown state than when the industry is under downside risk. The carbon market signals a green state. The tail dependence in the other two scenarios shows less asymmetry, being the tail carbon-sensitivities more similar in magnitude in this case. The results indicate a stronger dependence when the industry is under downside risk, and the carbon market signals a brown state than when the industry is under upside risk, and the carbon market signals a green state. Overall, our findings point out that the conditions in both markets determine investor awareness about carbon-related climate risk.

C0334: **The role of a green factor in stock prices: When Fama & French go green**
Presenter: **Clara Gonzalez**, Bank of Spain, Spain
Co-authors: Ricardo Gimeno

Concerns about climate change are now widespread, and the risks for financial assets have become more evident. Investors are increasingly aware of the need to incorporate climate-related considerations in their investment decisions. All this has had an impact on market valuations. We extend the framework of the factor models that explain the expected return of stock models to include a climate change exposure factor. To do so, we built a portfolio that is long on companies with low carbon emissions, and short on companies with high carbon emissions. We show that this factor is relevant in the market and allows for an approximation of the climate change exposure of firms with poor disclosure of their green performance.

Thus, the betas of this factor could be a useful tool for investors that wish to incorporate these aspects in the management of their portfolios and analysts interested in corporate exposure to climate change risks.

C0362: **Growth exposure to climate change risk: ClimaGRisk**
Presenter: **Laura Garcia-Jorcano**, Universidad de Castilla-La Mancha, Spain
Co-authors: Juan-Angel Jimenez-Martin, Dolores Robles

Climate change may alter the growth trajectory of a country, for example, by reducing productivity (particularly in agriculture), destroying productive assets (during extreme events), or altering investment priorities (from productive investment toward adaptation). It, therefore, remains a pertinent question of how the growth strategies of developing countries may have to be adjusted to account for climate change. The co-movement between climate risk and growth is explored through a Climate Risk exposure measure, ClimaGRisk, inspired by the financial systemic risk literature. We find evidence of a sign-switching sensitivity at different quantiles, which implies that the uncertainty surrounding climate change risk seems to have an asymmetric impact on the growth conditional distributions.

C0469: **Forecasting climate risk by extreme sea level rises and its impact on financial markets**
Presenter: **Lidia Sanchis-Marco**, University of Castilla-La Mancha, Spain
Co-authors: Laura Garcia-Jorcano

The demand for accurate predictions of sea-level rise in several affected sectors is increasing, but there is no standard method to measure it. Using the global and regional mean sea-level rise (mm) every 10 days (Dic/1992-Oct/2020), we propose two new measures, Extreme Sea-Level Value at Rise (ExSLVaR) and Extreme Sea-Level Expected Rise (ExSLER), to forecast extreme mean sea-level at several periods calculated for eight seas/oceans of the Earth. The method used is Extreme Value Theory and the Filtered Historical Simulation approach. Furthermore, we analyze the connection between our measures and financial risk in different sectors. The main evidence shows different regional and global forecasts, and both measures capture more risk in the energy and oil-gas sectors, especially in the current COVID-19 period. The extreme sea-level rise measures are relevant for regulators and investors for mitigation and adaptation strategies that reduce future physical costs of climate change.

C0899: **The impact of sustainability on financial risk: A firm level analysis**
Presenter: **Almudena Maria Garcia Sanz**, Complutense University of Madrid, Spain

Existing research on the effect of sustainability on different measures of financial risk is complemented and extended. Panel data of listed US and European firms for a period between 2000 and 2021, and the Thomson Reuters ESG scores have been used to directly assess such an impact avoiding the existing problems in the literature about the lack of standardized measures. Results show that sustainable activities carried out through the corporate social responsibility (CSR) policies of firms play a relevant role in mitigating total, idiosyncratic and systematic risk, as well as in mitigating the lack of synchronicity with the market while helping to increase the distance to default, even after controlling for characteristics such as size or profitability. Complementary analyses were carried out to know more about the differences between subsamples (European and American), and the effect of the covid pandemic.

CO036 Room BH (SE) 2.12 TOPICS IN QUANTITATIVE FINANCE Chair: Gabriele Torri

C0298: **Sentiment-based financial networks and the impact of monetary policy shocks**
Presenter: **Petre Caraiani**, Bucharest University of Economic Studies, Romania

Recent work in the transmission of monetary policy shocks on financial markets has incorporated the role of networks. This has been done generally in two ways, either using data from production networks, or by using data from networks of assets. Statistically, most of the work has been done based on simple correlations or derived from VAR models. We consider networks that are formed on the basis of the sentiments characterizing sectors or stocks on the networks. We also aim to propose new ways to construct statistical networks relevant to the financial markets and the sentiment characterizing them. We then study how the transmission of monetary policy works through such networks.

C1467: **Do lower ESG-rated companies have a higher systemic impact: Empirical evidence from Europe and the United States**
Presenter: **Sandra Paterlini**, University of Trento, Italy
Co-authors: Karoline Bax, Giovanni Bonaccolto

In recent years, companies have increasingly been characterized by environmental, social and governance (ESG) scores, and investors like academics have raised questions concerning financial performance and investment risks. Now, as the EBA has acknowledged that ESG risks can potentially impact the financial system, the debate on systemic risk has risen. While understanding the relationship between ESG merit and systemic risk is of utmost importance for the financial system's stability, only scarce knowledge still exists. Relying on real-world European and American data, we quantify the systemic risk impact by means of QL-CoVaR. Empirical analysis of the entire period from 2007-2021 shows that companies with a high ESG score tend to exhibit lower QL-CoVaR values than lower-rated companies indicating a positive effect of ESG scores. Such evidence is also confirmed by clustering the individual companies into ESG portfolios, and becomes clearer focusing on Covid-19

C1322: **The generalized precision matrix: Dependency in non-Gaussian settings, theory and financial applications**
Presenter: **Gabriele Torri**, University of Bergamo, Italy
Co-authors: Sandra Paterlini, Emanuele Taufer, Rosella Giacometti, Gyorgy Terdik

Multivariate financial time series are characterized by highly non-Gaussian distributions, showing fat tails and high levels of tail correlation. Due to these stylized facts, tools such as partial correlation networks cannot properly be used to characterize the interconnectivity structure of random variables. Starting from local dependency measures, we propose a generalization of the precision matrix that describes the interconnectivity structure of multivariate random variables in a single point of the probability space, in a region, or under any conditioning. We use a Gram-Charlier expansion of the density to show how this matrix is related to the traditional precision matrix, we then discuss several parametric cases, focusing on distribution with fat tails, and we illustrate financial applications.

C1605: **Improving index tracking and portfolio performance with regularisation**
Presenter: **Dietmar Maringer**, University of Basel, Switzerland
Co-authors: Sandra Paterlini

Historical data are often used for portfolio optimization when parametric distribution assumptions fail to capture key properties of asset returns. This, however, comes with the practical problem of choosing training window length: Long time series might include outdated observations that are no longer representing the immediate future, while short time series with more recent observations only are more prone to all sorts of inaccuracies typical for small sample sizes and lead to overfitting, in particular when the number of assets is large. Recent studies found that regularization is beneficial in the context of finding efficient portfolios. However, little guidance is usually given on calibration issues, and passive management techniques are rarely addressed. The aim is to fill these gaps. First, the elastic net, combining Lasso and ridge, is presented along different versions of index tracking and enhanced indexing. Next, an empirical study is performed analysing how different calibrations affect the reliability of optimized portfolios, considering different objectives, in-sample training window lengths, and out-of-sample investment horizons. We find that regularization can indeed improve the out-of-sample tracking error and other criteria, but the amount of improvement depends substantially on the calibration and the additional objectives considered on top of the tracking error.

C1632: **Maximum drawdown-optimized portfolios**
Presenter: **Philipp Staehli**, University of Basel, Switzerland

Co-authors: Dietmar Maringer

The maximum drawdown (MDD) is the largest cumulative loss from a peak to the following trough within a given period of time. It is one of the most widely used path-dependent risk measures in the fund management industry, and it is often used as an additional criterion to assess a portfolio or strategy. However, there is little empirical research on portfolios explicitly optimized for MDD. MDD is used as an objective for portfolio optimization. Based on S&P 500 Health Care stocks' data for 2012-2020, an empirical study is performed with 2000 random combinations of assets for different time windows. For each of these situations, gradient-based sequential least squares were used to minimize the in-sample MDD and the in-sample variance, respectively. These optimized portfolios were then analysed for their out-of-sample performance. As expected, the out-of-sample return of MDD-optimized portfolios was higher, and the out-of-sample MDD was lower than their minimum-variance (MVP) counterparts. At the same time, however, MDD-optimized portfolios typically outperformed their MVP counterparts by having lower Value-at-Risk and lower Expected Shortfalls for out-of-sample windows up until the end of 2019. These advantages do not prevail for out-of-sample windows in 2020, i.e., after the covid-crisis had begun while portfolios were optimized predominantly on pre-covid data.

CC750 Room BH (SE) 1.01 ECONOMETRIC MODELLING Chair: Alain Hecq

C1364: **Estimating nonparametric conditional frontiers and efficiencies: A new approach**

Presenter: **Camilla Mastromarco**, University of Calabria, Italy

Co-authors: Leopold Simar, Ingrid Van Keilegom

Conditional frontiers measures are a flexible and appealing approach to considering the role of environmental variables in the production process. Direct approaches estimate non-parametrically conditional distribution functions requiring smoothing techniques and the use of selected bandwidths. The statistical literature produces ways to derive bandwidths of optimal order, by using, e.g., least squares cross-validation techniques. However, it has been shown that the resulting order may not be optimal when estimating the boundary of the distribution function. As a consequence, the direct approaches may suffer from some statistical instability. We suggest a fully nonparametric approach which avoids the problem of estimating these bandwidths, by eliminating, in a first step, the influence of the environmental factors on the inputs and the outputs. By doing this, we produce pure inputs and outputs, which allow estimating a pure measure of efficiency, which is more reliable for ranking the firms, since the influence of the external factors has been eliminated. This can be viewed as an extension of the use of location-scale models (semi-parametric structure) to full nonparametric models, based on nonseparable, nonparametric models. We are also able to recover the frontier and efficiencies in original units. We describe the method, its statistical properties and we show in some Monte Carlo simulations, how our new method dominates the traditional direct approach.

C1620: **Identifying the elasticity of substitution with biased technical change: A structural panel GMM estimator**

Presenter: **Arvid Raknerud**, Statistics Norway, Norway

Co-authors: Thomas von Brasch, Trond Vigtel

The aim is to provide a structural panel GMM estimator (P-GMM) of the elasticity of substitution between capital and labour that does not depend on external instruments, and which can also be applied in the presence of biased technical change. We identify the conditions under which P-GMM yields unbiased estimates and compare it to a fixed effects estimator, which is unbiased when factor prices are exogenous. Using a Monte Carlo study, we examine the properties of the proposed P-GMM estimator. The fixed effects estimator is found to be heavily downward biased in the presence of simultaneity. In contrast, the P-GMM estimator is nearly unbiased, provided the number of time periods T is not too small (say, more than 10). We apply the estimator to a sample of manufacturing firms in Norway. In our application, with an unbalanced sample and T equal to 12, we estimate the elasticity of substitution to be 1.8 using P-GMM and 1.0 using a fixed effects estimator. Hence, neglecting simultaneity may lead to the conclusion that capital and labour are complements when, in fact, they are substitutes.

C1983: **Stochastic correlation modelling with Von Mises process**

Presenter: **Sourav Majumdar**, Indian Institute of Management Ahmedabad, India

Multi-asset financial derivatives can carry an intra-asset correlation risk. Previous empirical studies show that the correlation between assets is not stable, and there is a correlation risk premium. Pricing models for these derivatives should account for the time-varying nature of the correlation. Assuming constant correlation may lead to pricing and hedging risks. We propose a continuous-time model for correlation as a random variable on the circle. We consider a model based on the von Mises process whose stationary distribution is the von Mises distribution, the maximum entropy distribution on the circle. We discuss several estimation methods and results for the model. We apply the model to real-life financial data to study the correlation between equity and currency exchange rates.

C1682: **Discriminating direct from induced equilibrium mean shifts**

Presenter: **David Hendry**, University of Oxford, United Kingdom

Co-authors: Jennifer L Castle, Jurgen Doornik

Equilibrium mean, or location, shifts can result directly from changes in intercepts with constant dynamics, or be induced by shifts in dynamics (or other parameters) when data means are non-zero. The impacts of in-sample-induced shifts substantively modify previous taxonomies of forecast errors. Step-indicator saturation helps detect any resulting location shifts. However, even when all relevant variables in the data generation process (DGP) and all indicators matching DGP shifts are selected in the forecasting model, mis-forecasting can occur. To discriminate direct from induced shifts, we add to the model multiplicative indicators formed by interacting all selected step indicators with the lagged regressand. When equilibrium-mean or location shifts are induced by changes in dynamics, forecasts can be markedly improved when these interactive indicators are included.

C0671: **Linear panel regression models with non-classical measurement error: An application to investment equations**

Presenter: **Kazuhiko Hayakawa**, Hiroshima University, Japan

Co-authors: Takashi Yamagata

A new minimum distance estimator is proposed for linear panel regression models with measurement error and analyzes its theoretical properties. The model considered is more general than examined in the literature in that (i) measurement error is non-classical in the sense it is allowed to be correlated with true regressors, and (ii) measurement error and idiosyncratic error can be serially correlated. Notably, the proposed estimator does not require any instrumental variables to deal with the endogeneity. The finite sample evidence confirms that the proposed estimator has desirable performance. We revisit the investment model and theoretically illustrate that measurement error is negatively correlated with Tobin's marginal q, which is empirically supported by applying the proposed method to US manufacturing firm data for the period 2002-2016. Furthermore, we find that there is a structural break in 2008 and cash flow is insignificant before 2007 but becomes significant after 2009.

CC805 Room BH (SE) 1.06 FINANCIAL ECONOMETRICS I Chair: Leopold Soegner

C1512: **Exact likelihood for inverse gamma stochastic volatility models**

Presenter: **Blessings Majoni**, National graduate institute for policy studies, Japan

Co-authors: Roberto Leon-Gonzalez

A novel closed-form solution of the likelihood for the inverse gamma stochastic volatility (SV) model is obtained. It is shown that by marginalizing out the volatilities, the model that we obtain has the resemblance of a GARCH in the sense that the formulas that we get are similar, which simplifies computations significantly and permits maximum likelihood estimation. Recent literature has also attempted to obtain SV models that

are as simple as the GARCH for computational efficiency. However, the literature has only obtained this solution for gamma SV models and for restricted non-stationary models. We provide two empirical applications, one to UK exchange rate data and another to UK inflation data. We find that our proposed model has a better empirical fit than previously proposed models.

C1737: **A real-time monitoring test for common breaks and factors in panel data**
Presenter: **Cindy Shin Huei Wang**, HSBC Business School, Peking University, China
Co-authors: Silvia Bacci

A novel real-time monitoring test is established for detecting common breaks or factors within a panel-data framework via a panel autoregressive (PAR) approximation framework. The limiting distribution of this real-time monitoring test follows a Brownian bridge and is free of model parameters if there is no common break or factor within a panel. Its convincing finite sample performance has been confirmed through Monte Carlo simulations, even though there exist multiple breaks within a panel system. An empirical application to monitor the mean-variance convergence around the world during the period of the global crisis, including the 2007-2008 subprime crisis, the 2018-2019 Sino-US war trade, and the recent COVID-19 pandemic crisis, demonstrates the usefulness and feasibility of our real-time monitoring procedure.

C1909: **Tail index estimation in the presence of covariates: A systematic risk analysis**
Presenter: **Paulo Rodrigues**, Universidade Nova de Lisboa and Banco de Portugal, Portugal
Co-authors: Joao Nicolau, Marian Stoykov

Novel theoretical results are provided for the estimation of the conditional tail index of Pareto and Pareto-type distributions in a time series context. We show that both the estimators and relevant test statistics are normally distributed in the limit, when independent and identically distributed or dependent data are considered. Simulation results provide support for the theoretical findings and highlight the good finite sample performance of the approach in a time series context. The developed methodology is then used to evaluate systematic and unsystematic tail risk.

C0439: **Asymmetric intra-day volatility pattern and price jump detection: Evidence from international equity indices**
Presenter: **Ping Chen Tsai**, National Sun Yat-sen University, Taiwan

Current methods of price jump detection using high-frequency data typically assume a constant intra-day volatility pattern (IVP) over the sample period. We investigate the validity of this assumption by allowing IVP weights to depend on, for example, the sign of returns from day t-1 or the sign of overnight returns. Estimation results from 5-minute intra-day GARCH for four equity indices during 2019/01/03 - 2021/03/31 show that for those days with negative previous or overnight returns, squared-return-based IVP weights increase significantly in the early morning hours, suggesting a leverage effect or asymmetric IVP. For the jump-robust IVP estimator, a strong asymmetric response is found for days with realized variance increasing over the previous day. Our result is robust against a GARCH-t specification and complements recent studies on time-varying IVP. Price jumps obtained using the proposed IVP adjustment show markedly different distribution over trading hours.

C0600: **New insights on the foreign exchange risk premium through a portfolio-based approach**
Presenter: **Taejin Kim**, Korea University, Korea, South
Co-authors: Jinyong Kim, Kun Ho Kim

A simple and tractable framework is proposed for the forward premium regression to estimate foreign exchange risk premium. In contrast to previous work, we adopt the portfolio-based approach. Relevant methodology is applied to estimate the risk premiums for six currency portfolios. The empirical results show temporal variation and co-movement among estimates of the time-varying risk premiums. The study also illustrates that U.S. fundamentals have persistent power in predicting future risk premiums.

CC769 Room BH (S) 2.02 COMPUTATIONAL AND HIGH-DIMENSIONAL ECONOMETRICS Chair: Alessandra Amendola

C1905: **Machines do not go for lunch: A new diurnal adjustment for trade durations**
Presenter: **Markus Belfrage**, Hanken School of Economics, Finland

A new diurnal adjustment method is proposed for stock trade durations. A well-known feature of stock markets is the diurnal seasonality in the intensity of trading. Trade durations are often modelled by the class of Autoregressive Conditional Duration (ACD) models, where it is assumed that the seasonality factor acts multiplicatively on all durations. We show that this assumption is violated for ultra-high precision trade data when a large portion of the trades are executed by computer algorithms. A two-component mixture model with features that vary nonparametrically over time of the day is developed as a response to the heterogeneity in the diurnal seasonality caused by the mix of machines and regular traders. Furthermore, an estimation algorithm much in the flavor of expectation maximization is proposed and applied to a large set of Apple (AAPL) trades using data from Nasdaq Historical TotalView ITCH.

C0801: **Determining the number of factors in fractionally integrated factor models**
Presenter: **Dominik Ammon**, University of Regensburg, Germany
Co-authors: Tobias Hartl, Rolf Tschernig

Three different approaches are proposed to overcome limitations for factor selection in fractionally integrated factor models. Two of our methods for determining the number of factors include an approach that was designed for identifying the cointegration rank in VAR models. We extend their model selection approach by generalizing it to fractionally integrated factor models. In our two-step procedure, we first estimate the cointegration rank to obtain the non-stationary fractional factors. In the second step, we generalize the model selection criteria to fractionally integrated factors with memory smaller $1/2$ to obtain the number of asymptotically stationary factors. Before carrying out the second step, the non-stationary factors need to be removed from the data. We investigate two alternatives: i) subtract the estimated non-stationary part from the observable variables, ii) project out the non-stationary factors. In our third approach, we directly consider the model selection criteria without prior removing the non-stationary variation in the observable data. In the Monte-Carlo simulations, all three methods show satisfactory results; in particular, the third approach performs surprisingly well.

C1334: **Fast inference for high-dimensional one-factor copula models with additional Gaussian factors**
Presenter: **Alex Verhoijsen**, University of Melbourne, Australia
Co-authors: Pavel Krupskiy

Gaussian factor models allow the statistician to capture multivariate dependence between variables. However, they are computationally cumbersome in high dimensions and are not able to capture multivariate skewness in the data. We propose a copula model that allows for arbitrary margins, and multivariate skewness in the data by including a non-Gaussian factor whose dependence structure is the result of a one-factor copula model. Estimation is carried out using a two-step procedure: margins are modelled separately and transformed to the normal scale, after which the dependence structure is estimated. An estimation procedure is developed that allows for fast estimation of the model parameters in a high-dimensional setting. Theoretical results of the model with up to three Gaussian factors are proven, and simulation results confirm the results for increasing sample size and dimensionality. Finally, the model is applied to a selection of stocks of the SP500, demonstrating that the model can capture cross-sectional skewness in the stock market data.

C0354: **Parallel computing and software integration applied to financial models calibration**
Presenter: **Antonio Santos**, University of Coimbra, Portugal

Option pricing financial models usually belong to the set of testing problems associated with cutting-edge hardware and software developments

happening at a fast pace, namely the ones related to parallel computing. We demonstrate how parallel computing, possible through Graphical Processing Units, integrated with state-of-the-art software that implements numerical optimization, permits real-time calibration of option pricing models within a big data setting. We develop software that can calculate options prices in parallel, allowing a vast amount to be calculated simultaneously, reducing by a factor of several hundred the computational time needed to compute a fit function. Furthermore, the computation time reduction makes it viable that software packages full-tested in implementing standard optimization algorithms can be used to solve the calibration problem.

C1842: **Approximate Bayesian numerical method with product-Whittle-Matern-Yasuda kernel for Rosen's hedonic regression**
Presenter: **Andrej Srakar**, University of Ljubljana, Slovenia

Hedonic regression has featured an extensive amount of applications. It is estimated in a spatial equilibrium context in two stages which leads to a nonlinear partial differential equation framework. To date, its Bayesian extensions, while present, have still not adequately addressed features of its original proposal, which extends to many possible regression specifications. We develop an approximate Bayesian probabilistic numerical method with product-Whittle-Matern-Yasuda kernel, which extends existing literature in several aspects and is able to address different features of the original proposal: it is developed for nonlinear partial differential equations, is based on spatial kernels and a Gaussian process regression framework, and is applicable to any hedonic regression model specification. We develop a quasi-MC sampling algorithm and Bernstein-von Mises type asymptotic theorems to study the performance of the approach. Using Bayesian model comparison approaches, we compare its performance to several other parametric and nonparametric Bayesian priors (Zellner, Wasserstein, Dirichlet process and Dirichlet process mixtures, Bayesian additive regression trees, Bayesian causal forests) and apply it to simulated and real data examples from the areas of real estate and retail.

CC780 Room BH (S) 2.05 APPLIED ECONOMETRICS Chair: Gianluca Cubadda

C1731: **Decomposing earnings uncertainty using German SOEP data**
Presenter: **Friederike Schmal**, University of Muenster, Germany

A new approach is suggested to decompose earnings development in a predictable and unpredictable component, adjusted for the German labor market. We examine which part of income development was already predictable for the individual when he or she just left school. It is at this point that the decision about future education and thus its influence on future income is made. For this purpose, we use the link between expectations of future income variability, later realized income, and the educational decision regarding university attendance. The difficulty at this point is that only one income trajectory per individual can be observed at a time (income with university degree or without); however, we need both trajectories for our analysis. Therefore, we develop a new approach so that we can estimate the unobserved trajectory with a likelihood function for each individual. To avoid too strong restrictions on the covariance structures and to deal with the high number of parameters to be estimated and the counterfactual income trajectory, we will apply an MCMC algorithm and Gibbs sampling. A factor model will also be included, which will allow us to make statements about the predictable components. We will apply our approach to the German SOEP data that contains not only information on income history but also educational background and numerous demographic variables.

C1774: **The effect of retirement age reform on retirement behavior: Analysis and synthesis of several natural experiments**
Presenter: **Kevin Stabenow**, University of Muenster, Germany
Co-authors: Stella Martin

The 2007 retirement age reform from Germany is used to evaluate the impact of an incrementally increasing pension eligibility age on the likelihood of retiring. We use an administrative data set from the German public pension fund containing full employment biographies on a monthly basis. We estimate treatment effects for the natural experiment we observe at each marginal retirement age increase of one month. We then set up a meta-analytic random effects model to synthesize our findings from the natural experiments in order to elicit an overall treatment effect of a rise in the legal retirement age. We find that individuals with an increased pension age retire later and are less likely to be in early retirement prior to their legal retirement age than those in the control group are at a given age.

C1892: **Environmental and socioeconomic determinants of cardiovascular disease: The case of Poland**
Presenter: **Sylwia Roszkowska**, University of Lodz, Poland
Co-authors: Barbara Kula, Natalia Pawelec, Michal Swieczkowski, Adrian Kubis, Anna Tomaszuk-Kazberuk, Hanna Bachorzewska-Gajewska, Lukasz Kuzma

The aim is to assess the influence of environmental and socioeconomic factors on the prevalence of hospitalization for cardiovascular disease (CVD). The rate of CVD hospitalizations in Poland is one of the highest in Europe. The risk of hospitalization in patients with CVD is related not only to lifestyle, but also to environmental, economic, social factors and the organization of the health care system. Using the National Institute of Public Health, Central Statistical Office and Voivodeship Inspectorates for Environmental Protection in Poland data and panel data techniques (including spatial ones), the impact of environmental and socioeconomic factors on CVD in Poland has been estimated. The age- and sex-standardized data on the prevalence of hospitalization due to CVD in Poland in 2012-2019 on poviats level were explained using a set of air quality indicators (namely PM2.5, SO2, NO2), environmental factors, organization of the health care system and regional determinants of the socioeconomic situation. The results of the provided analyses indicate that regional indicators such as the share of physicians and healthcare expenditure, air quality, green areas density, working-age population structure, unemployment rate and GDP were associated with changes in CVD hospitalizations. The increasing number of CVD hospitalizations in the last decade is most noticeable in regions with low socioeconomic development.

C1005: **Time series forecasting of tourist arrival in Singapore**
Presenter: **Khay Boon Tan**, Singapore Institute of Management, Singapore

The aim is to apply various forecasting time series methods to forecast tourist arrival in Singapore. Deterministic time trend models, the Smoothing method and ARIMA modelling are used, and a one-step-ahead forecast is performed to compute the mean square forecast error. It is discovered that the log-linear model has the best goodness of fit among the deterministic time trend models, and the Holt and Winters smoothing method incorporating seasonality has the best performance among the class of smoothing methods. For ARIMA modeling, the best performance models are AR(2) model and ARIMA (1,1,1) models. When all the models are applied to generate forecasts, it is discovered that Holt and Winters exponential smoothing incorporating additive seasonality put up the best performance in the forecast of tourist arrival in Singapore.

C1095: **Leverage driven bubbles in the Chinese real estate corporate sector**
Presenter: **Isabel Catalina Figuerola-Ferretti Garrigues**, Universidad Pontificia Comillas, Spain

Motivated by the recent default of Evergrandel, the existence of bubble behaviour in equity prices of the Chinese real estate sector is analyzed. A previous methodology is applied for this purpose. Using weekly data of the benchmark real state index and six of its components, we find that there are three predominant periods of bubble behaviour in 2015, 2017-2018, 2021-2022, that can be linked to the value of natural fundamentals, the real dividend yield, and the ten years Chinese Treasury yield. We offer an additional dimension to the analysis of fundamentals that allows an explanation of the bubble process in terms of leverage-related metrics. Our results highlight the stress in the Chinese real estate sector shedding light on the potential future strain in global financial markets.

CC785 Room BH (SE) 2.09 MACHINE LEARNING Chair: Leopoldo Catania

C0352: **A new non-degenerate test for model selection based on maximum-mean-discrepancy**
Presenter: **Florian Brueck**, Technical University Munich, Germany
Co-authors: Jean-David Fermanian, Aleksey Min

The purpose is to investigate the influence of parameter estimation on the asymptotic distribution of the two-sample test based on the Maximum-Mean-Discrepancy. To circumvent the problem of determining quantiles of an infinite sum of Chi-squared distributed random variables, we propose a new two-sample test based on Maximum-Mean-Discrepancy, which solely requires to determine quantiles of the standard normal distribution, while approximately keeping the power of the original two-sample test. Moreover, we deduce a new test for model comparison based on Maximum-Mean-Discrepancy, which only requires determining quantiles of the standard normal distribution.

C0461: **Media influences on agricultural commodity pricing**
Presenter: **Xinquan Zhou**, Dublin City University, Ireland
Co-authors: Guillaume Bagnarosa, Jagadish Dandu, Michael Dowling

Textual machine learning is applied to 290,271 business news articles related to corn markets (2009-2020), to model the impact of news on corn pricing. Our approach allows the identification of seven distinct topics of corn news that well-describe the typical range of news coverage. We identify four topics on the fundamentals of corn markets around crop progress, weather impacts, exports, and USDA reports. We also identify three further topics on the relationship with wheat markets, soybean and biofuel markets, as well as financial market news related to corn. We first discuss these topics, and then show how the integration of news variables for each topic allows us to improve understanding of corn returns and price volatility. We demonstrate that news about financial markets, soybean-biofuels, crop progress, and exports, significantly contributes to explaining corn price dynamics. Our volatility analysis also demonstrates that soybean-biofuel news, especially, contributes to the level of uncertainty in corn pricing. Our study demonstrates the value of refined analysis of news flow in agricultural commodity markets.

C0237: **Man vs. machine learning to time markets: Who will win?**
Presenter: **Go Charles-Cadogan**, University of Leicester, United Kingdom

A market timing classifier (MTC) for high-dimensional data is introduced, which is based on a theory of portfolio manager market timing behaviour. This lies in stark contrast to the growing literature on the use of machine learning classifiers for data mining and market timing with big data in finance. The MTC separates statistically significant lower dimensional market timing portfolios from those with less timing ability. So, it falls in the class of dimension-reducing Neyman-Pearson classifiers. We applied the MTC to two different samples of high-dimensional asset pricing data, and the results show that the algorithm is able to separate statistically significant lower-dimension market timers from non-market timers. Fama-French size and value portfolios are included among significant market timers, as well as other portfolios predicated on anomalies. The MTC can be used as a pretest estimator in LASSO and machine learning classification schemes.

C1843: **A generative model of a limit order book using recurrent neural networks**
Presenter: **Hanna Hultin**, KTH Royal Institute of Technology, Sweden
Co-authors: Henrik Hult, Alexandre Proutiere, Ala Tarighati, Samuel Samama

A generative model based on recurrent neural networks for the complete dynamics of a limit order book is developed. The model captures the dynamics of the limit order book by decomposing the probability of each transition into a product of conditional probabilities of order type, price level, order size, and time delay. Each such conditional probability is modeled by a recurrent neural network. Several evaluation metrics for generative models related to trading execution are introduced. Using these metrics, it is demonstrated that the generative model can be successfully trained to fit both synthetic and real data from the Nasdaq Stockholm exchange.

E2035: **A text similarity-based algorithm for seed word generation in improving document classification**
Presenter: **Morteza Namvar**, The University of Queensland, Australia
Co-authors: Celeste Li

Topic modelling techniques typically use document-level co-occurrence information to group semantically related words into a single cluster or topic. Since the objective of these models is to maximize the probability of the observed data, the identified topics tend to explain only the most obvious aspects of a corpus and do not necessarily represent a construct. Interactive topic modelling techniques can be used as an alternative to unsupervised ones, as they can tackle the above issues by developing topics based on the initial seed words. As the performance of these interactive techniques heavily depends on the initial seed words, our study proposes how text features can be used to generate seed words in developing interactive topic models. We propose a method for seed word vector (SWV) generation. We provide initial SWVs for interactive topic modelling through qualitative content analysis. Then through several iterations, our developed algorithm updates SWVs from the corpus by considering document similarity. Our method of SWV generation, combined with interactive topic modelling, helps develop a probability vector of each document in the corpus, indicating their relevance to study constructs. To test our proposed method's validity and applicability in practice, we investigate the post-adoption use of contact tracing mobile applications during the COVID-19 pandemic. The results show a significant improvement in topic modelling using the generated SWVs.

Saturday 17.12.2022 13:35 - 15:40 Parallel Session D – CFE-CMStatistics

EV746 Room Virtual R01 STATISTICAL MODELLING Chair: Enrico Ripamonti

E1808: **Family of mixtures of multivariate Poisson log-normal distributions for clustering high dimensional count data**
Presenter: **Andrea Payne**, Carleton University, Canada
Co-authors: Anjali Silva, Steven Rothstein, Paul McNicholas, Sanjeena Dang

Multivariate count data encountered in bioinformatics are high dimensional and often exhibit over-dispersion. Mixtures of multivariate Poisson lognormal (MPLN) models have been used to analyze these multivariate count measurements efficiently. In the MPLN model, the latent variable comes from a multivariate Gaussian distribution and the counts, which are conditional on this latent variable, are modeled using a Poisson distribution. The MPLN model can account for over-dispersion and allows for correlation between the variables. We extend the mixture of multivariate Poisson-log normal distributions for high dimensional data by incorporating a factor analyzer structure in the latent space. A parsimonious family of mixtures of Poisson log-normal distributions are proposed by decomposing the covariance matrix and imposing constraints on these decompositions. We demonstrate the performance of the model using simulated and real datasets.

E1876: **Variance components testing in mixed effects models with small sample size**
Presenter: **Tom Guedon**, Inrae, France
Co-authors: Charlotte Baey, Estelle Kuhn

Mixed-effects models are latent variable models that allow for modeling intra and inter-individual variability in a population. Those models involve two types of effects: the fixed ones common to all individuals and the random effects that vary from one individual to another. Identifying the effects that can be modelled as fixed would reduce the number of parameters, and would also help to identify better the processes that cause the observed variability in the population. Formally, this question can be formulated as a test for the nullity of a block of components of the covariance matrix of the random effects. Since we are interested in an efficient testing procedure with small sample sizes, we propose a parametric bootstrap test procedure. The main issues are the fact that the true values of the variance parameters lie on the boundary of the parameter space, and that the Fisher information matrix is singular under the null hypothesis. Moreover, it is shown here that with a shrunk bootstrap parameter, the bootstrap test is consistent. The performance of the proposed methodology is highlighted through simulation studies.

E1900: **On integral part distribution models**
Presenter: **Violetta Piperigou**, University of Patras, Greece

Univariate integral part models have been considered in the area of risk control. It can be seen that these models are associated with weighted partial sum distributions. Moreover, bivariate discrete stochastic multiplicative models are introduced, and various properties are discussed. Some special cases are presented in detail.

E1937: **A finite mixture model for biclustering longitudinal trajectories: An application to Italian crime data**
Presenter: **Maria Francesca Marino**, University of Florence, Italy
Co-authors: Marco Alfo, Francesca Martella

Motivated by the analysis of data entailing the number of crime events that the Italian enforcement authorities (Polizia, Arma dei Carabinieri, Guardia di Finanza) reported to justice from 2012 to 2019, a biclustering approach is developed based on a finite mixture model. The data at hand represent a particular type of three- way data, where the modes correspond to Italian provinces (rows), crime-types (columns), and years (layers). A finite mixture of generalized linear models is built up to obtain a clustering of provinces. Further, within each cluster, we use a flexible and parsimonious parameterization of the linear predictor to obtain a partition of columns, such that each partition collects crime-types sharing a similar evolution over time. The aim is to identify geographical areas in the country that share common longitudinal trajectories for specific subsets of crime-types. Model parameter estimates are derived via a maximum likelihood approach based on the use of an extended EM-type algorithm. This is based on three separate steps: an expectation (E-), a classification (C-), and a maximization (M-) step. The efficacy of the proposal is also evaluated via a large-scale simulation study, based on varying sample sizes, number of partitions, ad model specifications.

E1570: **A function-based approach to model the measurement error in wearable devices**
Presenter: **Sneha Jadhav**, Wake Forest University, United States

Physical activity (PA) is an important risk factor for many health outcomes. Wearable devices such as accelerometers are increasingly used in biomedical studies to understand the associations between PA and health outcomes. Statistical analyses involving accelerometer data are challenging due to the following three characteristics (i) high dimensionality, (ii) temporal dependence, and (iii) measurement error. To address these challenges, we treat accelerometer-based measures of PA as a single function-valued covariate prone to measurement error. Specifically, in order to determine the relationship between PA and a health outcome of interest, we propose a regression model with a functional covariate that accounts for measurement error. Using regression calibration, we develop a two-step estimation method for the model parameters and establish their consistency. A test is also proposed to test the significance of the estimated model parameters. Simulation studies are conducted to compare the proposed methods with existing alternative approaches under varying scenarios. Finally, the developed methods are used to assess the relationship between PA intensity and BMI obtained from the National Health and Nutrition Examination Survey data.

EI009 Room Safra Lecture Theatre STATISTICAL METHODS IN NEUROIMAGING Chair: John Kornak

E0173: **Incorporation of brain spatial and connectivity-based information in statistical regularization**
Presenter: **Jaroslaw Harezlak**, Indiana University School of Public Health-Bloomington, United States
Co-authors: Aleksandra Steiner, Damian Brzyski, Timothy Randolph, Joaquin Goni

Prior information use in a principled manner can improve the quality of the regression coefficient estimation. Our proposal incorporates structural connectivity derived from Diffusion Weighted Brain Imaging and cortical spatial distance in the penalized approach. Extending previously developed methods informing the estimation of the regression coefficients, we incorporate such information via a Laplacian matrix based on the proximity measures. The penalty term is constructed as a weighted sum of structural connectivity and proximity between cortical areas. Simulation studies show improved estimation accuracy. We apply our approach to the data collected in the Human Connectome Project, where the cortical properties of the left hemisphere are found to be associated with vocabulary comprehension.

E0174: **Improved methods for biomarker detection and data harmonization in neuroimaging studies of Alzheimer's disease**
Presenter: **Dana Tudorascu**, University of Pittsburgh, United States

Biomarkers of Alzheimer's disease (AD), such as gray matter volume, white matter hyperintensities, amyloid and Tau are derived from Neuroimaging studies, including MRI and PET. There are many challenges in accurately quantifying these biomarkers, including technical (i.e., different scanners) as well as clinical manifestation of AD (i.e., atrophy) that require new or improved statistical methods for their quantification. We present improved statistical methods to address both: technical and clinical challenges. These improvements will be presented on a sample of AD subjects from studies at our institution.

E0175: **Modeling cortical brain network activity through concurrent EEG-FNIRS**
Presenter: **Hernando Ombao**, KAUST, Saudi Arabia

Co-authors: Marco Pinto

Brain activity is multi-faceted. Various imaging modalities aim to probe specific aspects of brain activity. In particular, electroencephalograms (EEGs) capture electrical activity, while functional near-infrared spectroscopy (fNIRS) captures cortical hemodynamic activity. Emerging technology now allows for the simultaneous collection of both signals. However, the interaction between EEG and fNIRS at a macroscopic scale is still unclear, and there is evidence that not all electrical recordings can be directly correlated to neurological activity. We will develop statistical models for joint EEG+fNIRS data, which we hope unveil a more comprehensive characterization of the inherent complex dynamics in the brain. Our proposed model for multimodal spatial-temporal brain dynamics has the potential to offer a description of the brain's self-regulating mechanisms. We will introduce some statistical spectral dependence measurements, including multimodal spectral causality and cross-frequency coupling metrics, that have the potential to contribute in distinguish between EEG components related to scalp (endothelial) vasomotion processes from neurological brain dynamics.

EO410 Room S-2.23 DIMENSION REDUCTION AND MODELING OF COMPLEX DATA STRUCTURES Chair: Joni Virta

E1069: **Higher order parametric inverse regression**

Presenter: **Daniel Kapla**, TU Wien, Austria

A method is proposed for sufficient dimension reduction of tensor-valued predictors (multi-dimensional arrays) for regression or classification. We assume the predictors conditional on the response follow a quadratic exponential family in a generalized linear model, where the relation via a link is multi-linear. Using a multi-linear relation allows us to perform per-axis reductions that drastically reduce the total number of parameters in regressions with higher-order tensor-valued predictors. We derive maximum likelihood estimates for the multi-linear sufficient dimension reduction of the tensor-valued predictors. Furthermore, we provide an estimation algorithm which utilizes the tensor structure allowing efficient implementations. The performance of the method is illustrated via simulations and real-world examples.

E0808: **Robust kernel PCA by weighting observations**

Presenter: **Lauri Heinonen**, University of Turku, Finland

Co-authors: Joni Virta

A robust version of kernel principal component analysis is presented, which tolerates outliers. Observations are weighted by iteratively calculating the mean and covariance matrix in feature space, calculating a weight using a function of the Mahalanobis distance and using that in calculating the next mean and covariance. The method is closely connected to classical k-step M-estimates. All calculations are done in the feature space with the kernel matrix. The convergence of the weights is discussed. The results are illustrated with examples and compared to other relevant methods.

E0908: **Nonlinear feature extraction for sufficient dimension reduction in the presence of categorical predictors**

Presenter: **Andreas Artemiou**, Cardiff University, United Kingdom

Co-authors: Ben Jones

One often encounters regression and classification problems where there is a high-dimensional continuous predictor along with a set of categorical predictors. Existing approaches to nonlinear sufficient dimension regression are unable to accommodate the categorical predictors, while methods that do take them into account seek only linear combinations of the continuous predictors. For the first time, we provide a nonlinear sufficient dimension reduction method which can handle categorical predictors. This is achieved by first extending measure-theoretic developments in sufficient dimension reduction and then adapting a generalised kernel-based version of sliced inverse regression.

E0988: **Dimension reduction techniques for conditional quantiles**

Presenter: **Eliana Christou**, University of North Carolina at Charlotte, United States

Quantile regression (QR) is becoming increasingly popular due to its relevance in many scientific investigations. There is a great amount of work on linear and nonlinear QR models. Specifically, nonparametric estimation of the conditional quantiles received particular attention, due to its model flexibility. However, nonparametric QR techniques are limited in the number of covariates. Dimension reduction offers a solution to this problem by considering low-dimensional smoothing without specifying any parametric or nonparametric regression relation. The existing dimension reduction techniques focus on the entire conditional distribution. We, on the other hand, turn our attention to dimension reduction techniques for conditional quantiles and introduce a new method for reducing the dimension of the predictor X. We propose both linear and nonlinear dimension reduction techniques for conditional quantiles, extend to alternative types of data (such as categorical and longitudinal predictors), and also consider functional predictors.

E0407: **Non-linear two-dimensional PCA**

Presenter: **Joni Virta**, University of Turku, Finland

Co-authors: Andreas Artemiou

Non-linear principal component analysis for matrix-valued data is developed. Our approach is based on applying non-linear transformations separately to the left and right singular vectors of the observed matrices, guaranteeing that the estimated latent components enjoy the left-right structure typically expected in matrix dimension reduction. We treat both population and sample-level estimation and also establish the convergence rates of the estimators. The results are illustrated with numerical examples.

EO623 Room S-2.25 EXPLAINABLE ARTIFICIAL INTELLIGENCE Chair: Emanuela Raffinetti

E0659: **Can ESG shape the cost of capital: A bibliometric review and empirical analysis through ML**

Presenter: **Niklas Bussmann**, Universita degli Studi di Pavia, Italy

Co-authors: Alessandra Tanda, Ellen Pei-yi Yu

The purpose is to investigate how environmental, social, and governance (ESG) related behaviour affects and contributes to the cost of capital, taken as a proxy for firm riskiness. The literature has devoted much effort lately to understanding if and how ESG and sustainable behaviour of companies enters the determination of companies' evaluation and riskiness. The aim is (i) to understand and analyse how this matter has been tackled by the previous empirical studies in the field; (ii) to employ AI methodologies to explain the impact of ESG behaviour on the cost of capital. To reach our research objectives, we first employ a bibliometric review tool to highlight the key variables that affect the cost of capital and secondly, we apply the XGBoost algorithm and SHAP framework to a sample of more than 1400 multinational companies located worldwide.

E0685: **A SAFE artificial intelligence approach**

Presenter: **Emanuela Raffinetti**, University of Pavia, Italy

Co-authors: Paolo Giudici

The growing availability of data and computational power has allowed innovative developments in the field of Artificial Intelligence (AI). Nevertheless, the consideration of the possible adverse consequences of activities with a high societal impact has led policymakers and regulators to a degree of suspicion towards AI applications. This concern was also addressed in the recent regulation for a trustworthy AI: to be trustworthy, AI methodologies have to be SAFE. A SAFE application of AI must fulfil four key principles: it should be robust in terms of data and computations (Sustainability); it should lead to accurate predictions (Accuracy); it should not discriminate by population groups (Fairness); it should be human interpretable in terms of its drivers (Explainability). In agreement with the previous requirements, the aim is to provide a concrete response. Specifically, we combined the notion of explainability with those of accuracy and robustness through the formalization of new statistical methods based

on the Shapley values and the Lorenz Zonoid tool. By means of our proposal, the most explainable variables can be detected for different groups of observations, allowing to narrow of the set of data to be analysed and consequently reducing the computational effort.

E0707: **Multidimensional financial connectedness**
Presenter: **Paolo Pagnottoni**, University of Pavia, Italy
Co-authors: Alessandro Celani

The increasing availability of high and multidimensional data generated over time in finance has put severe limitations on standard approaches in multivariate time series econometric models. While it is common to model vectors of observations through standard vector time series analysis, the potential matrix form of data often reflects different types of structures of time series observations which can be further exploited to model interdependencies across financial securities. We propose a novel autoregressive model in a bilinear form which is able to: a) handle high and multidimensional financial data; b) yield enhanced interpretability given by the autoregressive model in matrix form; c) unveil dependencies across different sources of risks, i.e. price, volatility and liquidity risks. We illustrate the properties of our model through a real example of cryptocurrency market data.

E1156: **Countering racial discrimination in algorithmic lending: A case for model-agnostic interpretation methods**
Presenter: **Parvati Neelakantan**, Dublin City University, Ireland

In respect to racial discrimination in lending, the usefulness of Global Shapley Value and Shapley-Lorenz methods to attain algorithmic justice is examined. Using 157,269 loan applications during 2017 in New York from the Home Mortgage Disclosure Act data set, we confirm that these methods, consistent with the parameters of a logistic regression model, reveal prima facie evidence of racial discrimination. We show, critically, that these explainable AI methods can enable a financial institution to select an opaque creditworthiness model which blends out-of-sample performance with ethical considerations.

E1194: **A novel interpretation method for explaining machine learning survival models**
Presenter: **Yujia Chen**, University of Edinburgh Business School, United Kingdom
Co-authors: Raffaella Calabrese, Belen Martin-Barragan

Machine learning models such as tree-based ensemble methods or neural networks have been adapted to handle survival data and have shown superior predictive performance compared to traditional statistical approaches. However, the lack of interpretability restricts the adoption of these machine-learning models in survival analysis. Along these lines, a novel interpretation method is proposed for explaining machine learning survival models. It extends the framework of the popular interpretation method LIME by applying the joint model to approximate the machine learning survival model at the local scale of a test example. The proposed method explains a machine learning survival model through the linear combination of covariates included in the joint model, such that coefficients of the covariates can be regarded as quantitative impacts on the prediction. Besides, by using the joint model, the proposed method has the advantage of handling the endogenous time-varying covariate, which is critical to survival analysis.

EO122 Room S-1.01 RECENT ADVANCES IN MODEL SPECIFICATION TESTING Chair: Bojana Milosevic

E1392: **On approximating eigenvalues of covariance operators with applications to goodness-of-fit tests**
Presenter: **Bruno Ebner**, Karlsruhe Institute of Technology, Germany
Co-authors: Bojana Milosevic, Maria Dolores Jimenez-Gamero

Methods are reviewed to approximate eigenvalues of covariance operators connected to limiting Gaussian processes. We present a new method to approximate the largest eigenvalue connected to the Rayleigh-Ritz method and apply it to limit distributions of statistics of weighted L^2 type in several goodness-of-fit settings of distribution-free type or classical testing problems as testing for exponentiality or normality. Finally, we apply the methods in different settings and show the impact on efficiency statements of the asymptotic Bahadur type.

E1263: **Test for multivariate normality based on new characterization**
Presenter: **Marko Obradovic**, University of Belgrade, Serbia
Co-authors: Bojana Milosevic, Wiktor Ejsmont

The standard multivariate normal distribution is characterized by a certain linear combination being constant on a unit n-sphere. Based on this characterization, some normality tests are constructed. The main emphasis is on the null hypothesis of multivariate normal distribution with diagonal covariance matrix. We explore the asymptotic properties and perform a simulation study. We also consider the case of a general covariance matrix. The tests perform well in comparison to some popular powerful competitors. Potential applications are also discussed.

E1056: **Quantile-based MANOVA: A new tool for inferring multivariate data in factorial designs**
Presenter: **Marc Ditzhaus**, Otto-von-Guericke University Magdeburg, Germany
Co-authors: Marlene Baumeister, Markus Pauly

Multivariate analysis-of-variance (MANOVA) is well known and applied in all kinds of areas to examine multivariate endpoints. While classical approaches depend on restrictive assumptions like normality and homogeneity, there is a recent trend toward more general and flexible procedures. We proceed on this path but do not follow the typical mean-focused perspective in statistical inference. In contrast, we consider general quantiles, in particular the median, for a more robust analysis against outliers. The resulting, flexible methodology is shown to be asymptotically valid and consistent. The theoretical results are complemented by an extensive simulation study for small and moderate sample sizes, and by illustrative data analysis.

E1115: **Tests for the comparison of distributions of the excess over a confidence level**
Presenter: **Julian Gerstenberg**, Goethe University Frankfurt, Germany
Co-authors: Daniel Gaigall

The difference between the two populations can be assessed by the two sample Cramer-von-Mises distance. This idea is adopted for the comparison of distributions of the excess over a confidence level of two populations. With respect to the underlying sampling procedure, we distinguish between the independent samples case and the paired sample case. In both situations, we suggest and justify new tests for testing homogeneity of distributions of the excess over a confidence level.

E1232: **Goodness-of-fit testing of survival models in the presence of Type II right censoring**
Presenter: **Marike Cockeran**, North-West University, South Africa

A variety of tests is considered for testing goodness of fit in a parametric Cox proportional hazards (PH) in the presence of Type II right censoring. The testing procedures considered can be divided into two categories: an approach involving transforming the data to a complete sample and an approach using test statistics that can directly accommodate Type-II right censoring. The power of the proposed tests is compared through a Monte Carlo study for various scenarios. It is found that both approaches are useful for testing exponentiality if the censoring proportion in a data set is lower than 30%, but it is recommended to use the approach that first transforms the sample to a complete sample when one encounters higher censoring proportions.

EO058 Room S-1.04 COPULAS AND DEPENDENCE MODELLING Chair: Piotr Jaworski

E0255: **Bivariate vine copula based quantile regression with applications in climate data analysis**
Presenter: **Marija Tepegjozova**, Technical University Munich, Germany
Co-authors: Claudia Czado

The statistical analysis of univariate quantiles is a well-developed research topic. However, there is a profound need for research in multivariate quantiles. We tackle the topic of bivariate quantiles and bivariate quantile regression using vine copulas. They are graph theoretical models identified by a sequence of linked trees, which allow for separate modelling of marginal distributions and the dependence structure. We introduce a novel graph structure model (given by a tree sequence) specifically designed for a symmetric treatment of two responses in a predictive regression setting. We establish the computational tractability of the model and a straightforward way of obtaining different conditional distributions. Using vine copulas, the typical shortfalls of regression, such as the need for transformations or interactions of predictors, collinearity or quantile crossings, are avoided. We illustrate the copula-based bivariate quantiles for different copula distributions and provide a climate data set example. Further, the data example emphasizes the benefits of the joint bivariate response modelling in contrast to two separate univariate regressions or by assuming conditional independence, for bivariate response data set in the presence of conditional dependence.

E0277: **Singular value decomposition based low-rank representations of copulas**
Presenter: **Oliver Grothe**, Karlsruhe Institute of Technology, Germany
Co-authors: Jonas Rieger

Optimal low-rank approximations of arbitrary discretized (checkerboard) copulas are analyzed. Methodologically, we make use of truncated singular value decompositions of bistochastic matrices representing the copulas. The resulting (truncated) representations of the dependence structures are sparse, and memory usage decreases significantly. Due to the simple structure, essential statistical functionals of the copula's dependence structure are still readily available. The low-rank approximations conserve the uniform margins properties of the copulas but might lack non-negativity if the copula density has high peaks. For cases where non-negativity is crucial, we calculate the (Frobenius)-nearest valid discretized copula as a corrected low-rank representation. Copulas with stronger monotone dependence generally correspond to bistochastic matrices with larger ranks. Therefore, truncation leads to higher approximation errors than for copulas near the independence copula. We show how centering around the diagonals of the copulas compensates for this effect, leading to good low-rank representations in these cases as well. We illustrate the low-rank representation for various copula examples and families and derive some analytical results. We also discuss important general properties of the approximations and link our analysis to continuous decompositions of copula CDFs and copula-generating algorithms.

E0281: **Concordance measures generalizing Kendall's tau**
Presenter: **Martynas Manstavicius**, Vilnius University, Lithuania

Generalizations of a popular concordance measure, namely, Kendall's tau, to the bivariate case, are discussed. This is motivated by an open question left in previous work and the desire to have examples of polynomial-type concordance measures of any degree, which we hope will stimulate research on their characterization as was done previously for degree-one concordance measures.

E0624: **Dedekind-MacNeille completion of multivariate copulas via ALGEN method**
Presenter: **Matjaz Omladic**, Institute of Mathematics, Physics, and Mechanics, Slovenia

The problem of the Dedekind-MacNeille completion of the class of $d$copulas originated in 2005 and is well-known to the experts in the field. Unlike in the bivariate case, where the solution is just the class of 2-quasi-copulas, in the case that $d > 2$, this class is simply too big. The presented solution to the problem identifies an appropriate concrete subclass together with concrete meet and join operations so that the requirements for the desired completion are satisfied. The construction is made so that the induced order coincides with the starting pointwise order on d-quasi-copulas. However, this causes the two operations to be adjusted accordingly. The solution is based on a method called Algebraic Obstacles in the Geometry of Negative Volumes (ALGEN for short). This technique has been developed by the same authors to solve some questions in imprecise probability. Since the presented solution is based on a small subclass of d-quasi-copulas, a natural question arises whether this class is too big as the completion of d-copulas with respect to order suprema and infima. An answer to this question is also given.

E0635: **On certain bivariate copulas with a trapezoid support**
Presenter: **Piotr Jaworski**, University of Warsaw, Poland

A family of bivariate copulas given by: for $v+2u<2, C(u,v)=F(2F^{-1}(v/2)+F^{-1}(u))$, where F is a strictly increasing cumulative distribution function of a symmetric, continuous random variable, and for $v+2u\geq 2$, $C(u,v)=u+v-1$, is introduced. The basic properties and necessary conditions for the absolute continuity of C are discussed.

EO696 Room S-1.06 RECENT ADVANCES IN MIXTURE MODELS Chair: Konstantinos Perrakis

E1489: **Sparse estimation in heterogeneous varying coefficient regression models**
Presenter: **Abbas Khalili**, McGill University, Canada

Statistical methodologies are presented based on the regularized local-kernel likelihood for parameter estimation and feature selection in a sparse finite mixture of varying coefficient regressions. These models are commonly used to learn heterogeneous effects of features on a response variable where there is unobservable heterogeneity in data, and features' effects also vary according to an index variable such as time or location. Although complex, this situation frequently occurs in real data applications, which we demonstrate using a genetic data set. We will discuss the large-sample properties of the proposed methods, and we also evaluate their finite-sample performance via simulations. Finally, we will discuss the results of our real data analysis.

E1045: **Regularized joint mixture models**
Presenter: **Konstantinos Perrakis**, Durham University, United Kingdom
Co-authors: Thomas Lartigue, Frank Dondelinger, Sach Mukherjee

Regularized regression models are well studied and, under appropriate conditions, offer fast and statistically interpretable results. However, large data in many applications are heterogeneous in the sense of harboring distributional differences between latent groups. Then, the assumption that the conditional distribution of response Y given features X is the same for all samples may not hold. Furthermore, in scientific applications, the covariance structure of the features may contain important signals and its learning is also affected by latent group structure. We propose a class of mixture models for paired data (X,Y) that couples together the distribution of X (using sparse graphical models) and the conditional $Y|X$ (using sparse regression models). The regression and graphical models are specific to the latent groups and model parameters are estimated jointly (hence the name "regularized joint mixtures"). This allows signals in either or both of the feature distribution and regression model to inform learning of latent structure and provides automatic control of confounding by such structure. Estimation is handled via an expectation-maximization algorithm, whose convergence is established theoretically. We illustrate the key ideas via empirical examples. An R package is available on github.

E0316: **Dynamic mixture of finite mixtures of factor analysers with automatic inference on number of clusters and factors**
Presenter: **Margarita Grushanina**, WU Vienna University of Economics and Business, Austria
Co-authors: Sylvia Fruehwirth-Schnatter

Mixtures of factor analysers represent a popular tool for finding structure in data. While in most applications the number of clusters and latent

factors within clusters is fixed in advance, recently models with automatic inference on both the number of clusters and factors have been introduced. The automatic inference is usually done by assigning a nonparametric prior and allowing the number of clusters and factors potentially be infinite. The MCMC estimation is performed via an adaptive algorithm, in which the parameters associated with redundant factors are discarded as the chain moves. Besides its clear advantages, this approach also has drawbacks. Running a separate factor-analytical model for each cluster involves matrices of changing dimensions, which makes the model and programming cumbersome. Also, discarding the parameters associated with the redundant factors could lead to a bias in estimating cluster covariance matrices. The contribution to the MFA literature is to allow automatic inference on the number of clusters and factors while keeping both cluster and factor dimensions finite. Thus, some of the abovementioned drawbacks of infinite models are avoided. For the automatic inference on cluster structure, we employ the dynamic mixture of finite mixtures. Automatic inference on cluster-specific factors is performed via an extension of the cumulative shrinkage process prior to using its representation as an ordered version of the Indian buffet process.

E1338: **Spying on the prior of the number of data clusters and the partition distribution in Bayesian cluster analysis**
Presenter: **Jan Greve**, WU Vienna University of Economics and Business, Austria
Co-authors: Bettina Gruen, Gertraud Malsiner-Walli, Sylvia Fruehwirth-Schnatter

Recently in Bayesian Model-Based Clustering, the use of mixture models with an unknown number of clusters and/or components such as Dirichlet Process Mixtures (DPMs), Pitman-Yor Mixtures (PYMs) and Mixture of Finite Mixtures (MFMs) is getting increasingly common. A major empirical challenge involving these models is the characterisation of the prior on the partition space they each induce. An approach is introduced to compute descriptive statistics of the prior on the partitions for several influential mixture models employed in Bayesian Model-Based Clustering (specifically, DPMs and two classes of MFMs). The proposed methodology involves computationally efficient enumeration of the prior on the number of clusters and determining the first two prior moments of symmetric additive statistics characterising the partitions. The accompanying reference implementation is made available in the R package fipp. Finally, ongoing work to generalize this approach to a broader class of models is briefly mentioned.

EO452 Room S-1.27 MISSING DATA ANALYSIS AND ITS APPLICATION Chair: Wang Miao

E0704: **Causal and counterfactual views of missing data models**
Presenter: **Razieh Nabi**, Emory University, United States

It is often said that the fundamental problem of causal inference is a missing data problem – the comparison of responses to two hypothetical treatment assignments is made difficult because, for every experimental unit, only one potential response is observed. We consider the implications of the converse view: that missing data problems are a form of causal inference. We make explicit how the missing data problem of recovering the complete data law from the observed law can be viewed as the identification of a joint distribution over counterfactual variables corresponding to values had we (possibly contrary to fact) been able to observe them. Drawing analogies with causal inference, we show how identification assumptions in missing data can be encoded in terms of graphical models defined over counterfactual and observed variables. The validity of identification and estimation results using such techniques rely on the assumptions encoded by the graph holding true. Thus, we also provide new insights on the testable implications of a few common classes of missing data models, and design goodness-of-fit tests around them.

E1017: **A self-censoring model for multivariate nonignorable nonmonotone missing data**
Presenter: **Yilin Li**, Peking University, China
Co-authors: Wang Miao, Ilya Shpitser, Eric Tchetgen Tchetgen

An itemwise modeling approach, called self-censoring, is introduced for multivariate nonignorable non-monotone missing data, where the missingness process of each outcome is affected by its own value and is associated with missingness indicators of other outcomes, while conditionally independent of the other outcomes. The self-censoring model complements previous graphical approaches for the analysis of multivariate nonignorable missing data. It is identified under a completeness condition stating that any variability in one outcome can be captured by variability in the other outcomes among complete cases. For estimation, we propose a suite of semiparametric estimators, including doubly robust estimators that deliver valid inferences under partial misspecification of the full-data distribution. We also provide a novel and flexible global sensitivity analysis procedure anchored at the self-censoring. We evaluate the performance of the proposed methods with simulations and apply them to analyze a study about the effect of highly active antiretroviral therapy on preterm delivery of HIV-positive mothers.

E1236: **A stableness of resistance model for nonresponse adjustment with callback data**
Presenter: **BaoLuo Sun**, National University of Singapore, Singapore
Co-authors: Wang Miao, Xinyu Li

The survey world is rife with nonresponse, and in many situations, the missingness mechanism is not at random, which is a major source of bias for statistical inference. Nonetheless, the survey world is rich with paradata that track the data collection process. A traditional form of paradata is callback data that record attempts to contact. Although it has been recognized that callback data are useful for nonresponse adjustment, they have not been used widely in statistical analysis until recently. In particular, there have been a few attempts that use callback data to estimate response propensity scores, which rest on fully parametric models and fairly stringent assumptions. We propose a stableness of resistance assumption for identifying the propensity scores and the outcome distribution of interest, without imposing any parametric restrictions. We establish the semiparametric efficiency theory, derive the efficient influence function, and propose a suite of semiparametric estimation methods, including doubly robust ones, which generalize existing parametric approaches. Application to a Consumer Expenditure Survey dataset suggests an association between nonresponse and high housing expenditures.

E1446: **Semi-automated estimation of weighted rates for e-commerce catalog quality monitoring**
Presenter: **Mauricio Sadinle**, University of Washington, United States

E-commerce product catalogs are constantly evolving, and close monitoring of quality metrics is needed, which often requires identifying whether the product attributes contain defects. When such identification requires human auditing, catalog monitoring is extremely expensive to conduct frequently. We investigate approaches for tracking weighted rates over time, defined as the fraction of customer attention that goes to products with a certain defect. We assume that the gold standard for detecting such defects comes from human auditors, but to avoid collecting audited data at each point in time, we leverage automated procedures, such as classifiers. However, simply replacing human auditor decisions with automated predictions generally leads to large biases in the estimated weighted rates. We leverage automated procedures while obtaining approximately unbiased and low variance estimators of the rate of interest. We rely on being able to evaluate the quality of the automated procedure using audits at a baseline time or domain, and then extrapolate the performance of the procedure to the time or domain of interest. We perform extensive simulation studies to stress-test our proposed estimation approaches under a variety of scenarios representative of our actual use cases. Our proposed estimation approach is related to the task of quantification in machine learning, and so we draw connections throughout.

E1683: **General MANOVA with missing data: A resampling-based solution**
Presenter: **Lubna Amro**, TU Dortmund University, Germany
Co-authors: Markus Pauly, Burim Ramosaj

Repeated measure designs and split plot plans are widely employed in scientific and medical research. The analysis of such designs is typically based on MANOVA models, requiring complete data, and certain assumptions on the underlying parametric distribution, such as normality or covariance matrix homogeneity. Several nonparametric multivariate methods have been proposed in the literature. They overcome the distributional

assumptions, but the issue of missing data remains. The aim is to develop asymptotic correct procedures that are capable of handling missing values without assuming normality and allowing for covariance matrices that are heterogeneous between groups. This is achieved by applying a proper resampling method in combination with quadratic form-type test statistics. An extensive simulation study is conducted, exemplifying the tests for finite sample sizes under different missingness mechanisms. Finally, an illustrative data example is analyzed.

EO192 Room K0.16 SPECTRAL METHODS IN STATISTICAL NETWORK INFERENCE Chair: Vince Lyzinski

E0295: **Robust spectral clustering with rank statistics**
Presenter: **Joshua Cape**, University of Wisconsin, Madison, United States

Traditional non-robust approaches for spectral clustering and embedding exhibit severe performance degradation in the presence of outliers, heavy-tailed distributions, and heterogeneous noise variances. In this talk, we address these challenges by studying the problem of robust spectral clustering using rank statistics. We highlight ongoing work spanning methodology, theory, and applications, with a focus on statistical guarantees for user-friendly dimensionality reduction techniques.

E0706: **Matching embeddings via shuffled total least squares regression**
Presenter: **Daniel Sussman**, Boston University, United States

A frequently used approach for graph matching is first to embed the networks as points in Euclidean space and then match the embeddings. We consider the case that the two graphs have related but not identical distributions that necessitate a more complex alignment in the matching step. This is related to the problem known as shuffled linear regression. We consider a modified shuffled regression setting where there is noise in both the response and the predictor variables. This setting better matches the graph matching problem and we provide convergence rates for a shuffled total least squares method in terms of the normalized Procrustes quadratic loss.

E0385: **The importance of being correlated: Implications of dependence in joint spectral inference across multiple networks**
Presenter: **Konstantinos Pantazis**, Johns Hopkins University, United States
Co-authors: Avanti Athreya, Jesus Arroyo, William Frost, Evan Hill, Vince Lyzinski

Spectral inference on multiple networks is a rapidly-developing subfield of graph statistics. Recent work has demonstrated that joint, or simultaneous, spectral embedding of multiple independent networks can deliver more accurate estimation than individual spectral decompositions of those same networks. Little attention has been paid, however, to considering multiple networks with inherent correlation, and even less, to the network correlation that such joint embedding procedures naturally induce. We present a generalized omnibus embedding methodology and we provide a detailed analysis of this embedding across both independent and correlated networks. We also describe how this omnibus embedding can itself induce correlation which leads us to distinguish between inherent correlation—that is, the correlation that arises naturally in multisample network data—and induced correlation, which is an artifice of the joint embedding methodology. We show that the generalized embedding procedure is flexible and robust, and we prove both consistency and a central limit theorem for the embedded points. We examine how induced and inherent correlation can impact inference for network time series data, and we provide network analogues of classical questions such as the effective sample size for more generally correlated data. We construct an appropriately calibrated omnibus embedding that can detect changes in real biological networks that previous embedding procedures could not discern.

E1438: **Population-level inference for networks via graph embeddings**
Presenter: **Jesus Arroyo**, Texas A&M University, United States
Co-authors: Avanti Athreya, Vince Lyzinski

The problem of inferring population properties from observed network samples is considered. We approach this problem via dimensionality reduction by projecting the data onto a low-dimensional space. It is shown that this procedure can yield accurate inferences; however, in the presence of shared structure across the networks, classical approaches such as PCA are sub-optimal and can have reduced power. We show that a graph embedding that exploits a common low-rank structure can yield improvements, and we present a central limit theorem for the resulting network projections. Applications of this methodology are introduced in simulations and real data, including two-sample testing, anomaly detection, and network classification.

E0715: **Discovering underlying dynamics in time series of networks**
Presenter: **Avanti Athreya**, Johns Hopkins University, United States
Co-authors: Zachary Lubberts, Youngser Park, Carey Priebe

Understanding dramatic changes in the evolution of networks are central to statistical network inference, as underscored by recent challenges of predicting and distinguishing pandemic-induced transformations in organizational and communication networks. We consider a joint network model in which each node has an associated time-varying low-dimensional latent vector of feature data, and connection probabilities are functions of these vectors. Under mild assumptions, the time-varying evolution of the constellation of latent vectors exhibits a low-dimensional manifold structure under a suitable notion of distance. This distance can be approximated by a measure of separation between the observed networks themselves, and there exist consistent Euclidean representations for underlying network structure, as characterized by this distance, at any given time. These Euclidean representations permit the visualization of network evolution and transform network inference questions such as change-point and anomaly detection into a classical setting. We illustrate our methodology with real and synthetic data, and identify change points corresponding to massive shifts in pandemic policies in a communication network of a large organization.

EO250 Room K0.18 COUNT DATA MODELS: DEVELOPMENTS AND APPLICATIONS Chair: Jochen Einbeck

E0751: **Bayesian modelling of underreported count data**
Presenter: **Michaela Dvorzak**, Joanneum Research Forschungsgesellschaft mbH, Austria
Co-authors: Helga Wagner

Regression models for count data subject to underreporting in a Bayesian framework are considered. We specify a joint model for the data-generating process of true counts and the fallible reporting process, where the outcomes in both processes are related to a set of potential covariates. Identification of this joint model is achieved through additional information on the reporting process provided by validation data and incorporation of variable selection in both parts of the model. For posterior inference, an MCMC sampling scheme is implemented, which is based on data augmentation and auxiliary mixture sampling techniques. The proposed method is illustrated using simulated data and applied to a real data set.

E0465: **Testing for the generalized Poisson distributions**
Presenter: **Apostolos Batsidis**, University of Ioannina, Greece
Co-authors: Maria Dolores Jimenez-Gamero, Bojana Milosevic

The family of generalized Poisson (GP) distributions, which contain, among many others as special cases, the compound Poisson and Katz distributions, is a flexible family of distributions for modelling count data. The probability generating function (PGF) of the GP is the unique PGF satisfying a certain differential equation. This property leads us to propose and study a goodness-of-fit test for the family of GP distributions. The test is consistent against fixed alternatives, and its null distribution can be consistently approximated by a parametric bootstrap. The goodness of the bootstrap estimator and the power for finite sample sizes are numerically assessed. Apostolos Batsidis acknowledges support of this work by the project: Establishment of capacity building infrastructures in Biomedical Research (BIOMED-20) (MIS 5047236) which is implemented under the

Action Reinforcement of the Research and Innovation Infrastructure, funded by the Operational Programme: Competitiveness, Entrepreneurship and Innovation (NSRF 2014-2020) and co-financed by Greece and the European Union (European Regional Development Fund).

E0922: ***p*-values and quantiles for count data**

Presenter: **Paul Wilson**, University Of Wolverhampton, United Kingdom
Co-authors: Jochen Einbeck

For a given continuous test statistic T and observed value t, traditionally "right–tail" p–values are defined as $P_0(T \geq t)$, where P_0 is the distribution of T under the null hypothesis, "left–tail" p–values being defined analogously. However, this is not the only way of defining p–values. For discrete data, $\hat{\alpha}_{T,\lambda}(t) = P_0[T > t] + \lambda P_0[T = t]$ defines the "λ-p-value" $\lambda = 1$ corresponding to the "traditional p-value", and $\lambda = 0.5$ to the "mid p–value". Following similar lines of reasoning, one can motivate and define the λ-*quantile*. In a discrete setting, λ-p-values and quantiles may be shown to have superior properties to their traditional counterparts, for example, unlike traditional p-values, the expected value of mid-p-values is 0.5 under the null hypothesis, and their use in hypothesis tests with discrete test statistics may, at least in certain cases, be shown to lead to better power and attainment rates. We explore the use of λ-p-values and quantiles, discussing their advantages and disadvantages.

E1295: **Modelling the impact of Covid-19 pandemic on health insurance-related claims**

Presenter: **Amanda Fernandez-Fontelo**, Universitat Autonoma de Barcelona, Spain
Co-authors: Pedro Puig, Montserrat Guillen, David Morina

In many developed countries -including Spain- health insurance is one of the insurance categories with the greatest penetration in the market. In 2020 and 2021, the Covid-19 pandemic impacted the claim rate in health insurance, particularly for consultations and medical events that could be postponed. Mobility issues resulted in insurance usage by policyholders declining, and patient-health professional interactions dramatically changed due to a rise in telephone consultations. In the current work, we specifically investigate whether the number of claims made by policyholders during the Covid-19 pandemic was under-reported but over-represented after such a pandemic. For this purpose, we present a new model of integer-valued time series for either under-reporting or over-reporting identification and estimation. We apply our model to different time series of the number of claims made by policyholders for several pathologies.

E0275: **Time-varying dispersion integer-valued GARCH models**

Presenter: **Wagner Barreto-Souza**, University College Dublin, Ireland
Co-authors: Luiza Piancastelli, Konstantinos Fokianos, Hernando Ombao

A general class of INteger-valued Generalized AutoRegressive Conditionally Heteroskedastic (INGARCH) processes is proposed by allowing time-varying mean and dispersion parameters, which we call time-varying dispersion INGARCH (tv-DINGARCH) models. More specifically, we consider mixed Poisson INGARCH models and allow for dynamic modeling of the dispersion parameter (as well as the mean), similarly to the spirit of the ordinary GARCH models. We derive conditions to obtain first and second-order stationarity, and ergodicity as well. Estimation of the parameters is addressed, and their associated asymptotic properties are established as well. A restricted bootstrap procedure is proposed for testing constant dispersion against time-varying dispersion. Monte Carlo simulation studies are presented for checking point estimation, standard errors, and the performance of the restricted bootstrap approach. The inclusion of covariates is also addressed and applied to the daily number of deaths due to COVID-19 in Ireland. Insightful results were obtained in the data analysis, including a superior performance of the tv-DINGARCH processes over the ordinary INGARCH models.

EO532 Room K0.20 RECENT ADVANCES IN STATISTICAL MODELING IN GENETICS AND BIOLOGICAL RESEARCH Chair: Linxi Liu

E0349: **A unified quantile framework reveals nonlinear heterogeneous transcriptome-wide associations**

Presenter: **Tianying Wang**, Tsinghua University, China
Co-authors: Iuliana Ionita-Laza, Ying Wei

Transcriptome-wide association studies (TWAS) are powerful tools for identifying putative causal genes by integrating genome-wide association studies and gene expression data. Most existing methods are based on linear models and, therefore, may miss or underestimate nonlinear associations. We propose a robust, quantile-based, unified framework to investigate nonlinear transcriptome-wide associations in a quantile process manner. Through extensive simulations and the analysis of multiple psychiatric and neurodegenerative disorders, we showed that the proposed framework gains substantial power over conventional approaches and leads to insightful discoveries on nonlinear associations between gene expression levels and traits, thereby providing a complementary approach to existing literature. In doing so, we applied the proposed method for 797 continuous traits from the UK Biobank, and the results are available in a public repository.

E0392: **Integration of multidimensional splicing data and GWAS summary statistics for risk gene discovery**

Presenter: **Ran Tao**, Vanderbilt University Medical Center, United States

A common strategy for the functional interpretation of genome-wide association study (GWAS) findings has been the integrative analysis of GWAS and expression data. Using this strategy, many association methods have been successful in identifying trait-associated genes via mediating effects on RNA expression. However, these approaches often ignore the effects of splicing, which can carry as much disease risk as expression. Compared to expression data, one challenge to detecting associations using splicing data is the large multiple testing burden due to multidimensional splicing events within genes. Here, we introduce a multidimensional splicing gene (MSG) approach, which consists of two stages: 1) we use sparse canonical correlation analysis (sCCA) to construct latent canonical vectors (CVs) by identifying sparse linear combinations of genetic variants and splicing events that are maximally correlated with each other; and 2) we test for the association between the genetically regulated splicing CVs and the trait of interest using GWAS summary statistics. Simulations show that MSG has proper type I error control and substantial power gains over existing multidimensional expression analysis methods under diverse scenarios. When applied to the Genotype-Tissue Expression Project data and GWAS summary statistics of 14 complex human traits, MSG identified much more significant genes than existing approaches.

E0775: **Generative mediation models for microbiome data analysis**

Presenter: **Kris Sankaran**, University of Wisconsin, United States

Mediation methods add nuance to causal inferences, making it possible to attribute causal effects to specific intermediate changes. Recently, these methods have been applied to microbiome studies, where a mechanistic understanding has otherwise proven elusive. Strategies are introduced for adapting generative models of microbiome data – including topic, zero-inflated, and logistic-normal-multinomial models – to support mediation analysis. We will investigate the properties of these methods using a semi-synthetic simulation study and a posterior predictive visual analysis. Finally, we will study longitudinal data from a randomized control trial on the effects of a mindfulness intervention on gut microbiome composition, drawing mediators from host physiological, behavioral, and diet measurements. All code needed to reproduce the results is available in an accompanying R package.

E1020: **A simple method for removing batch effects from single-cell RNA-sequencing data**

Presenter: **Jun Li**, University of Notre Dame, United States
Co-authors: Dailin Gan

Integrative analysis of multiple single-cell RNA-sequencing datasets allows for more comprehensive characterizations of cell types. Still, systematic technical differences between datasets, known as "batch effects", need to be removed before integration to avoid misleading interpretation of the data. Although many batch-effect-removal methods have been developed, there is still a large room for improvement: most existing methods only

give dimension-reduced data instead of expression data of individual genes, are based on computationally-demanding models, and are black-box models and thus difficult to interpret or tune. We present a new batch-effect-removal method called SCIBER and study its performance on real datasets. SCIBER matches cell clusters across batches according to the overlap of their differentially expressed genes. As a simple algorithm that has better scalability to data with a large number of cells and is easy to tune, SCIBER shows comparable and sometimes better accuracy in removing batch effects on real datasets compared to the state-of-the-art methods, which are much more complicated. Moreover, SCIBER outputs the expression of individual genes, which can be used directly for downstream analyses. Additionally, SCIBER is a reference-based method, which assigns one of the batches as the reference batch and keeps it untouched during the process, making it especially suitable for integrating user-generated datasets with standard reference data.

E1562: **Accelerated algebraic functions for statistical genomics**
Presenter: **Alexander Freudenberg**, University of Mannheim, Germany

Genomic datasets used in empirical research are steadily growing in size and advances in computing power can only partially offset the associated computational demand. A suboptimal utilization of hardware resources can lead to significant increases in costs and computation times. We explore the benefits of highly finetuned GPU routines for the calculation of important population statistics, which utilize instruction sets of modern NVIDIA hardware. We showcase examples as part of our implementation in the R package miraculix.

EO232 Room K0.50 NEW CHALLENGES IN DESIGN OF EXPERIMENTS Chair: John Stufken

E1527: **Subdata selection with a large number of variables**
Presenter: **John Stufken**, George Mason University, United States
Co-authors: Rakhi Singh

With ever larger datasets, computational challenges have led to a vast literature on using only some of the data (subdata) for estimation or prediction. This raises the question of how subdata should be selected from the entire dataset (full data). One possibility is to select the subdata randomly from the full data, but this is typically not the best method. The literature contains various suggestions for better alternatives. Most of these alternatives focus on situations where the number of variables is small to modest. We introduce a method that can be used for big data with a large number of variables in the context of linear regression.

E1528: **Deriving nearly-optimal subdata**
Presenter: **Min Yang**, University of Illinois at Chicago, United States

Big data brings the unprecedented challenge of analyzing such data due to its extraordinary size. One strategy for analyzing such massive data is data reduction. Instead of analyzing the full dataset, a selected subdata set is analyzed. Various subdata selection methods have been proposed. While the trade-off between computation complexity and statistical efficiency has been studied, little is known about how efficient the selected subdata is in terms of statistical efficiency. To answer this question, we need to find an optimal subdata. Deriving an optimal subdata, however, is an N-P hard problem. A novel framework to derive a nearly-optimal subdata, under any given statistical model, regardless of optimality criterion or parameters of interest, will be introduced. This framework has three benefits: (i) it shows us the structure of a nearly-optimal subdata for any given full data under various set-ups (model, optimality criterion, parameter of interest); (ii) it measures highly accurate statistical efficiency; and (iii) it provides a tool of deriving a nearly optimal subset in active learning where statistical efficiency is the main concern.

E1529: **Predictive subdata selection for large-scale deterministic computer models**
Presenter: **Ming-Chung Chang**, Academia Sinica, Taiwan

Computer models are implementations of complex mathematical models using computer codes. Tremendous amounts of data generated from computer models are becoming ubiquitous owing to advanced technology. Such data richness, however, may yield an inability to conduct statistical analysis in terms of time cost. Recently, increased attention has focused on solving this data reduction problem. We will introduce a new subdata selection method for large-scale deterministic computer models. The proposed method takes advantage of the information of the output values and adaptively updates the current subdata with affordable computational cost. Simulated examples and real data analyses are provided.

E1612: **Optimal design for penalized estimators**
Presenter: **Jonathan Stallrich**, North Carolina State University, United States
Co-authors: Maria Weese, Kade Young, Byran Smucker, David Edwards

The experimental design community has gravitated towards penalized estimation (e.g., Dantzig selector and lasso) to analyze data from screening experiments that may assume factor sparsity. However, the optimal design framework is largely based on statistical properties of unpenalized least-squares estimation. The purpose is to review the current theory relating design properties to the support recovery properties of the lasso and Dantzig selector. An optimal design framework is then proposed that better tailors design selection to maximizing the probability of support recovery. A local optimality approach is considered first and demonstrates that the optimal design assuming all positive model coefficients is not orthogonal. We then use the proposed framework to justify popular heuristic measures for constructing supersaturated designs in which the number of runs is less than the number of factors. A more robust optimal design framework is then considered that assumes less knowledge of the true model. To optimize the computationally-expensive criteria, we propose a construction algorithm that starts with many efficient designs according to the heuristic measures.

EO204 Room S0.03 SPATIAL STATISTICS Chair: Soutir Bandyopadhyay

E0855: **A graphical lasso model for Hermitian matrices to detect global time-lagged teleconnections**
Presenter: **Indranil Sahoo**, Virginia Commonwealth University, United States
Co-authors: Joseph Guinness, Brian Reich

Teleconnections refer to spatially and temporally connected large-scale anomalies that influence the variability of atmospheric phenomena. Since teleconnections influence the global climate system, it is important to understand the abnormal behavior and interactions of these phenomena and identify them accurately. We provide a mathematical definition of teleconnections based on a spatio-temporal model using spherical needlet functions. Spherical needlets are exactly localized at several overlapping intervals corresponding to different frequencies in the frequency domain and form a tight frame. This ensures the perfect reconstruction property of an orthonormal basis. We also extend the famous graphical Lasso algorithm to incorporate Hermitian matrices and use it to estimate the inverse covariance matrix of needlet coefficients after projecting them onto the Fourier domain. The proposed method is demonstrated by simulation studies and detection of possible global teleconnections in the HadCM3 model output air temperature data.

E1607: **Machine learning in geo-spatial mixed models**
Presenter: **Abhi Datta**, Johns Hopkins Bloomberg School of Public Health, United States

Spatial generalized linear mixed models, consisting of a linear covariate effect and a Gaussian Process (GP) distributed spatial random effect, are widely used for analyses of geospatial data. We consider the setting where the covariate effect is non-linear and propose modeling it using a flexible machine learning algorithm like random forests or deep neural networks. We propose well-principled extensions of these methods, for estimating non-linear covariate effects in spatial mixed models where the spatial correlation is still modeled using GP. The basic principle is guided by how ordinary least squares extend to generalized least squares for linear models to account for dependence. We demonstrate how the same extension can

be done for these machine learning approaches like random forests and neural networks. We provide extensive theoretical and empirical support for the methods and show how they fare better than nave or brute-force approaches to use machine learning algorithms for spatially correlated data. We demonstrate the RandomForestsGLS R-package that implements this extension for random forests.

E1630: **Regridding uncertainty for statistical downscaling of solar radiation**
Presenter: **Soutir Bandyopadhyay**, Colorado School of Mines, United States
Co-authors: Douglas Nychka, Maggie Bailey

As the photovoltaic (PV) industry moves to extend plant lifetimes to 50 years, the changing climate may have an effect on PV production and assumptions that current solar radiation patterns are representative of the future may not be appropriate. A key step in aiding the prediction of PV production is projecting solar radiation for future years based on a changing climate. This involves downscaling future climate projections for solar radiation to spatial and temporal resolutions that are useful for building PV plants. Initial steps in downscaling involve being able to closely predict observed data from regional climate models (RCMs). This prediction requires (1) regridding RCM output from their native grid on differing spatial resolutions to a common grid in order to be comparable to observed data and (2) bias correcting solar radiation data, via quantile mapping, for example, from climate model output. The uncertainty associated with (1) is not always considered for downstream operations in (2). This uncertainty, which is not often shown to the user of a regridded data product, is examined. This analysis is applied to data from the National Solar Radiation Database housed at the National Renewable Energy Lab, and a case study of the mentioned methods in California is presented.

E0214: **A nonparametric approach to deal with spatio temporal quantile regression problems**
Presenter: **Soudeep Deb**, Indian Institute of Management Bangalore, India
Co-authors: Subhrajyoty Roy, Claudia Neves

The focus is on the quantile regression problems in spatio-temporal data. Such data appears in many economic and environmental applications. We develop a nonparametric technique that requires minimal assumptions and, in fact, can be used even when explicit information on the location is not available. Detailed asymptotic theory of our method is then derived. We also show the usefulness of the method through an extensive simulation study. Finally, a real-life application of electricity demand forecasting is carried out.

EO541 Room S0.12 ROBUSTNESS AND RELATED TOPICS I Chair: Ana Maria Bianco

E0296: **On robustness for spatio-temporal data**
Presenter: **Alfonso Garcia-Perez**, Universidad Nacional de Educacion a Distancia (UNED). Department of Statistics, Spain

The spatio-temporal variogram is an important factor in spatio-temporal analysis because it is the key element in kriging prediction. However, the traditional spatio-temporal variogram estimator, which is commonly used for this purpose, is extremely sensitive to outliers. We address this problem in two different ways. First, defining new robust spatio-temporal variogram estimators, which are defined as M-estimators or trimmed estimators of an original data transformation, estimators for which we obtain accurate approximations for their distributions. Second, we compare the classical estimate against a robust one, identifying spatio-temporal outliers in this way. In these two approaches, we use a multivariate scale-contaminated normal model framework. In the contribution, we also define and study a new class of M-estimators and include real-world applications. We finally determine whether there are significant differences in the spatio-temporal variogram between two temporal lags reducing, if so, the number of lags considered in the spatio-temporal analysis.

E0303: **Robust estimation in mixture models under case-wise and cell-wise contamination**
Presenter: **Claudio Agostinelli**, University of Trento, Italy
Co-authors: Giovanni Saraceno, Ayanendranath Basu

Classical contamination models consider as possible outliers the statistical units, that is, the whole observations (case-wise contamination). This approach has limits when it is applied to data sets with a large number of variables. A more general approach to robustness is to consider the possible presence of both case-wise and cell-wise contamination. The last, also known as the independent contamination model, identified as possible outliers in the cells. One approach is to filter out the contaminated cells from the data set and then apply a robust procedure able to handle case-wise outliers and missing values. We introduce filters in any dimension based on statistical data depth functions, and we show how we can use them in the robust estimation of parameters in a finite mixture model.

E0771: **A robust alternative to the Poisson hurdle model**
Presenter: **Conceicao Amado**, Universidade de Lisboa, Portugal
Co-authors: Manuela Souto de Miranda

Hurdle Poisson models are mixed models that can deal with an excess of zeroes by considering two separate components, namely, a binary process and a truncated discrete distribution. They are particularly adequate for modelling counting processes when the occurrence of zero observations does not depend on the main generating process of strictly positive counting. It is often the case when counting low-probability incidents, as they appear in some health econometric statistics, or extreme-value events per unit of time. Maximum likelihood is used to fit the Poisson hurdle model under exact and strict stochastic assumptions. However, the performance of these estimators degrades when these assumptions are not verified. The aim is to compare robust estimators, particularly considering minimum distance estimators for the parameters of the hurdle model when the positive component is modelled by a truncated Poisson. We investigate the performance of the estimators through a simulation study.

E0807: **Direct methodology for adjusting ROC curves: A robust approach**
Presenter: **Ana Maria Bianco**, Universidad de Buenos Aires, Argentina
Co-authors: Jesica Charaf

Receiver Operating Characteristic (ROC) curve is a graphical tool that became a key tool for evaluating a diagnostic test based on a continuous marker. In practice, several factors, such as age, gender or blood pressure, may improve the discriminatory ability of the marker. When this is the case, it seems wise to assess the possible covariates effects in the ROC analysis to avoid oversimplification. To incorporate this additional information, we follow the direct method, where the effect of the covariates is directly evaluated on the ROC curve by means of a generalized linear model and pseudo-observations. In this framework, the estimator of the ROC curve is obtained through a stepwise procedure. The aim is twofold. On the one hand, we illustrate the instability of the classical method to estimate the conditional ROC curve in presence of outliers, and we also provide a robust alternative. The proposal combines robust estimators of the coefficients of the involved parametric models with an adaptive weighted empirical estimator. Through a Monte Carlo study, we compare the performance of the proposed estimators with that of the classical ones in clean and contaminated samples.

E1745: **Robust and efficient Breusch-Pagan test: A beta-score LM test for heteroscedasticity in linear regression models**
Presenter: **Nirian Martin**, Complutense University of Madrid, Spain

In Econometrics, the Breusch-Pagan test-statistic has become an iconic application of the Lagrange multipliers (LM) test, also recognized as the Raos scores test for composite null hypotheses. We shall introduce beta-score LM tests for heteroscedasticity in linear regression models, for which the degree of robustness and efficiency is regulated through a non-negative tuning parameter, being $\beta = 0$ the classical Breusch-Pagan test-statistic, the most efficient one under absence of outliers. A very elegant expression is obtained, with a similar interpretation to the one for the classical case. The test statistic is constructed by extending the methodology from identically distributed to nonidentically distributed individuals, for composite null hypotheses. Detailed theoretical justifications for robustness and efficiency properties are given. A simulation study illustrates the finite-sample

behaviour of several Breusch-Pagan beta-score LM test statistics.

EO793 Room S0.13 SPORT ANALYTICS Chair: Andreas Groll

E0601: **A varying coefficient state-space model for investigating betting behaviour within in-play markets**
Presenter: **Marius Oetting**, Bielefeld University, Germany
Co-authors: Rouven Michels, Roland Langrock

The potential effects of in-game dynamics on betting behaviour are investigated. Considering two comprehensive data sets from the 2017/18 Bundesliga season comprising in-play betting volumes and match events, we use state-space models to analyse the dynamics and drivers of betting volumes. Within this state-space framework, we use (penalised) B-splines to model the potentially time-varying effect of in-game dynamics as implied by measurable events such as shots and passes. Preliminary results suggest that volumes in the in-play market are driven by such in-game dynamics and that this effect varies throughout a match.

E0680: **Statistical and machine learning methods in modern football analysis**
Presenter: **Alexander Gerharz**, TU Dortmund University, Germany

In today's world, professional football clubs are collecting lots of data. The reasons for this are manifold, but the main reasons are (i) to evaluate the fitness of the players, (ii) to scout players from other clubs and (iii) to analyze tactics. The ultimate goal for every football club is to be successful and each of the mentioned reasons can contribute to this goal: it is important to keep your best players healthy, to find the best new players for your club and to help the coaches find the winning tactic for the upcoming game. In recent years, especially technological advancements for generating and collecting data in professional football have risen. To create useful knowledge from this mass of data, the techniques for statistical data analysis have also advanced, and new approaches have been developed.

E0995: **Survival analysis in basketball: An analysis of the NBA players' offensive performance**
Presenter: **Ambra Macis**, University of Brescia, Italy
Co-authors: Marica Manisera, Paola Zuccolotto

Over the years, statistics have been widely used in sport analytics. Many questions have been answered concerning, for example, player performance, outcome prediction of a match or a tournament and factors identifying successful and unsuccessful teams. Moreover, many studies have been developed to study athletes' dropout rate or their return to sport after an injury. These last studies have been developed through survival analysis. However, up to now, to the best of our knowledge, survival methods have not been used for analyzing player performance. The offensive performance of NBA players is investigated in terms of the amount of points gained during a season segment using statistical methods for survival analysis. More in detail, the available players' statistics in the pre-All-Stars game segment have been used as baseline covariates. Then, the post-All-Stars game season segment has been observed to analyze whether the players exceeded or not a given amount of points and how many minutes they played for experiencing or not the event of interest. The final aim is to identify the features associated with the probability of exceeding the given threshold during a fixed time period.

E1136: **Estimating the risk of time-loss injuries in football players through recurrent time-to-event methods**
Presenter: **Lore Zumeta-Olaskoaga**, BCAM - Basque Center for Applied Mathematics, Spain
Co-authors: Andreas Bender, Helmut Kuechenhoff, Dae-Jin Lee

Sports injury prevention research has gained increased interest in professional sports, including professional football. Players are constantly exposed to high competition demands, and subsequently, they are repeatedly exposed to injury risk, which greatly impacts on their individual and team performance. Thus modelling and understanding the injury occurrence is important to help to prevent them, in addition, to maximising players' performance. The aim is to estimate the risk of time-loss injuries in an elite male football team participating in LaLiga. We propose the use of piece-wise exponential additive mixed models for modelling such data and for studying the correlation between recurrent events (injuries) and within-player variability in injury risk.

E1032: **Plus-minus a couple of millions: a model for transfer fee evaluation**
Presenter: **Arne Maes**, BNP Paribas Fortis, Belgium
Co-authors: Christophe Ley, Dries Goossens, Lars Magnus Hvattum, Senthil Murugan

Soccer players contribute to their respective teams both on and off the field, to the extent that clubs are willing to pay a significant fee to purchase the contractual rights of a player from another club. Tough negotiations usually take place about the size of this fee, but what is a good price? Our models estimate transfer fees for players using detailed information on the player, his remaining contract duration, and the selling and buying club. We believe these models can give practitioners a better estimate of the expected transfer fee a player would command. As such, it can be used to guide negotiations, and it allows for more sophisticated planning of investment decisions. Finally, it may serve as a tool for transfer clearing houses in their battle against fraud, allowing them to focus their attention on those transfers where the fee differs substantially from what was predicted. We improve on the existing literature in three ways. First of all, we include domain knowledge to identify niches of the transfer market where a model-based approach can add most value. Secondly, we combine various data sources better to capture the financial and sportive aspects of each transfer. Lastly, we consider several feature-selection (Baruto, Barutoshap, and Statistical FS) and modelling techniques (Decision Tree, Linear Regression, XGBoost, and RF) in a structural fashion to arrive at the most suitable specified models.

EO627 Room Virtual R02 RECENT DEVELOPMENTS IN STATISTICAL MACHINE LEARNING Chair: Fei Xue

E0672: **Bayesian analysis of multiway data with applications to nba game analysis**
Presenter: **Weining Shen**, UC Irvine, United States

A Bayesian nonparametric matrix clustering approach is proposed to analyze the latent heterogeneity structure in the shot selection data collected from professional basketball players in the National Basketball Association (NBA). The proposed method adopts a mixture of finite mixtures framework and fully utilizes the spatial information via a mixture of matrix normal distribution representation. We propose an efficient Markov chain Monte Carlo algorithm for posterior sampling that allows simultaneous inference on both the number of clusters and the cluster configurations. We also establish large-sample convergence properties for the posterior distribution. The compelling empirical performance of the proposed method is demonstrated via simulation studies and an application to shot chart data from selected players in the NBAs 20172018 regular season.

E0696: **Efficient learning of optimal individualized treatment rules**
Presenter: **Weibin Mo**, Purdue University, United States

Recent development in data-driven decision science has seen great advancement in individualized decision-making. Existing methods typically require the initial estimation of some nuisance models. To protect consistency from nuisance model misspecification, the double robustness property has been widely advocated, while the concern of estimation efficiency is rarely studied. Efficiency is critical for stable and reliable predictions, as well as high power to justify treatment benefits. To improve the efficiency of the estimated individualized treatment rule (ITR), we propose an Efficient Learning (E-Learning) framework for finding an optimal ITR in the multi-categorical treatment setting. We establish the optimality of the proposed E-Learning in the presence of regression model misspecification and heteroscedasticity.

E0772: **A non-asymptotic framework for approximate message passing**
Presenter: **Yuting Wei**, University of Pennsylvania, United States

Approximate message passing (AMP) emerges as an effective iterative paradigm for solving high-dimensional statistical problems. However, prior AMP theory — which focused mostly on high-dimensional asymptotics — fell short of predicting the AMP dynamics when the number of iterations surpasses $o(\log n/\log\log n)$ (with n the problem dimension). To address this inadequacy, a non-asymptotic framework is developed for understanding AMP in spiked matrix estimation. Built upon a new decomposition of AMP updates and controllable residual terms, we lay out an analysis recipe to characterize the finite-sample behavior of AMP in the presence of an independent initialization, which is further generalized to allow for spectral initialization. As two concrete consequences of the proposed analysis recipe: (i) when solving Z_2 synchronization, we predict the behavior of spectrally initialized AMP for up to $O(n/\mathrm{poly}\log n)$ iterations, showing that the algorithm succeeds without the need of a subsequent refinement stage (as conjectured recently); (ii) we characterize the non-asymptotic behavior of AMP in sparse PCA (in the spiked Wigner model) for a broad range of signal-to-noise ratio.

E1287: **The power of contrast for feature learning: A theoretical analysis**
Presenter: **Linjun Zhang**, Rutgers University, United States

Contrastive learning has achieved state-of-the-art performance in various self-supervised learning tasks and even outperforms its supervised counterpart. Despite its empirical success, the theoretical understanding of why contrastive learning works is still limited. We provably show that contrastive learning outperforms autoencoder, a classical unsupervised learning method, for both feature recovery and downstream tasks. Moreover, we also illustrate the role of labeled data in supervised contrastive learning. This provides theoretical support for recent findings that contrastive learning with labels improves the performance of learned representations in the in-domain downstream task, but it can harm the performance in transfer learning. We verify our theory with numerical experiments.

E1709: **Continuous-time recommender system with evolutionary temporal feature process embedding**
Presenter: **Xiwei Tang**, University of Virginia, United States

Large volumes of temporal event data are drawing increasing attention in a wide variety of applications, such as in analyzing social media data, healthcare records, online consumption, and product recommendation. For the recommender system, traditional models based on static latent features or discretized time epochs usually fail to capture the important temporal dynamics in user-item interactions. We propose a novel evolutionary recommender system by leveraging the temporal mechanism on the continuous-time user-item interactive events. The proposed approach can effectively capture the long- and short-term preferences from the sequential historical data with informative dynamic feature embeddings. We develop an efficient algorithm for learning the model parameters with outstanding scalability and computational effectiveness. Using both synthetic and real-world datasets, we show the outperformance of the proposed model in learning sequential user behaviors and achieving better predictive power in the recommendation.

EO615 Room Virtual R03 RECENT ADVANCES IN CAUSAL INFERENCE AND HIGH-DIMENSIONAL STATISTICS Chair: Lin Liu

E0423: **Optimal transport weights for causal inference**
Presenter: **Eric Dunipace**, UCLA, United States

Imbalance in covariate distributions leads to biased estimates of causal effects. Weighting methods attempt to correct this imbalance but rely on specifying models for the treatment assignment mechanism, which is unknown in observational studies. This leaves researchers to choose the proper weighting method and the appropriate covariate functions for these models without knowing the correct combination to achieve distributional balance. In response to these difficulties, we propose a nonparametric generalization of several other weighting schemes found in the literature: Causal Optimal Transport. This new method directly targets distributional balance by minimizing optimal transport distances between any source and target population. Our approach is semiparametrically efficient and model-free but can also incorporate any important functions of covariates that a researcher desires to balance. Moreover, we show how this method can provide nonparametric imputations of the missing potential outcomes and give rates of convergence for this estimator. Finally, optimal transport methods also allow researchers diagnosis when distributions are not balanced. We find that Causal Optimal Transport outperforms competitor methods when both the propensity score and outcome models are misspecified.

E1031: **Two-sample testing of high-dimensional linear regression coefficients via complementary sketching**
Presenter: **Fengnan Gao**, Fudan University and SCMS, China
Co-authors: Tengyao Wang

A new method is introduced for two-sample testing of high-dimensional linear regression coefficients without assuming that those coefficients are individually estimable. The procedure works by first projecting the matrices of covariates and response vectors along directions that are complementary in sign in a subset of the coordinates, a process which we call 'complementary sketching'. The resulting projected covariates and responses are aggregated to form two test statistics, which are shown to have essentially optimal asymptotic power under a Gaussian design when the difference between the two regression coefficients is sparse and dense, respectively. Simulations confirm that our methods perform well in a broad class of settings. An application to a large single-cell RNA sequencing dataset demonstrates its utility in the real world.

E1209: **Long-term causal inference under persistent confounding via data combination**
Presenter: **Yuhao Wang**, Tsinghua University, China
Co-authors: Guido Imbens, Nathan Kallus, Xiaojie Mao

The identification and estimation of long-term treatment effects are studied when both experimental and observational data are available. Since the long-term outcome is observed only after a long delay, it is not measured in the experimental data, but only recorded in the observational data. However, both types of data include observations of some short-term outcomes. We uniquely tackle the challenge of persistent unmeasured confounders, i.e., some unmeasured confounders that can simultaneously affect the treatment, short-term outcomes and the long-term outcome, noting that they invalidate identification strategies in previous literature. To address this challenge, we exploit the sequential structure of multiple short-term outcomes, and develop three novel identification strategies for the average long-term treatment effect. We further propose three corresponding estimators and prove their asymptotic consistency and asymptotic normality. We finally apply our methods to estimate the effect of a job training program on long-term employment using semi-synthetic data. We numerically show that our proposals outperform existing methods that fail to handle persistent confounders.

E1229: **Superoptimal regimes for decision-making assisted by algorithms**
Presenter: **Mats Stensrud**, Ecole polytechnique federale de Lausanne, Switzerland

Healthcare providers desire to implement decision rules that, when applied to individuals in the population of interest, yield the best possible outcomes. For example, the current focus on precision medicine reflects the search for individualized treatment decisions, adapted to a patient's characteristics. We will introduce superoptimal regimes, which are guaranteed to outperform conventional optimal regimes. Importantly, identification of superoptimal regimes and their values require exactly the same assumptions as identification of conventional optimal regimes in several common settings, including instrumental variable settings. The superoptimal regimes can also be identified in data fusion contexts, in which experimental data and (possibly confounded) observational data are available. We will present two examples that have appeared in the optimal regimes literature, illustrating that the superoptimal regimes perform better than conventional optimal regimes.

E1744: **Estimating and improving individualized treatment rules and dynamic treatment regimes with an instrumental variable**
Presenter: **Bo Zhang**, Fred Hutchinson Cancer Center, United States

Estimating individualized treatment rules (ITRs) and dynamic treatment regimes (DTRs) from retrospective observational data is challenging as some degree of unmeasured confounding is often expected. We develop a framework for estimating properly defined "optimal" ITRs and DTRs with a possibly time-varying instrumental variable (IV) when unmeasured covariates confound the treatment and outcome, rendering the potential outcome distributions possibly partially identified. We define a generic class of estimands (termed IV-optimal ITRs/DTRs)and study the associated estimation problem. We then extend the IV-optimality framework to tackle the policy improvement problem, delivering IV-improved ITRs/DTRs that are guaranteed to perform no worse and potentially better than a pre-specified baseline ITR/DTR. Importantly, our IV-improvement framework opens up the possibility of strictly improving upon DTRs that are optimal under the no unmeasured confounding assumption (NUCA). We demonstrate via extensive simulations the superior performance of IV-optimal and IV-improved ITRs/DTRs over the ITRs/DTRs that are optimal only under the NUCA. In a real data example, we embed retrospective observational registry data into a natural, two-stage experiment with noncompliance using a differential-distance-based, time-varying IV and estimate useful IV-optimal DTRs that assign mothers to a high-level or low-level neonatal intensive care unit based on their prognostic variables.

EO234 Room Virtual R04 RESAMPLING METHODS IN MODERN SETTINGS Chair: Miles Lopes

E0691: **Bootstrapping Whittle estimators**

Presenter: **Efstathios Paparoditis**, University of Cyprus, Cyprus

Co-authors: Jens-Peter Kreiss

Fitting parametric models by optimizing frequency domain objective function is an attractive approach to parameter estimation in time series analysis. Whittle estimators are a prominent example in this context. Under weak conditions and the assumption that the true spectral density of the underlying process does not necessarily belong to the parametric class of spectral densities fitted, the distribution of Whittle estimators typically depends on the difficulty in estimating characteristics of the underlying process. This makes the implementation of asymptotic results for the construction of confidence intervals or for assessing the variability of estimators, difficult in practice. A frequency domain bootstrap method is proposed to estimate the distribution of Whittle estimators, which is asymptotically valid under assumptions that not only allow for possible model misspecification but also for weak dependence conditions which are satisfied by a wide range of stationary stochastic processes. Adaptions of the bootstrap procedure developed to incorporate different modifications of Whittle estimators proposed in the literature, like, for instance, tapered, de-biased or boundary-extended Whittle estimators, are also considered. Simulations demonstrate the capabilities of the bootstrap method proposed and its good finite sample performance. A real-life data analysis also is presented.

E0814: **High-dimensional CLT with general covariance structure**

Presenter: **Yuta Koike**, University of Tokyo, Japan

Co-authors: Xiao Fang

The focus is on the problem of bounding the normal approximation error over rectangles for a sum of n independent d-dimensional random vectors. We aim to establish such a bound with poly-logarithmic dependence on the dimension d. It is known that such a bound is available with a nearly $n^(-1/2)$ convergence rate when the covariance matrix of the sum is non-degenerate. We show that, under some additional distributional assumptions such as log-concavity, we can derive error bounds with nearly $n^(-1/2)$ convergence rates and poly-log dependence on d without any restriction on the covariance matrix. We also discuss whether these improved normal approximation error rates can be transferred to bootstrap approximation.

E1205: **A cheap bootstrap method for fast inference**

Presenter: **Henry Lam**, Columbia University, United States

A bootstrap methodology is presented that uses minimal computation in terms of resampling effort, namely as low as one Monte Carlo replication, while maintaining desirable statistical guarantees. We present the theory of this method that uses a simple twist from the standard bootstrap principle. We illustrate how this methodology can be used for fast inference across different estimation problems, and its relevance and generalizations, especially to large-scale statistical problems and computational simulation.

E1298: **Inference for high-dimensional exchangeable arrays**

Presenter: **Harold Chiang**, University of Wisconsin-Madison, United States

Co-authors: Kengo Kato, Yuya Sasaki

Inference for high-dimensional separately and jointly exchangeable arrays is considered where the dimensions may be much larger than the sample sizes. For both exchangeable arrays, we first derive high-dimensional central limit theorems over the rectangles and subsequently develop novel multiplier bootstraps with theoretical guarantees. These theoretical results rely on new technical tools such as Hoeffding-type decomposition and maximal inequalities for the degenerate components in the Hoeffiding-type decomposition for the exchangeable arrays. We exhibit applications of our methods to uniform confidence bands for density estimation under joint exchangeability and penalty choice for l1-penalized regression under separate exchangeability. Extensive simulations demonstrate precise uniform coverage rates. We illustrate this by constructing uniform confidence bands for international trade network densities.

E1434: **Statistical inference for streaming PCA in high dimensions**

Presenter: **Robert Lunde**, Washington University in St Louis, United States

Co-authors: Purnamrita Sarkar, Rachel Ward

The problem of quantifying uncertainty is considered for the estimation error of the leading eigenvector from Oja's algorithm for streaming principal component analysis, where the data are generated IID from some unknown distribution. By combining classical tools from the U-statistics literature with recent results on high-dimensional central limit theorems for quadratic forms of random vectors and concentration of matrix products, we establish a weighted chi-squared approximation result for the sin-squared error between the population eigenvector and the output of Oja's algorithm. Under certain structural assumptions on the covariance matrix, we show that the error of the weighted chi-squared approximation goes to zero even when $p >> n$. Furthermore, to facilitate statistical inference, we propose a multiplier bootstrap algorithm that may be updated in an online manner. We establish conditions under which the bootstrap distribution is close to the corresponding sampling distribution with high probability, thereby establishing the bootstrap as a consistent inferential method in an appropriate asymptotic regime.

EO733 Room Virtual R05 CAUSAL INFERENCE: CHALLENGES IN COMPLEX SETTINGS Chair: Laura Pazzagli

E1396: **Selection bias and multiple inclusion criteria**

Presenter: **Ingeborg Waernbaum**, Uppsala University, Sweden

Spurious associations between an exposure and outcome not describing the causal estimand of interest can be the result of the selection of the study population. Recently, sensitivity parameters and bounds have been proposed for selection bias, along the lines of sensitivity analysis previously proposed for bias due to unmeasured confounding. The basis for the bounds is that the researcher specifies values for sensitivity parameters describing associations under additional identifying assumptions. We extend the previously proposed bounds to give additional guidance for practitioners to construct i) the sensitivity parameters for multiple selection variables and ii) an alternative assumption-free bound, producing only logically feasible values. The results show that the assumption-free bounds can be both smaller and larger than the previously proposed bounds and, therefore, can serve as an indicator of settings when the former bounds do not produce feasible values. We derive the bounds in a study of perinatal

risk factors for childhood-onset type 1 diabetes mellitus where the selection of the study population was made by multiple inclusion criteria. It may be difficult for the researcher to give plausible input values for the sensitivity parameters for selection bias under multiple selection and to provide further guidance for practitioners, we provide a data learner in R where both the sensitivity parameters and the assumption free bounds are implemented.

E0803: **Elaborated ontologies for causal inference in resource-limited settings**
Presenter: **Aaron Sarvet**, EPFL, Switzerland

Emerging scarcity requires new policies for triaging limited resources. However, common-sense counterfactual targets are often impossible to articulate under standard causal models. We will briefly review these standard causal models and discuss their limitations. Then, to make progress, we will elaborate a general potential outcomes-based framework for evaluating the effects of strategies for allocating a fixed supply of limited resources in a longitudinal setting. We will provide non-parametric conditions that allow the identification of counterfactual outcomes from the observation of a single cluster ($n = 1$) of patients, and motivate semi-parametric estimators based on likelihood ratio weights. As an illustration, we will consider the estimation of survival under counterfactual rules for ventilator triage (including both initiation and termination) in an intensive care unit over the course of a COVID-19 epidemic.

E1010: **Causal inference methods for multiple treatment group evaluations**
Presenter: **Hongwei Zhao**, Texas A&M University, United States

Causal inference methods are discussed for comparing treatment effects when multiple treatment groups are present. Key ideas of causal inference and the potential outcome framework will be reviewed, and several propensity score-based methods will be considered. Additionally, machine-learning methods will be applied to increase the flexibility of the model. Simulation studies will be conducted to compare the consistency and efficiency of different causal inference methods. Finally, these methods will be demonstrated using a real example where multiple treatment groups are involved.

E0927: **Causal inference with competing events**
Presenter: **Jessica Young**, Harvard Medical School and Harvard Pilgrim Health Care Institute, United States

In failure-time settings, a competing risk event is any event that makes it impossible for the event of interest to occur. For example, cardiovascular disease death is a competing event for prostate cancer death because an individual cannot die of prostate cancer once he has died of cardiovascular disease. Various statistical estimands have been posed in the classical competing risks literature. These include the cause-specific hazard, subdistribution hazard, marginal hazard, cause-specific cumulative incidence and marginal cumulative incidence. We will place these estimands within a counterfactual framework for causal inference in order to define, interpret and identify counterfactual contrasts in each of these estimands under different treatment interventions in a given study. We discuss limitations in the interpretation of these existing estimands when a causal treatment effect on the event of interest is the goal, and the treatment affects the competing event. Finally, we introduce separable effects for causal inference, which overcome these interpretational limitations, coincide with effects often cited to justify the clinical relevance of an analysis of path-specific effects and that rely only on assumptions that are testable in a future experiment.

E2019: **Single time-series conditional causal effect**
Presenter: **Ivana Malenica**, Harvard University, United States
Co-authors: Mark van der Laan

Causal estimands are defined in a single time-series setup with observational data. We consider a sequential setting, where at each time t, a data record $O(t)$ is observed, which consists of treatments $A(t)$, outcomes $Y(t)$, and time-varying covariates $W(t)$. We assume that the conditional distribution of $O(t)$ can be described by a function $Co(t)$ only depending on a fixed dimensional summary of the past. Intensive longitudinal data is collected on a single individual, where data recorded at t carries information about a causal effect of treatment on the proximal outcome defined by $Co(t)$. Our approach allows the estimation of a broad class of estimands, including a class of summaries of the conditional causal parameters defined by the current context over time. The proposed target parameters are pathwise differentiable with an efficient influence function that is doubly robust. We propose a targeted maximum likelihood estimator (TMLE) of these causal parameters, and present results on the asymptotic consistency and normality of the TMLE. The limit distribution of the proposed estimator is characterized under a sequential Donsker condition, and expressed in terms of a notion of bracketing entropy adapted to martingale settings. Our methodology is inspired by financial and health care applications (e.g., chronic disease monitoring), where data is collected frequently over time, and holds immense promise for improving prevention and treatment allocation.

EO280 Room BH (SE) 1.01 TIME SERIES WITH CHANGES IN REGIME Chair: Maddalena Cavicchioli

E0254: **Inference for Markov switching GARCH(1,1) models using sequential Monte Carlo**
Presenter: **Feng Chen**, UNSW Syd, Australia
Co-authors: Damien Wee, William Dunsmuir

Markov switching (MS-)GARCH(1,1) models allow for structural changes in volatility dynamics between a finite number of regimes. Since the regimes are not observed, computation of the likelihood requires integrating over an exponentially increasing number of regime paths, which is intractable. An existing smooth likelihood estimation procedure for sequential Monte Carlo (SMC), which is currently limited to hidden Markov models with a one-dimensional state variable, is modified to enable likelihood estimation and maximisation for MS-GARCH(1,1) models, a model which requires two dimensions, volatility and regime, to evolve its hidden state process. Furthermore, the modified SMC procedure is shown to be easily adapted to fitting MS-GARCH(1,1) models even when there are missing observations. The proposed methodology is validated with simulated data and is also illustrated with an analysis of two financial time series, the daily returns on the S&P 500 index and on the Henry Hub natural gas spot price, with the latter series containing a gap caused by the shutdown in response to Hurricane Rita in 2005.

E0289: **Regime switching processes and their long time behaviors**
Presenter: **Abhishek Pal Majumder**, University of Reading, United Kingdom

Regime-switching processes have proved to be indispensable in the modeling of various phenomena in econometrics and physical science, allowing model parameters that traditionally were considered to be constant to fluctuate in a Markovian manner in line with empirical findings. We study diffusion processes of the Ornstein-Uhlenbeck type where the drift and diffusion coefficients are functions of an underlying Markov process with a stationary distribution on a countable state space. The exact long-time behavior of the process is determined for the three regimes corresponding to the expected drift strictly greater, equal or strictly less than zero, respectively. The time asymptotic behaviors are naturally expressed in terms of solutions to the well-studied distributional affine fixed-point equation $X = AX + B$ in law, where X is independent of (A, B). Additional applications will be discussed with findings in terms of Cox-Ingersoll-Ross diffusion, Geometric Brownian motions under Markovian and semi-Markovian environments. Long-term behaviors change in the transient cases for semi-Markov switching due to the difference in the tail behaviors of the sojourn times. Diffusion processes of Ornstein-Uhlenbeck types are continuous versions of the well-known auto-regressive processes so that the results can be translated for discrete time series as well.

E1264: **Time series forecasting using hybrid regime switching ANN models**
Presenter: **Jie Cheng**, Keele University, United Kingdom

Regime-switching models have been widely used for both economic and financial time series. However, their out-of-sample forecasting perfor-

mance is frequently inferior to simple benchmark models for standard loss functions. The purpose is to determine whether improvements can be achieved in forecasting performance by using a hybrid regime-switching model with a set of Machine Learning techniques and incorporating financial and/or economic variables. The empirical results with some real data sets indicate the effectiveness of the new combinatorial model in obtaining more accurate forecasting as compared to existing regime-switching models.

E0963: **Three states of the French business cycle**
Presenter: **Anna Petronevich**, Bank of France, France
Co-authors: Catherine Doz
The aim is to infer the current state of the French business cycle and nowcast the French GDP using Markov Switching Dynamic Factor Model. The multifactor model is estimated on a large database of macroeconomic series. We show that the first factor approximates the French business cycle and is best described as having three states - expansion, mild recession and severe recession, with severe recessions characterized as sharp but brief deterioration of the economic situation. The resulting smoothed probabilities provide a timely dating of past French business cycles and detect well the waves of the Covid-19 crisis. Additionally, we find that the second factor describes the global business cycle. We observe that global recessions often spill over on the French economy in recent history, illustrating the high integration of France into the world economy. Incorporating the two factors into the nowcasting equation allows producing accurate nowcasts of the French GDP.

E0428: **Impulse response analysis in Markov switching vector autoregressions**
Presenter: **Maddalena Cavicchioli**, University of Modena and Reggio Emilia, Italy
The regime-dependent impulse response functions are exactly derived for a Markov switching vector autoregression (VAR) model in terms of neat matrix expressions in closed form. The key is to recognize that the latent first-order Markov switching process in the model has a VAR(1) representation, and that the model can be cast into a state-space form. Using such a representation, the regime-dependent impulse response function analysis can be processed with respect to either an asymmetric discrete shock or to a symmetric continuous shock.

EO718 Room BH (S) 2.03 EMPIRICAL PROCESSES AND THEIR APPLICATIONS Chair: Eric Beutner

E0861: **Testing monotonicity of mean potential outcomes in a continuous treatment with high-dimensional data**
Presenter: **Ying-Ying Lee**, University of California, Irvine, United States
Co-authors: Martin Huber, Yu-Chin Hsu, Chu-An Liu
While most treatment evaluations focus on binary interventions, a growing literature also considers continuously distributed treatments. We propose a Cramer-von Misestype test for testing whether the mean potential outcome given a specific treatment has a weakly monotonic relationship with the treatment dose under a weak unconfoundedness assumption. In a nonseparable structural model, applying our method amounts to testing monotonicity of the average structural function in the continuous treatment of interest. To flexibly control for a possibly high-dimensional set of covariates in our testing approach, we propose a double-debiased machine learning estimator that accounts for covariates in a data-driven way. We show that the proposed test controls asymptotic size and is consistent against any fixed alternative. These theoretical findings are supported by the Monte-Carlo simulations. As an empirical illustration, we apply our test to the Job Corps study and reject a weakly negative relationship between the treatment (hours in academic and vocational training) and labor market performance among relatively lowtreatment values.

E0894: **Estimation of systemic risk in semi-parametric dynamic models based on the empirical distribution of residuals**
Presenter: **Jean-Michel Zakoian**, CREST, France
Co-authors: Loic Cantin, Christian Francq
In semi-parametric conditional location-scale models, commonly used systemic risk measures (such as the CoVaR) involve conditional moments and conditional quantiles of the joint distribution of two innovation processes. To estimate the latter quantiles, we rely on the estimation of the empirical joint distribution of two residuals series. We establish stochastic equicontinuity of the empirical joint cdf, allowing us to prove the consistency and asymptotic normality of estimators of the innovations quantiles. As an application, asymptotic confidence intervals for systemic risk measures are derived. Numerical illustrations based on simulated and real data will also be presented.

E1429: **Recursive quantile estimation: Non-asymptotic confidence bounds**
Presenter: **Likai Chen**, Washington University in Saint Louis, United States
Co-authors: Georg Keilbar, Wei Biao Wu
The recursive estimation of quantiles is considered using the stochastic gradient descent (SGD) algorithm with Polyak-Ruppert averaging. The algorithm offers a computationally and memory-efficient alternative to the usual empirical estimator. The focus is on studying the non-asymptotic behavior by providing exponentially decreasing tail probability bounds under mild assumptions on the smoothness of the density functions. This novel non-asymptotic result is based on a bound of the moment-generating function of the SGD estimate. We apply our result to the problem of best-arm identification in a multi-armed stochastic bandit setting under quantile preferences.

E1768: **Two-sample smooth test for the equality of distributions for dependent data**
Presenter: **Eric Beutner**, Vrije Universiteit Amsterdam, Netherlands
The two-sample problem is considered, i.e. testing whether two cumulative distribution functions are equal. Various proposals have been made to construct test statistics for this testing problem. These proposals include kernel-based methods, characteristic function type approaches, Cramer-von Mises type statistics, energy statistics and modifications of Neyman's smooth test. Common to these approaches is that they assume the two samples to be independent. We focus on modifications of Neyman's smooth test and extend them to dependent data. Bootstrapping the test statistics is also considered.

EO442 Room BH (S) 2.05 BAYESIAN MODELING AND APPLICATIONS Chair: Christopher Hans

E1202: **Spatio-temporal Bayesian modeling of crime in Philadelphia**
Presenter: **Shane Jensen**, The Wharton School of the University of Pennsylvania, United States
Urban data analysis has been recently improved through publicly available high-resolution data, allowing us to investigate theories in criminology and urban design empirically. We will focus on a particular direction: spatial-temporal modeling of the change in crime over the past decade in the city of Philadelphia. We will explore different parametric and non-parametric Bayesian approaches for finding regions of the city that share similar crime dynamics. Within this context, we have developed a methodology for non-parametric clustering of regions simultaneously across multiple levels of spatial resolution. We will also provide an interpretation of our results in the context of the geography and built environment of the city of Philadelphia.

E1371: **TD-CARMA: Painless, accurate, and scalable estimates of gravitational-lens time delays with flexible CARMA processes**
Presenter: **Antoine Meyer**, Imperial College London, France
Co-authors: David van Dyk, Aneta Siemiginowska
The gravitational field of a galaxy can act as a lens and deflect the light emitted by a more distant object such as a quasar. Strong gravitational lensing causes multiple images of the same quasar to appear in the sky. Cosmological parameters encoding our current understanding of the expansion history of the Universe can be constrained by accurate estimation of time delays. We propose TD-CARMA, a Bayesian method to estimate cosmological time delays by modelling the observed and irregularly sampled light curves as realizations of a CARMA process. Our

model accounts for heteroskedastic measurement errors and microlensing, a source of independent extrinsic long-term variability. The CARMA formulation admits a linear state-space representation, allowing for efficient and scalable likelihood computation via the Kalman Filter. We obtain a sample from the joint posterior distribution using nested sampling. This allows for painless Bayesian Computation, dealing with the expected multi-modality of the posterior distribution in a straightforward manner and not requiring starting values for the time delay, unlike existing methods. In addition, the proposed sampling procedure automatically evaluates the Bayesian evidence, allowing us to perform principled Bayesian model selection. TD-CARMA is parsimonious, and typically includes no more than a dozen unknown parameters.

E0990: **Distance-to-set priors for Bayesian constraint modeling**
Presenter: **Jason Xu**, Duke University, United States

Distance-to-set penalties provide a flexible way to incorporate a broad range of constraints for tasks cast as an optimization problem, especially as they apply to majorization-minimization (MM) algorithms. However, their use in statistical modeling is largely limited, and few results are available pertaining to inference after obtaining a point estimate. We consider a class of distance-to-set priors to facilitate constrained Bayesian inference. We draw connections between the existing MM and optimization approaches and Bayesian constraint relaxation, and we show that this class of priors has desirable theoretical properties for constrained Bayesian inference. Moreover, we elucidate why distance-to-set priors are particularly amenable to gradient-based sampling algorithms and can succeed in sampling the posterior in situations in which one is limited in the ability to relax the constraints. Finally, we demonstrate our results in various simulated and real-world settings.

E1459: **Density regression with Bayesian additive regression trees**
Presenter: **Alexander Volfovsky**, Duke University, United States

Flexibly modeling how a density changes with covariates is an important but challenging generalization of mean and quantile regression. While existing methods for density regression primarily consist of covariate-dependent discrete mixture models, we consider a continuous latent variable model in general covariate spaces, which we call DR-BART. The prior mapping of the latent variable to the data is constructed via a novel application of BART. We prove that the posterior induced by our model concentrates quickly around true generative functions that are sufficiently smooth. We also analyze DR-BART's performance on a set of challenging simulated examples, where it outperforms various other methods for Bayesian density regression. Lastly, we apply DR-BART to a U.S. census dataset to study returns to education. Our proposed sampler is efficient and allows one to take advantage of BARTs flexibility in many applied settings where the entire response distribution is of interest. Furthermore, our scheme for splitting on latent variables within BART facilitates its application to other models that can be described via latent variables, such as those involving hierarchical or network data.

EO374 Room K2.31 (Nash Lec. Theatre) METHODOLOGY FOR SEMIPARAMETRIC AND CAUSAL INFERENCE Chair: Charles Doss

E0498: **Disentangling confounding and nonsense associations due to dependence**
Presenter: **Elizabeth Ogburn**, Johns Hopkins University, United States

Nonsense associations can arise when an exposure and an outcome of interest exhibit similar patterns of dependence. Confounding is present when potential outcomes are not independent of treatment. The purpose is to describe how confusion about these two phenomena results in shortcomings in popular methods in three areas: causal inference with multiple treatments and unmeasured confounding; causal and statistical inference with social network data; and causal inference with spatial data. For each of these three areas, we will demonstrate the flaws in existing methods and describe new methods that were inspired by careful consideration of dependence and confounding.

E0971: **Debiased nonparametric inference for covariate-adjusted regression functions**
Presenter: **Ted Westling**, University of Massachusetts Amherst, United States
Co-authors: Kenta Takatsu

The problem of obtaining valid nonparametric inference for a covariated-adjusted (also known as G-computed) regression function with a continuous scalar exposure is discussed. We propose a debiased local linear estimator, and demonstrate that this estimator converges pointwise to a mean-zero normal limit distribution. We use this result to construct asymptotically valid pointwise confidence intervals for function values and differences thereof. Finally, we use recent finite-sample approximation results for the suprema of empirical processes to construct asymptotically valid uniform confidence bands, highlighting, in particular, the technical challenge associated with obtaining faster-than-usual rates of convergence for an empirical process remainder term.

E1968: **Adversarial Monte Carlo meta-learning of conditional average treatment effects**
Presenter: **Alex Luedtke**, University of Washington, United States

The meta-learning of conditional average treatment effect estimators is framed as a search for an optimal strategy in a two-player game. In this game, nature selects a prior over distributions that generate labeled data consisting of covariates, treatment, and an associated outcome, and the estimator observes data sampled from a distribution drawn from this prior. The estimator's objective is to learn a function that maps from a new feature to an estimate of the conditional average treatment effect. We establish that, under reasonable conditions, the estimator's has an optimal strategy that is equivariant to shifts and rescalings of the outcome and is invariant to permutations of the observations and to shifts, rescalings, and permutations of the features. We introduce a neural network architecture that satisfies these properties.

E0592: **Optimal estimation of heterogeneous causal effects**
Presenter: **Edward Kennedy**, Carnegie Mellon University, United States

Estimation of heterogeneous causal effects - i.e., how effects of policies and treatments vary across units - is fundamental to medical, social, and other sciences, and plays a crucial role in optimal treatment allocation, generalizability, subgroup effects, and more. Many methods for estimating conditional average treatment effects (CATEs) have been proposed in recent years. Still, there have remained important theoretical gaps in understanding if and when such methods make optimally efficient use of the data at hand. This is especially true when the CATE has a nontrivial structure (e.g., smoothness or sparsity). Work across two recent papers in this context is surveyed. First, we study a two-stage doubly robust estimator and give a generic model-free error bound, which, despite its generality, yields sharper results than those in the current literature. The second contribution is aimed at understanding the fundamental statistical limits of CATE estimation. We resolve this long-standing problem by deriving a minimax lower bound, with a matching upper bound obtained via a new estimator based on higher-order influence functions. Applications in medicine and political science are considered.

EO174 Room K2.40 RECENT ADVANCES IN STATISTICS FOR HEALTH Chair: Sophie Dabo

E0805: **Adaptive functional principal components analysis**
Presenter: **Sunny Wang**, ENSAI, France
Co-authors: Valentin Patilea

Kernel estimators are built for the mean and the covariance functions of functional data, and they are used for functional PCA. The random trajectories are, not necessarily differentiable, have unknown, possibly non-constant regularity, and are measured with possibly heteroscedastic error, at discrete design points. We propose specific bandwidth rules for the eigenvalues and the eigenfunctions, respectively. The bandwidth adapts to the local regularity of the trajectories, and minimises the mean squared error between our eigenelements' estimates and the ideal ones,

which would be obtained if the curves were observed in continuous time, without noise. They can be applied with both sparsely or densely sampled curves, are easy to calculate and update, and perform well in simulations. Simulations illustrate the effectiveness of the new approach.

E1024: **A Bayesian shared-frailty spatial scan statistic model for time-to-event data**
Presenter: **Camille Frevent**, University of Lille, France
Co-authors: Mohamed Salem Ahmed, Sophie Dabo, Michael Genin

Spatial scan statistics are well-known and widely used methods for the detection of spatial clusters of events. In the field of spatial analysis of time-to-event data, several models of scan statistics have been proposed. However, these models do not take into account the potential intra-unit spatial correlation of individuals nor a potential correlation between spatial units. To overcome this problem, we propose a scan statistic based on a Cox model with shared frailty that takes into account the spatial correlation between spatial units. In simulation studies, we have shown that (i) classical models of spatial scan statistics for time-to-event data fail to maintain the type I error in the presence of intra-spatial unit correlation, and (ii) our model performs well in the presence of both intra-spatial unit correlation and inter-spatial unit correlation. Our method has been applied to epidemiological data and the detection of spatial clusters of mortality in patients with end-stage renal disease in northern France.

E1300: **The patient pathway in a hospital environment**
Presenter: **Rim Essifi**, INRIA, France
Co-authors: Sophie Dabo, Cristian Preda, Christophe Biernacki

European healthcare systems are faced with multiple challenges, including an aging population, an increase in chronic diseases and patients with multi-morbidity, and limited financial and human resources. The response to these challenges is based, in particular, on the organization of care into care pathways. Namely, once the data necessary for the construction of a care pathway are acquired and processed, one has to model the patient pathway mathematically in a generic way. After that, using clustering algorithms, one can identify patients' subgroups, then, mine for common treatments, predict the future of patient pathways and answer clinicians' questions. All these steps would lead to an automated process which has to be evaluated by medical experts. Available statistical methods remain limited and inefficient in constructing care pathways. Indeed, data obtained from health care providers and insurance companies are all time-dependent, highly heterogeneous, qualitative in part, with several thousand possible modalities and mainly made up of missing data. We propose an approach based on functional data analysis combined with longitudinal data analysis in order to construct care pathways.

E1361: **Classification of multivariate functional data (defined on different domains) using PLS approach**
Presenter: **Issam-Ali Moindjie**, Inria, France

Classification of multivariate functional data is of interest. We propose two classification methods using the partial least squares approach relying on partial least squares regression. The first one uses the equivalence between linear discriminant analysis and linear regression. The second is a decision tree based on the first method. Moreover, we prove that multivariate functional components can be estimated using the univariate counterparts. This offers an alternative way to calculate PLS for multivariate functional data (potentially defined on different domains). Simulation studies and real data applications show that the proposed methods are competitive with linear discriminant on principal components scores and black-boxes models. Since they give interpretable results, our methods are suitable for sensitive application areas such as medicine, and biology.

EO539 Room K2.41 NEW ADVANCEMENTS IN UNDIRECTED GRAPHICAL/NETWORK MODELING Chair: Andriette Bekker

E0415: **Graphical ridge with sparsity in high-dimension**
Presenter: **Azam Kheyri**, University of Pretoria, South Africa
Co-authors: Andriette Bekker, Mohammad Arashi

The focus is on the estimation of the precision matrix in a high-dimensional Gaussian graphical model. Since the accuracy of the maximum likelihood approach can be improved by penalization, we consider the elastic net type penalty, which combines the L1 and L2 penalties while taking the target matrix into estimation consideration. We suggest a novel 2-step estimator that combines alternative ridge and graphical lasso estimators to improve precision estimation. Numerical results support our findings and demonstrate the superior efficiency of our proposal compared to the alternatives.

E0570: **Fully symmetric graphical lasso for dependent data**
Presenter: **Saverio Ranciati**, Universita di Bologna, Italy
Co-authors: Alberto Roverato, Alessandra Luati

A method is proposed to analyze multivariate data with intrinsic symmetrical structures and, in general, to solve problems belonging to the class of dependent samples inference, such as case-control studies, matched and paired data. To this aim, we propose the fully symmetric graphical lasso, a penalized likelihood method with a fused type penalty function that takes into explicit account the natural symmetrical structure within and between symmetrical blocks of the data (or samples). The implementation leverages an alternating directions method of multipliers algorithm to solve the corresponding convex optimization problem. The procedure is applied to various real-world datasets, concerning air pollution and brain fMRI scans.

E0584: **Distributionally robust formulation of the graphical Lasso**
Presenter: **Sang-Yun Oh**, University of California, Santa Barbara, United States
Co-authors: Alexander Petersen, Pedro Cisneros-Velarde, Chau Tran

Building on a recent framework for distributionally robust optimization, the inverse covariance matrix estimation is considered for multivariate data. We provide a novel notion of a Wasserstein ambiguity set specifically tailored to this estimation problem. A Special case includes penalized likelihood estimator for Gaussian data, specifically the graphical lasso estimator. As a consequence of this formulation, the radius of the Wasserstein ambiguity set is directly related to the regularization parameter in the estimation problem. Taking advantage of this finding, we develop a simple algorithm to determine a regularization parameter for the graphical lasso, using only the bootstrapped sample covariance matrices, avoiding repeated evaluation of the graphical lasso algorithm during regularization parameter tuning, for example, with cross-validation. We also establish a theoretical connection between the confidence level of graphical model selection via the DRO formulation and the asymptotic family-wise error rate of estimating false edges.

E1321: **Network-based clustering of pancancer data accounting for clinical covariates**
Presenter: **Fritz Bayer**, ETH Zurich, Switzerland
Co-authors: Giusi Moffa, Niko Beerenwinkel, Jack Kuipers

Cancer progresses in diverse ways leading to a heterogeneous landscape of mutations within and across cancer types. This heterogeneity is a considerable challenge for precision medicine, and the task of leveraging genomic data to predict survival and treatment outcomes. We focus on learning the diverse probabilistic relationships among mutations and clinical covariates. We propose a novel network-based clustering method that allows us to learn distinct mutational patterns while accounting for covariate effects. Using probabilistic graphical models, we cluster the mutations and covariates based on their distinct probabilistic relationships. Since the covariates should not drive the clustering of mutational patterns but are necessary to accurately model the mutations, we propose a covariate-adjusted clustering framework. Our framework allows us to detach the effects of covariates on the clustering, by exploiting causal relationships among the variables. Over a broad range of simulations, we demonstrate that our method outperforms standard clustering methods in correlated data. We apply our method to a large-scale genomic dataset, including

the mutational profiles and clinical covariates of 8085 patients, where we identify novel clusters based on mutational patterns. These clusters are significantly predictive of survival beyond clinical information and could serve as biomarkers for targeted treatment.

E1550: **Sparse graphs based on exchangeable random measures: properties, models and examples**
Presenter: **Francois Caron**, University of Oxford, United Kingdom
Co-authors: Francesca Panero, Judith Rousseau, Adrien Todeschini, Xenia Miscouridou

Random simple and multigraph models based on exchangeable random measures, also called graphex processes or generalised graphon models, have recently been proposed as a flexible class of sparse random graph models. This class of models can be seen as a generalisation of the popular graphon models. We will present this class of models, and discuss some of their asymptotic properties, in particular, the asymptotic behaviour of the degree distribution. We will also present some particular models within this class and their use for discovering latent communities in sparse real-world networks.

EC813 Room S-1.22 SURVIVAL ANALYSIS II Chair: Stefanie Biedermann

E0533: **Estimation in the Cox survival regression model with covariate measurement error and a changepoint**
Presenter: **Sarit Agami**, The Hebrew University of Jerusalem, Israel

The Cox regression model is a popular model for analyzing the relationship between a covariate vector and a survival endpoint. The standard Cox model assumes a constant covariate effect across the entire covariate domain. However, in many epidemiological and other applications, the covariate of main interest is subject to a threshold effect: a change in the slope at a certain point within the covariate domain. Often, the covariate of interest is subject to some degree of measurement error. The measurement error correction is discussed in the case where the threshold is known. Several bias correction methods are examined: two versions of regression calibration (RC1 and RC2, the latter of which is new), two methods based on the induced relative risk under a rare event assumption (RR1 and RR2, the latter of which is new), a maximum pseudo partial likelihood estimator (MPPLE), and simulation-extrapolation (SIMEX). The theoretical properties of these methods and a simulation comparing the methods are discussed. An illustrative example of the relationship between chronic air pollution exposure to particulate matter PM10 and fatal myocardial infarction (Nurses Health Study (NHS)) is presented.

E0203: **Smoothed bootstrap method for double-censored data**
Presenter: **Asamh Al Luhayb**, Qassim University, Saudi Arabia

A smoothed bootstrap method is introduced for double-censored data based on a generalization of Hill's A(n) assumption. The smoothed bootstrap method is compared to Efron's method for double-censored data through simulations. The comparison is conducted in terms of the coverage of percentile confidence intervals for the quartiles. From the study, it is found that the smoothed bootstrap method mostly performs better than Efron's method, in particular for small data sets. We also illustrate the use of the method for survival function inference.

E0895: **Regression analysis with censored covariates in the presence of a cured fraction**
Presenter: **Bella Vakulenko-Lagun**, University of Haifa, Israel

Many health surveys collect data in a cross-sectional way and record only a current value of a Patient Reported Outcome. Any such study encounters a problem in that at the time of data collection, some of the important events, related to the studied outcome, had happened for some of the survey participants, but not for others. In addition, the incompleteness in this time-to-event covariate might be complicated by the presence of a cured fraction (those patients who do belong to the target population but will never experience this event). An example of such data is the data from the Web-based Adult Perthes Survey, which was launched in order to collect data on the physical functioning of patients who had a rare Perthes' disease in childhood. A total hip replacement (THR) in adulthood is a life-changing event for those Perthes patients who need it, but it might not be needed for some of the Perthes patients. We aim to estimate trajectories of the physical functioning of Perthes patients in adulthood as functions of their Perthes history, including THR and age-at-THR. There are few available approaches for cross-sectionally measured outcomes with censored covariates, and none of them accounts for a cured fraction. We propose an approach based on a pseudo-likelihood and assess its finite sample performance in simulations. We derive its asymptotic properties and apply our approach to the data on the long-term outcomes of Perthes' disease.

E1647: **Assessing patient benefit with semi-Markov multi-state models**
Presenter: **Abdul Haris Jameel**, University of Nottingham, United Kingdom
Co-authors: Christopher Fallaize, Joachim Grevel, Blesson Chacko, Christopher Brignell, Gilles Stupfler

In the context of oncological drug trials, the Cox proportional hazards model is traditionally used to establish treatment efficacy. However, a treatment which causes a high rate of premature discontinuation due to adverse side effects could be considered clinically effective by the Cox model, despite being too toxic for most patients to tolerate. A patient-focused modelling approach would instead seek to answer whether a particular treatment can treat cancer while being sufficiently tolerable for patients. Such information is crucial for patients, their doctors, and caretakers to assess and make better-informed decisions about optimally using the patients' limited remaining lifespans. Multi-state models are well known and can alternatively be used to model the entire event history of patients in drug trials. Furthermore, assuming an underlying semi-Markov process allows for arbitrary holding times before transitioning to other states. This allows for an alternative set of tools to quantify the various effects of a treatment on patients, and thus whether patients can potentially benefit. The main focus is on the methods used to formally define and quantify the patient benefit, namely the survival function of the holding time in a given state. A proposed hypothesis test is also discussed and validated.

E1934: **A joint model for multiple longitudinal responses with informative time measurement**
Presenter: **Ines Sousa**, Minho University, Portugal

In longitudinal studies, individuals are measured repeatedly over a period of time for a response variable of interest. In classical longitudinal models, the longitudinal observed process is considered independent of the times when measurements are taken. However, in a medical context, it is common that patients in the worst health conditions are more often observed, whereas patients under control do not need to be seen so many times. Therefore, longitudinal models for data with this characteristic should allow for an association between longitudinal and time measurement processes. We consider a response longitudinal variable with Gaussian distribution. We propose a model where the follow-up time process is stochastic. The model is described through the joint distribution of the observed process and the follow-up time process. Estimation of model parameters is through maximum likelihood. We conducted a simulation study of longitudinal data where model parameter estimates are compared, when using the model proposed and ignoring the association between processes. Finally, the model proposed is applied to a real data set when monitoring for biomarkers CEA and CA15.3 on breast cancer progression. In this case, the follow-up time process should be considered dependent on the longitudinal outcome process.

EC775 Room K0.19 NON- AND SEMI-PARAMETRIC STATISTICS Chair: Michelle Carey

E1509: **An interval-valued random forests model**
Presenter: **Paul Gaona Partida**, Utah State University, United States
Co-authors: Chih-Ching Yeh, Yan Sun, Adele Cutler

Analyzing soft interval data for uncertainty quantification has attracted much attention recently. Within this context, regression methods for interval data have been extensively studied. As most existing works focus on linear models, it is important to note that many problems in practice

are nonlinear in nature and the development of nonlinear regression tools for interval data is crucial. An interval-valued random forests model is proposed that defines the splitting criterion of variance reduction based on an L_2 type metric in the Banach space of compact intervals. The model simultaneously considers the centers and ranges of the interval data as well as their possible interactions. Unlike most linear models that require additional constraints to ensure mathematical coherences, the proposed random forests model estimates the regression function in a nonparametric way, and so the predicted length is naturally nonnegative without any constraints. Simulation studies show that the new method outperforms typical existing regression methods for various linear, semi-linear, and nonlinear data archetypes and under different error measures. A real data example is presented to demonstrate the applicability where the price range data of the Dow Jones Industrial Average index and its component stocks are analyzed.

E1532: **Targeted weight estimation in the heteroscedastic partially linear model**
Presenter: **Elliot Young**, University of Cambridge, United Kingdom
Co-authors: Rajen D Shah

Heteroscedasticity is common in many regression settings, a prime example being (partially) linear mixed effects models. In order to estimate regression coefficients in such models accurately, we must account for this via weighted estimation. Classical approaches use weights derived from parametric models for the conditional covariance function (CCF). We show, however, that in the inevitable case where these models are misspecified, both (restricted) maximum likelihood and regressing squared residuals onto covariates to estimate the CCF can lead to poor estimation of target parameters that may even be substantially worse than using no weighting at all. Instead, we propose to choose weights to directly minimise an estimate of the asymptotic variance of the parameter of interest. When used with a potentially misspecified model for the weights, we argue that in contrast to classical approaches, this always yields an asymptotically optimal estimate of the target parameter subject to the modelling constraints imposed. We introduce a computationally efficient boosting scheme to perform this optimisation that can leverage flexible machine learning methods to approximate the unknown CCF. We demonstrate the effectiveness of our approach on real and simulated data.

E1634: **Towards dynamic quantile function models for anomaly detection**
Presenter: **Lekha Patel**, Sandia National Laboratories, United States
Co-authors: Peter Jacobs

Quantile regression provides a powerful nonparametric framework for learning the relationship between predictors and specific quantiles of a target variable, while simultaneously resisting anomalous observations' effect. Such a robust method is thus able to consider a wide range of data structures at different quantiles. Due to the monotonicity constraint of the underlying quantile function, the standard quantile regression set-up induces major challenges when concurrently considering multiple quantiles of interest in the data. As a result, the quantile crossing problem, where estimated quantiles violate this constraint, is frequently encountered when several quantiles of the data are analyzed at once. We study a class of densities able to characterize time-varying monotonic quantile functions underlying observations collected through time. We detail an approach for relating observations with such densities in a Markovian fashion which leverages a recently developed family of distributions for quantile regression. Further, fast approximate Bayesian inferential schemes are discussed for parameter estimation of the quantile functions along the stream. Estimated quantiles with credible bounds are finally compared against labeled observations of anomalies in an exemplary case study.

E1710: **More efficient exact permutation testing: using a representative subgroup**
Presenter: **Nick Koning**, Erasmus University Rotterdam, Netherlands
Co-authors: Jesse Hemerik

Non-parametric tests based on permutation, rotation or sign-flipping are examples of so-called group-invariance tests. These tests test the invariance of the null distribution under a set of transformations that has a group structure, in the algebraic sense. Such groups are often huge, which makes it computationally infeasible to test using the entire group. Hence, it is standard practice to test using a randomly sampled set of transformations from the group. This random sample still needs to be substantial to obtain good power and replicability. We improve upon the standard practice by using a well-designed subgroup of transformations instead of a random sample. The resulting subgroup-invariance test is still exact, as invariance under a group implies invariance under its subgroups. We illustrate this in a generalized location model and find that it can yield a more powerful and fully replicable test with the same number of transformations. For the special case of a normal location model and a particular design of the subgroup, we show that the power improvement is equivalent to the power difference between a Monte Carlo Z-test and a Monte Carlo t-test. In our simulations, we find that our test has the same power as a test based on sampling that uses twice as many random transformations.

E1622: **Bayesian projection pursuit regression**
Presenter: **Gavin Collins**, Ohio State University, United States

In projection pursuit regression (PPR), an unknown response function is approximated by the sum of M "ridge functions", which are flexible functions of one-dimensional projections of a multivariate input space. Traditionally, optimization routines are used to estimate the projection directions and ridge functions via a sequential algorithm, and M is typically chosen via cross-validation. We introduce the first Bayesian version of PPR, which has the benefit of accurate uncertainty quantification. To learn the projection directions and ridge functions, we apply novel adaptations of methods used for the single ridge function case ($M = 1$), called the Single Index Model, for which Bayesian implementations do exist; then use reversible jump MCMC to learn the number of ridge functions M. We evaluate the predictive ability of our model in 20 simulation scenarios and for 23 real datasets, in a bake-off against an array of state-of-the-art regression methods. Its effective performance indicates that Bayesian Projection Pursuit Regression is a valuable addition to the existing regression toolbox.

EC778 Room S0.11 METHODOLOGICAL AND APPLIED STATISTICS Chair: Sabrina Giordano

E0373: **Assessment of the performances of new linear profile memory-type schemes under fixed explanatory variables**
Presenter: **Majika Jean-Claude Malela**, University of Pretoria, South Africa

Classical monitoring schemes are designed to monitor one or several quality characteristics that do not depend on other variables. When there is a functional relationship between the quality characteristic and one or several exogeneous variables, the use of classical monitoring schemes is not appropriate. In this case, the literature proposes the use of simple or general profiles depending on the number of exogeneous variables. A new multivariate exponentially weighted moving average (MEWMA) scheme for general linear profiles is proposed. The performance of the new MEWMA scheme is investigated using extensive simulations, and it is compared to that of the extended versions of the MEWMA, that is, the multivariate double, triple and quadruple EWMA (MDEWMA, MTEWMA and MQEWMA) schemes for general linear profiles in terms of the zero-and steady state run-length properties. It is found that the proposed MEWMA scheme is superior to the competing schemes in many cases under the zero-state and outperforms them in the steady-state regardless of the situation. A numerical real-life example is also provided to demonstrate the practicability and superiority of the MEWMA scheme over the considered competing schemes.

E1800: **Multiple-choice log-linear cognitive diagnostic model framework and its application**
Presenter: **Kentaro Fukushima**, The University of Tokyo, Japan
Co-authors: Kensuke Okada

Polytomous response models have attracted attention in the literature on diagnostic classification models (DCMs) for efficiently extracting diagnostic information on the levels of learners. In addition, the DCMs for multiple-choice items in which each option has its own Q-vector make effective use of observed information in distractors to infer the attribute mastery status of examinees. A novel framework of DCMs for multiple-choice items is proposed from a unified perspective. This framework is derived from the log-linear cognitive diagnostic model (LCDM) in binary DCMs and is

hereby referred to as multiple-choice LCDM (MC-LCDM). It expresses models by the main effects and interactions of attribute mastery states and elements of Q-vectors. Existing DCMs for multiple-choice can be interpreted as the combinations of these effects in the proposed framework by introducing appropriate parameter constraints. In addition, novel sub-models can be derived within the framework based on rational assumptions. Essentially, these reduced models are suitable for empirical analysis with a modest sample size since they can save the number of parameters used without sacrificing the empirical adequacy of the model. Moreover, real data illustration of several sub-models within the proposed framework is likewise provided to demonstrate the applicability and interpretability of the proposal.

E1648: **Effectiveness's comparison of longitudinal imputation methods for wave nonresponse applied to LFS data, ICBS**
Presenter: **Fatina Awad**, Central Bureau of Statistics, Israel
Co-authors: Louiza Burk, Tzahi Makovky

In a longitudinal panel (LP) survey, wave nonresponse occurs when responses are obtained for some but not all waves. The Labor Force Survey (LFS), ICBS, is an LP survey in which wave nonresponse manifests in two aspects: person wave nonresponse and household wave nonresponse, and both are handled by Nonresponse Weighting Adjustment in a Cross-Sectional approach. Longitudinal imputation methods use information obtained from previous waves in the missing wave data's imputation process in order to reduce estimation bias. Three longitudinal single imputation methods: Randomized Hot-Deck, K-Nearest Neighbors, and Classification and Regression Trees, were implemented in a multivariate approach to the LFS data through a simulations process. 500 random samples were drawn from the original non-missing data, and the original values of the target variables were removed. Thus, the values of the target variable were imputed according to each of the above methods to examine each method's performance. The effectiveness of these methods was tested by comparing the Precision metric, and measures for estimation quality, such as Root Mean Square Error, Mean Absolute Error and the fit between the distribution of the imputed values versus the original values.

E1726: **Statistical inference for the virtual age imperfect repair model**
Presenter: **Mosa Alsabhi**, Durham University, United Kingdom

The repairable systems are usually analysed using Nonhomogeneous Poisson Process (NHPP), which represents minimal repairs or as bad as old and the Renewal process (RP) which represents perfect repairs or as good as new. However, some repairs are imperfect, which will be between minimal repairs and perfect repairs. Virtual age models are imperfect repair models used for repairable systems after each repair or maintenance. After each repair, a system can either be rated as bad as old, as good as new, or somewhere between the two. Two methods for estimating virtual age models will be presented based on the Weibull distribution: maximum likelihood estimation (MLE) and the Bayesian estimator method. These two methods are compared using the Monte Carlo simulation by generating different random sample sizes. Then, an example will be provided that will illustrate MLE and Bayesian estimation inference.

C2041: **Combining extreme value theory with martingale regression in market risk analytics and portfolio management**
Presenter: **Wei Dai**, Shenzhen Institute of Information Technology, China
Co-authors: Tze Leung Lai

A new econometric approach is introduced to the estimation of VaR and convex risk measures in financial risk management, which uses a martingale regression for asset pricing and the extreme value theory (EVT) for the martingale difference residuals. Using decoupling inequalities, we show that the MLGEV (maximum likelihood for generalized extreme value distributions) and the peaks over threshold (POT) method in EVT can apply to the dynamically scaled residuals. This approach, therefore, addresses the pitfalls of EVT techniques while enabling them to realize their widely recognized potential for estimating extreme quantiles and probabilities.

CI019 Room BH (S) 1.01 Lecture Theatre 1 NEW ADVANCES IN INFERENCE Chair: Indeewara Perera

C0179: **Bootstrap inference in the presence of bias**
Presenter: **Giuseppe Cavaliere**, University of Bologna, Italy
Co-authors: Silvia Goncalves, Morten Nielsen

Bootstrap inference is considered for estimators which are (asymptotically) biased. We show that, even when the bias term cannot be consistently estimated, valid inference can be obtained by proper implementations of the bootstrap. Specifically, we show that the prepivoting approach, originally proposed to deliver higher-order refinements, restores bootstrap validity by transforming the original bootstrap p-value into an asymptotically uniform random variable. We propose two different implementations of prepivoting (plug-in and double bootstrap) and provide general high-level conditions that imply the validity of bootstrap inference. To illustrate the practical relevance and implementation of our results, we discuss five applications: (i) a simple location model for i.i.d. data, possibly with infinite variance; (ii) regression models with omitted controls; (iii) inference on a target parameter based on model averaging; (iv) ridge-type regularized estimators; and (v) dynamic panel data models.

C0180: **Asymptotics and inference for autoregressive conditional duration models**
Presenter: **Anders Rahbek**, University of Copenhagen, Denmark
Co-authors: Giuseppe Cavaliere, Thomas Mikosch, Frederik Vilandt

New asymptotic distributional results are presented for likelihood-based estimators in autoregressive conditional duration (ACD) models. We show the unexpected result that the large sample behavior of the estimators strongly depends on the tail behavior of the duration data, and hence on the finiteness of moments. In particular, we show that asymptotic normality breaks down when the tail index of the durations is smaller than one. In this case, the estimators are shown to be asymptotically mixed Gaussian with a non-standard rate of convergence depending on the tail index. Our results are particularly surprising when compared to the analysis of ARCH models: while ARCH and ACD likelihood functions have the same form, standard asymptotic arguments apply to ARCH models but not to ACD models. The crucial difference between the two types of models is that for ACD models, the number of observations within any given sample period is random. This feature, rather than being innocuous, requires as demonstrated new, non-standard theory.

C0259: **Extending the scope of inference about predictive ability to machine learning methods**
Presenter: **Juan Carlos Escanciano**, Universidad Carlos III de Madrid, Spain

Though out-of-sample forecast evaluation is routinely recommended with modern machine learning methods, and there exists a well-established classic inference theory for predictive ability, such theory is not directly applicable to modern machine learners such as the Lasso in the high dimensional setting. We investigate under which conditions such extensions are possible. Two key properties for standard asymptotic inference are: (i) a zero mean condition for the score of the loss function (a locally robust property); and (ii) a fast rate of convergence for the machine learner.

CO128 Room BH (SE) 1.05 DYNAMIC CONDITIONAL SCORE MODELS Chair: Dario Palumbo

C0511: **A Lucas critique compliant SVAR model with observation-driven time-varying parameters**
Presenter: **Giacomo Bormetti**, University of Bologna, Italy
Co-authors: Fulvio Corsi

An observation-driven time-varying SVAR model is proposed where, in agreement with the Lucas critique, structural shocks drive both the evolution of the macrovariables and the dynamics of the VAR parameters. Contrary to existing approaches where parameters follow a stochastic process with random and exogenous shocks, our observation-driven specification allows the evolution of the parameters to be driven by realized past structural

shocks, thus opening the possibility to gauge the impact of observed shocks and hypothetical policy interventions on the future evolution of the economic system.

C0918: **Dynamic partial correlation models**
Presenter: **Enzo DInnocenzo**, VU University Amsterdam, Netherlands
Co-authors: Andre Lucas

A new nonlinear dynamic model is introduced for dynamic conditional correlation matrices. To generate correlation matrices that satisfy the constraints of positive (semi) definiteness and ones on the diagonal, we parameterize the correlation matrix using a sequence of partial correlations. Each partial correlation is built recursively from previous partial correlations and pairwise correlations using the so-called D-vine copula structure in a static framework for random correlation matrices. The main advantages of this strategy are that (i) it ensures positive definite correlation matrices with a sequence of simple transformations; (ii) the method is easily scalable to higher dimensions without losing computational stability (which off-sets it from other parameterizations); (iii) the recursive structure of the parametrization allows for much a much simplified asymptotic analysis of the process and the maximum likelihood estimator; (iv) the formulation allows us to easily impose (theoretical) restrictions such as zero restrictions on some of the partial correlations during the filtering stage. We provide conditions for stationarity, ergodicity and invertibility of our model and prove strong consistency and asymptotic normality of the maximum likelihood estimator. An extensive Monte Carlo simulation and an empirical in-sample and out-of-sample analysis of stock return data show that the new approach outperforms a range of recent alternatives.

C1308: **Dynamic combination and calibration of forecasts**
Presenter: **Dario Palumbo**, University Ca' Foscari of Venice, Italy
Co-authors: Roberto Casarin, Francesco Ravazzolo

A density calibration and combination model is proposed that dynamically calibrates and combines predictive distributions. The time-varying calibration and combination weights are fitted by an observation-driven model with dynamics inferred by the score of the assumed conditional likelihood of the data-generating process. Through simulations, we show that the model is very flexible and can handle different shapes, instability and model uncertainty. An empirical application to short-term wind speed predictions documents the large instability of individual model performance and their calibration properties, favouring our model in terms of predictive accuracy.

C1353: **Measuring the information content of trades in a time-varying setting**
Presenter: **Francesco Campigli**, Scuola Normale Superiore, Italy
Co-authors: Fabrizio Lillo, Giacomo Bormetti

The estimation of the market impact of trades on prices is important for measuring the information content of trades, for optimal execution, and for transaction cost analysis. Originally, a Structural-VAR (S-VAR) model was proposed to be used, but more recent literature has highlighted some pitfalls of this approach. S-VAR models are misspecified, and the estimates can be contradictory when the permanent impact function has a nonlinear relationship with the trade sign. They are not flexible to parsimoniously exploit the long memory of the order flow variable. The instantaneous impact, which is a measure of market liquidity, is constant, while liquidity is known to be highly fluctuating. A nonlinear modified score-driven version of the original model is proposed. To measure the long-term effect of trade on prices, we compute the asymptotic cumulative impulse response function using Monte Carlo simulations. The analysis indicates that the trade information content varies and it is conditional on past trades and prices. Real-time knowledge of the permanent impact is important for transaction cost analysis. We derive an expression for the permanent impact from the estimated parameters. Simulations and empirical applications suggest that the approach provides reliable estimates of the cost of an optimal execution.

C1375: **On the optimality of score-driven models**
Presenter: **Alessandra Luati**, University of Bologna, Italy
Co-authors: Paolo Gorgi, Christopher Sacha Aristide Lauria

Score-driven models are shown to be optimal with respect to a novel, intuitive, high dimensional and global optimality criterion, defined as Conditional Expected Variation optimality. The property formalises the use of the score as the driving force in updating models for time-varying parameters. To prove the aforementioned property, a point of contact between the econometric literature and the time-varying optimisation literature is established. As a matter of fact, Conditional Expected Variation optimality can be naturally viewed as a generalisation of the monotonicity property of the gradient descent scheme. Differently from information-theoretic optimality criteria based on the Kullback-Leibler divergence between the model density and the true density, the Conditional Expected Variation regards Euclidean distances in the parameter space, holds on the whole parameter space and is trivially extended to the case in which the time-varying parameter is multidimensional.

CO593 Room BH (SE) 2.05 STRUCTURAL, PREDICTIVE INFERENCE IN NONLINEAR MACROECONOMETRICS Chair: Niko Hauzenberger

C0273: **Model specification for Bayesian neural networks in macroeconomics**
Presenter: **Karin Klieber**, Oesterreichische Nationalbank, Austria
Co-authors: Niko Hauzenberger, Florian Huber, M. Marcellino

Relations in macroeconomic data are often nonlinear and subject to structural breaks. This is commonly captured through appropriate models. However, by choosing a specific model, the researcher takes a strong stance on the nature and degree of nonlinearities. This gives rise to substantial model and specification uncertainty. We develop Bayesian neural networks that remain agnostic on the precise form of nonlinearities. Our model flexibly adjusts to the complexity of the dataset. This is achieved through Bayesian regularization techniques that adequately select the appropriate network structure without the necessity for using cross-validation. To investigate the degree and nature of nonlinearities in macroeconomic data, we train our neural network to four commonly used datasets in macroeconomics and finance. Our empirical results suggest that for cross-sectional data, a linear approximation works well in predictive terms, whereas for time series data, nonlinearities are important and especially so during turbulent times.

C0355: **Nonparametric methods for measuring asymmetries in monetary policy transmission**
Presenter: **Anna Stelzer**, University of Salzburg, Austria
Co-authors: Michael Pfarrhofer, Florian Huber, M. Marcellino

Nonparametric methods are considered to assess asymmetries in the transmission of monetary policy shocks. The preceding literature typically relies on variants of parametric (nonlinear) time series methods. While this allows for studying differentials in impulse response functions over time - such as varying persistence of the shocks - transmission channels are still linear conditional on each point in time. This rules out asymmetric responses with respect to the sign and size of the shock. Our proposed frameworks circumvent this limitation by relying on flexible nonparametric methods for the conditional mean of the underlying VAR model, which allows us to study asymmetric monetary policy responses along three dimensions: timing, sign and size of the shock.

C0729: **The nonlinear effects of shocks to bank capital vulnerability in Euro area countries**
Presenter: **Sharada Davidson**, University of Strathclyde, United Kingdom
Co-authors: Diego Moccero

When capital in the banking system is depleted, financial intermediation is impaired, and the broader economy suffers. However, the degree to which financial intermediation is weakened likely depends on the financial and macroeconomic environment. With a limited number of observations,

existing macroeconometric studies assume that the impact of bank capital shocks is linear or nonlinearities are examined using single-equation panel data models which ignore feedback effects to the macroeconomy. The focus is on the nonlinear propagation of bank capital shocks by estimating three Bayesian Panel Threshold VAR models with macroeconomic and aggregate banking variables for Germany, France, Italy and Spain. Evidence shows that when banks become vulnerable to a depletion in capital, bank lending supply and economic activity are adversely affected. Crucially, these effects are stronger when banks are already highly vulnerable to losing capital, and the policy interest rate is low. The state of the business cycle, however, does not appear to amplify the impact of bank capital shocks. We conclude that the financial environment (bank vulnerability and the monetary policy stance) is more important than the macroeconomic environment (business cycles) in amplifying adverse bank capital shocks.

C1029: **The impact of carry trade activity on the transmission of monetary policy**
Presenter: **Thomas Zoerner**, Oesterreichische Nationalbank, Austria
Co-authors: Maximilian Boeck, Alina Steshkova

Currency markets are shaped by a variety of puzzles and influenced, inter alia, by the current monetary policy stance. This gives rise to carry trade strategies that seek to exploit existing global interest rate differentials. We study how carry trade activity affects the transmission of monetary policy on currency markets. After monetary and non-monetary news from the central bank, investors may quickly unwind their positions in currencies featuring high-carry trade activity. To infer these effects from data, a threshold vector autoregressive model is fitted to discriminate between different regimes of carrying trade activity and estimate the effects of monetary policy shocks for a large set of currencies vis-a-vis the US dollar. Empirical evidence shows that carry trade activity indeed affects the transmission of monetary policy to exchange rate markets, causing different adjustment dynamics. Thus, we conjecture that the exchange rate channel and its behavioral component is an important dimension of monetary policy and need to be taken properly into account.

C1933: **What can be learned about the future**
Presenter: **Philippe Goulet Coulombe**, Universita du Quabec a Montraal, Canada
Co-authors: Maximilian Goebel

Most of the economic forecasting literature focuses on increasing the predictability of key indicators by tirelessly developing new models and algorithms – often met with modest success, if any. The inverse of the prototypical forecasting problem is investigated. Given an information set and a particular model, we find the transformation and combination of many variables' future realizations to maximise the composite's predictability. This is implemented through a multivariate generalization of the ACE algorithm (Alternating Conditional Expectations) that we inevitably call the MACE. The approach mechanizes many manual interventions that have populated the time series econometrics practice in recent and less recent years. Among others, notable special cases include finding predictability in tails of distributions, modeling volatilities and covariances, and core inflation. MACE also allows for the algorithmic discovery of new predictable features of the macroeconomy.

CO693 Room BH (SE) 2.10 LATEST DEVELOPMENTS IN FINANCIAL ECONOMETRICS Chair: Roxana Halbleib

C0236: **A multivariate perturbation robust test against spurious long memory**
Presenter: **Vivien Less**, Leibniz Universitaet Hannover, Germany
Co-authors: Philipp Sibbertsen

A multivariate extension to the local Whittle with noise estimator, as well as a modified score-type test against spurious long memory, are introduced. The test statistic is based on the weighted sum of the partial derivatives of the multivariate local Whittle with noise estimator. Explicitly addressing the noise term when approximating the spectral density near the origin improves the efficiency of the estimator and the size and power properties of the test. We prove the consistency and asymptotic normality of the local Whittle estimator, and we derive the limiting distribution and show the consistency of the procedure. An empirical example on the squared returns and the realised volatilities from the Spanish IBEX, Nikkei 250, Swiss Market Index, and KOSPI Composite Index is conducted and shows the usefulness of the procedures.

C0477: **Common factors in large panels of option prices**
Presenter: **Maria Grith**, Erasmus University Rotterdam, Netherlands
Co-authors: Paolo Santucci de Magistris, Francesco Violante, Pierluigi Vallarino

A new factor model is proposed for multivariate tensor-valued data that describes the joint dynamics of a large cross-section of option prices containing over 200 equities of the SPX 500 Index with different strike prices and expiration dates. The factors explain the common variation of all the options in the cross-section, and we model their dynamics in a standard time series context. In contrast, the factor loadings express the heterogeneous response to common shocks and are two-dimensional arrays (i.e., tensors). We propose an inference framework to test the significance of the loadings. Furthermore, we implement a tensor-counterpart version of the multivariate principal component model to deal with the high-parametrization of the factor loadings, which enables us to extract trading signals, which we use to design a dynamic trading strategy. Our results show that this strategy yields significantly higher profits than a mean-variance investment strategy, even when controlling transaction costs.

C0501: **Finite sample performance of generalized covariance estimator**
Presenter: **Joann Jasiak**, York University, Canada

The finite sample performance of the Generalized Covariance (GCov) estimator is examined for semi-parametric dynamic models with independent identically distributed errors. The GCov estimator is obtained by minimizing a residual-based multivariate portmanteau statistic. It has an interpretable objective function, circumvents the inversion of high-dimensional matrices, and achieves semi-parametric efficiency in one step. We study its finite sample properties in application to the mixed causal-noncausal Vector Autoregressive (VAR) model and examine the effect of the error distribution, the number of autocovariance conditions and the lag.

C0718: **Multiple forecast comparisons in unstable environments and high dimensions**
Presenter: **Ekaterina Smetanina**, University of Chicago, United States

A methodology is developed for the forecast evaluation of various models in potentially unstable environments. We first propose a new measure of forecast performance in unstable environments and then develop a methodology for selecting the best forecasting model from a pool of models while allowing for general types in potential instabilities (e.g. best forecasting model that changes over time). At a given point in time, the methodology allows practitioners to construct a set of best models - models that are indistinguishable from each other given our new metric.

C0798: **Bagged value-at-risk forecast combination**
Presenter: **Ekaterina Kazak**, University of Manchetser, United Kingdom
Co-authors: Roxana Halbleib, Winfried Pohlmeier

Recent developments in financial econometrics literature on joint scoring functions for Value-at-Risk and Expected Shortfall allowed for consistent implementation of statistical tests based on the Model Confidence Set (MCS). MCS is shown to be a great tool for model comparison, both in-sample and out-of-sample. Another branch of literature focused on the superior performance of convex forecast combinations, which often outperform stand-alone forecasting models. Both results are combined, and a novel approach is proposed to a forecast combination of Value-at-Risk and Expected Shortfall based on the MCS. We exploit the statistical properties of bootstrap aggregation (bagging) and combine competing models based on the bootstrapped probability of the model being in the Confidence Set. The resulting forecast combination allows for a flexible and smooth switch between the underlying models and outperforms the corresponding stand-alone forecasts.

CO398 Room BH (SE) 2.12 RECENT ADVANCES IN QUANTITATIVE FINANCE Chair: Shixuan Wang

C1417: **Implied willow tree method for the term structure of moments risk premia**
Presenter: **Zhenyu Cui**, Stevens Institute of Technology, United States
Co-authors: Bing Dong, Wei Xu

A data-driven implied willow tree method is proposed to extract the joint implied risk-neutral density functions of asset prices at several future time points from market-observable options prices with various strikes and maturities. The implied risk-neutral density functions are then transformed into the density functions under the physical measure, and the accuracy and robustness of our method are examined in recovering the pricing kernel on both synthetic and empirical data sets. Finally, we study the forward-term structure of moment risk premia based on the S&P 500 options data from 2006 to 2019 to explore the impact of market jumps and investors' behaviors.

C1119: **Intraday foreign exchange rate volatility forecasting: Univariate and multilevel functional GARCH models**
Presenter: **Yuqian Zhao**, University of Kent, United Kingdom
Co-authors: Fearghal Kearney, Han Lin Shang

The aim is to predict conditional intraday volatility in foreign exchange (FX) markets using functional Generalized AutoRegressive Conditional Heteroscedasticity (GARCH) models. We contribute to the existing functional GARCH-type models by accounting for the stylised features of long-range conditional heteroscedasticity and cross-dependence in major FX currencies through estimating the models with long-range dependent and multi-level functional principal component basis functions, as well as incorporating the intraday bid-ask spread microstructure information. Overall, we demonstrate the statistical and economic superiority of appropriately modelling FX volatility using various functional GARCH-based models. Remarkably, we find that taking account of cross-dependency dynamics between the major currencies can significantly improve intraday conditional volatility forecasting in the FX market. Intraday risk management benefits and inter-daily asset allocation applications are presented to highlight the practical benefits of our proposed approaches.

C0525: **Distinctly large eigenvalue with approximately diagonal eigenvector of covariance matrix**
Presenter: **Yifan Zhang**, Renmin University of China, China
Co-authors: Zhenya Liu, Bo Li

Covariance matrices of financial returns are often observed with distinctly large eigenvalues and approximately diagonal eigenvectors. With a geometric illustration, we show that how such characteristic occurs for specific covariance matrices.

C1185: **Does the less sophistication investors intended to prefer simple signals to choose mutual fund**
Presenter: **Jinge Fu**, University of Glasgow, United Kingdom

It is shown that investors tend to use some simple signal (Morningstar rating) as an indicator when they doing a mutual fund invest. Two tests have been used to prove our results. Using proxies for investor sophistication (distribution channels, and periods of investor sentiment), we prove that sophisticated investors will consider complex factor models when assessing a fund managers skill. We also investigate which factor investors focused on by using flow and factor-alpha.

C0528: **Asset pricing model with functional principal component analysis**
Presenter: **Bo Li**, Beijing International Studies University, China
Co-authors: Zhenya Liu, Shixuan Wang, Yifan Zhang

A functional principal component analysis (fPCA) procedure is proposed to construct characteristic portfolios for estimating risk factors in asset pricing models. It sorts individual returns on univariate characteristics to attain a balanced panel and utilizes fPCA to extract statistical factors. We empirically verify that the first is equivalent to the market factor, and the second has properties suitable to be a pricing factor, named the fPCA factor. Then, we suggest constructing characteristic portfolios using the second eigenfunction as weights. Further empirical study confirms that the fPCA factors substantially improve the pricing performance of the multi-factor model for anomalies. Their mean-variance efficiency portfolios achieve higher out-of-sample Sharpe ratios than the conventional factors. Specifically, the fPCA size and momentum factors with the market factor attain a Sharpe ratio of 2.12.

CC761 Room BH (SE) 1.02 FORECASTING Chair: Alessandra Amendola

C1651: **The role of central bank forecasts in the turbulent times**
Presenter: **Jacek Kotlowski**, Narodowy Bank Polski, Poland

The aim is to examine whether the public information provided by the central bank affects the forecasts formulated by professional forecasters. We investigate empirically whether the projection of GDP and CPI inflation published by Narodowy Bank Polski (the central bank of Poland) reduces the disagreement in professional forecasters' expectations and acts as an attractor for medium-term forecasts. We find that central bank forecasts are the focal point for professional forecasters. We also document that the role of the central bank forecast for the coordination of professional forecasters' expectations strengthens in turbulent times when the uncertainty is high and when inflation deviates from the central bank target.

C1874: **A procedure for upgrading linear-convex combination forecasts with an application to volatility prediction**
Presenter: **Verena Monschang**, University of Muenster, Germany
Co-authors: Wilfling Bernd

Mean-squared-forecast-error (MSE) accuracy improvements are investigated for linear-convex combination forecasts, whose components are pre-treated by a procedure called Vector Autoregressive Forecast Error Modeling (VAFEM). Assuming that the forecast-error series of the individual forecasts are governed by a stable VAR process under classic conditions, we obtain the following results: (i) VAFEM treatment bias-corrects all individual and linear-convex combination forecasts. (ii) Any VAFEM-treated combination has a smaller theoretical MSE than its untreated analogue, if the VAR parameters are known. (iii) In empirical applications, VAFEM gains depend on (1) in-sample sizes, (2) out-of-sample forecast horizons, and (3) the biasedness of the untreated forecast combination. We demonstrate the VAFEM capacity for realized-volatility forecasting, using S&P 500 data.

C2008: **Censored ExponenTial smoothing to solve lost-sales demand forecasting problems**
Presenter: **Diego Pedregal**, University of Castilla-La Mancha, Spain
Co-authors: Juan R Trapero, Enrique Holgado

Sales data are used as a good approximation to the true demand in many supply inventory management contexts, even though it is well-known that such an approximation is not as good as it may seem when lost sales occur. In these cases, crucial elements for inventory management, such as safety stocks or reorder points, are miscalculated. A censored general ExponenTial Smoothing algorithm is developed to estimate correctly demand when only sales are available in contexts where lost sales are known to occur. The algorithm is based on a general dynamic linear innovations state space system with censoring levels at the output. Once the model is set up, the estimation of parameters and initial conditions are estimated as usual by Maximum Likelihood. Such generality provides a solution with several advantages. First, the solution is general enough to be able to implement other methodologies straightforwardly, like time-varying regression or ARIMA models with censoring levels. Second, time-varying censoring levels may be considered, actually, the usual situation in real life, where censoring levels depend on forecasts themselves and, consequently, change over time. Finally, models with all sorts of components (like the trend, seasonal, coloured noise, and exogenous variables) can be used.

C1897: **Forecasting the conditional Covariance: Using daily, intraday returns, or both**
Presenter: **Yongdeng Xu**, Cardiff University, United Kingdom

The intraday returns are used to construct the "realized" covariance, which provides a more precise measurement of covariance than the daily returns. When working with high-frequency data from markets that operate during a reduced time, an approach is needed to correct the missing overnight information. Typically, overnight covariance is added to the realized covariance, or the realized covariance is rescaled to the daily covariance. A recently developed high frequency-based multivariate GARCH (HF-MGARCH) model makes use of both daily and intraday returns, and provides more accurate forecasts of daily covariance. We compare the forecasting performance daily realized covariance models with the HF-MGARCH models. Statistically, we find that HF-MGARCH models perform better than the realized covariance models augmented with overnight returns. The daily rescaled realized covariance performs almost as well as the HF-MGARCH model. Economically, we find that a risk-averse investor would be willing to pay 20 to 80 basis points per year to capture the observed gains in portfolio performance by switching from intraday realized covariance to HF-MGARCH modelling of the daily conditional covariance.

C1564: **Minimizing post-shock forecasting error through aggregation of outside information**
Presenter: **Daniel Eck**, University of Illinois, United States

A forecasting methodology is developed to provide credible forecasts for time series that have recently undergone a shock. We achieve this by borrowing knowledge from other time series that have undergone similar shocks for which post-shock outcomes are observed. Three shock effect estimators are motivated with the aim of minimizing average forecast risk. We propose risk-reduction propositions that provide conditions that establish when our methodology works. Bootstrap and leave-one-out cross-validation procedures are provided to prospectively assess the performance of our methodology. Several simulated data examples and two real data examples of forecasting Conoco Phillips and Apple stock price are provided for verification and illustration.

CC781 Room BH (SE) 1.06 FINANCIAL MODELLING Chair: Douglas Hodgson

C1304: **Dynamic comovement and spillover among global financial markets during Covid19: Evidence from wavelet coherence analysis**
Presenter: **Luiz Felix**, VU University, Netherlands
Co-authors: Limiao Bai, Antonio Eduardo Clark Peres, Virgilio Gaggiato

The purpose is to investigate how the Covid-19 pandemic impacted co-movements and lead-lag relationships among global financial markets. Large unexpected events, such as the September 11th terrorist attack, generally transmit transient shocks to financial markets capable of changing the direction of their causal relationships. The outbreak of Covid-19 pandemic was one such exogenous shock, though with a long-lasting effect, as it propagated waves of turbulence across financial markets and generated time-varying risk contagion among different assets and geographies. Wavelet coherence analysis provides us with an excellent tool to identify these dynamics, i.e., spillovers and risk contagion. Our empirical analysis, utilizing daily data from January/2018 to November/2021, reveals that co-movements between U.S. stocks and other major markets have increased substantially around March/2020 relative to the pre-Covid time. U.S. lead-lag relationship with other equity markets has also changed dramatically post-Covid, especially relative to the Chinese market. Changes in the lead-leg dynamics were also observed across equity, USD, oil, gold and Bitcoin markets. For instance, our results indicate that gold was still a better asset to hedge equity risk than Bitcoin, contrary to the hypothesized by recent studies. Our findings contribute to the domain of portfolio optimization and risk management at different horizon bands.

C1453: **The Abelian sandpile model in financial crises**
Presenter: **Wayne Tarrant**, Rose-Hulman Institute of Technology, United States

The situation of bank failure is modelled using the Abelian sandpile model from mathematical physics. The model leads to the situation of self-organized criticality, and we find consistent results over a wide range of probability distributions. This gives us a chance to estimate bank failures' natural rate without regulators' intervention. Because of the consistent results, we are able to determine the effectiveness of Central Bank regulation and interventions from countries which publish records of bank failures.

C1504: **A simplified decomposition model for asset return predictability**
Presenter: **Arsene Brou**, Laval University, Canada
Co-authors: Richard Luger

A previous modelling approach decomposes asset returns into their signs and absolute values, and obtains the joint distribution by specifying a multiplicative error model for the absolute values, a dynamic binary choice model for the signs, and a bivariate copula for their interaction. With this copula-based approach, each component is specified conditional on past information. We propose a simplified approach that recovers the dynamics of the joint distribution with a specification for the sign component that conditions on past information and the contemporaneous absolute value. In sharp contrast to the original decomposition model, the simplified approach avoids the need to specify a copula for joining a continuous margin with a discrete one. Simulation results demonstrate that the simplified approach does as well as copula-based specifications, and empirical findings show that a larger degree of stock return predictability is revealed by decomposition modelling than by traditional predictive regressions.

C1915: **An analysis of the return-volume relationship in decentralised finance (DeFi)**
Presenter: **Stephen Chan**, American University of Sharjah, United Arab Emirates

The decentralised finance sector has recently experienced a surge in popularity and emerged from the cryptocurrency space's shadows. Although the purposes of the currencies used in this new sector differ from traditional cryptocurrencies, they still possess monetary value and can be traded using at currencies on specialised decentralised exchanges. The aim is to investigate the dynamic volume-return relationship of the five largest decentralized finance tokens, to understand better this relationship given the similarities with cryptocurrencies and the possible benefits for traders and practitioners. We implement the quantile-on-quantile regression and an extreme value theory approach to examine the relationship between the daily returns of the prices and trading volumes of decentralised finance tokens at varying quantiles and at the extreme tails. Our results suggest that when trading volume is experiencing large increases, the returns of the prices of tokens appear to be significantly positive for some cases but negative for others. The extreme volume-return dependence is found to be asymmetric in the extreme negative and positive tails of the distributions, where the dependence below extreme negative thresholds is essentially non-existent, but above extreme positive thresholds, it is significant. This extreme dependence between return and volume may be beneficial for developing trading strategies that incorporate trading volume data.

C2020: **Stock market volatility and expected returns**
Presenter: **Yawen Zheng**, Durham University, United Kingdom
Co-authors: Sam Pybis

An empirical evaluation of US industry portfolio predictability is provided using excess stock return volatility as a predictor. Out-of-sample, excess stock return volatility is a poor predictor of the equity risk premium, but a strong predictor of industry portfolios. An industry rotation portfolio highlights the usefulness of this strategy to investors who time the market.

CC798 Room BH (SE) 2.09 INFLATION Chair: Leopold Soegner

C1573: **Back to the past: Long memory properties (again) matter for inflation forecasting**
Presenter: **Todd Prono**, Federal Reseve Board, United States

Emerging from the high inflation period of the late-1970s and mid-1980s, the headline CPI YOY rate was characterized as a long memory process

with hyperbolic decay. Over the next several decades, that long memory process degraded towards one of short memory with geometric decay. So stark was this degradation that over this period, the best out-of-sample forecasts, including multi-step-ahead forecasts, were produced by a model that ignores long memory properties altogether. By analyzing a series of multi-step, out-of-sample forecasts coming out of the COVD lock downs, it is demonstrated that long memory is back as a property affecting inflation dynamics. For instance, an ARFIMA model estimated over the high inflation period of the late-1970s and mid-1980s is shown to produce better medium-horizon CPI YOY rate forecasts than either an ARFIMA model estimated over a longer period that includes the Great Moderation or an ARIMA model estimated over the same period. An implication of this result is that the decay rate for headline inflation is time varying and slows down during periods of high inflation. Consequently, while the Great Moderation ushered in less 'sticky' inflation dynamics, that lost 'stickiness' has now returned.

C1652: **Consumer price index differentials among Polish households**

Presenter: **Aleksandra Halka**, Narodowy Bank Polski, Poland

Co-authors: Damian Stelmasiak

The aim is to investigate how the inflation perceived by the households differs with respect to the level of income. In order to estimate the scale of inflation differentiation, we divide the sample of households in Poland according to income into four quartile groups and calculate CPI for the respective groups. We focus on the difference in inflation between the group with the lowest and the highest income, which we call inflation inequality. Our results indicate that in Poland, the lowest-income households generally experienced higher inflation than the highest-income households. This is mainly due to the higher growth of food prices with the persistently significant disproportion in the share of food in the consumption basket of income groups. At the same time, inflation inequality is limited mainly by the prices of services. Inflation inequality increased after the outbreak of the pandemic in 2020, partly also due to a relatively greater increase in income in the fourth quartile group compared to the. However, in 2021 the inflation inequality was relatively low, which we attribute to the similarity of baskets of various income groups as a result of the restrictions related to COVID-19. However, the decline in inequality was short-lived, and in 2022 we observe a sharp rise in inflation inequality. We also find that the social policy (the Family 500+ program) aimed at reducing income inequalities lead to a significant reduction in the scale of income inequality.

C1810: **The risk of inflation dispersion in the Euro area**

Presenter: **Stephane Lhuissier**, Banque de France, France

Co-authors: Fabien Tripier, Aymeric Ortmans

The cross-country dispersion of inflation risks in the euro area and their macroeconomic origins are explored. The approach builds on the concept of inflation-at-risk developed, which is itself highly related to that of Growth-at-Risk. The inflation risk approach aims at forecasting shifts in the tails of inflation distribution. An in-depth analysis of inflation-at-risk in the euro area and the U.S. grounded has been recently provided on a quantile Phillips curve. We are not interested in the inflation risk for one country per see, but in the dispersion of these risks between euro area countries. We document five facts. (i) While the dispersion of inflation rates mainly concerns upside inflation risks during the first decade of the euro area, it shifted to downside inflation risks during the second decade. (ii) The dispersion of downside and upside risks to inflation reaches record levels in the wake of the COVID crisis. (iii) The main determinant of the dispersion at the bottom of the distribution is the development of financial stress. (iv) In the wake of the COVID crisis, value chain pressures drove the dispersion of inflation risks. (v) Overall, the dispersion of inflation rates is largely caused by heterogeneous Phillips curves between countries rather than by different national economic contexts.

C1945: **Testing hyperinflation data against Makochekanwa's model**

Presenter: **Laura Vaughan**, Vanderbilt University, United States

Hyperinflation has become an increasingly common problem in the last century. In 2007, Zimbabwe entered a period of extreme hyperinflation that led to the collapse of the Zimbabwean dollar. We evaluate Makochekanwa's model, which formed hypotheses surrounding Zimbabwean inflation from 1999-2006. Finding data on semi-recent situations in Zimbabwe to be sparse, we work to find appropriate estimators for missing pieces of data. We apply linear regression and the Toda-Yamamoto variant of Granger causality to check Makochekanwa's findings.

Saturday 17.12.2022 16:10 - 17:50 Parallel Session E – CFE-CMStatistics

EI007 Room Safra Lecture Theatre NEW DEVELOPMENTS FOR TIME SERIES ANALYSIS FOR COMPLEX DATA Chair: Christian Francq

E0182: **Model diagnostics for discretely sampled functional data**
Presenter: **Siegfried Hoermann**, Graz University of Technology, Austria
Co-authors: Fatima Jammoul

In practice, functional data are recorded on a discrete set of observation points. A common assumption in the literature is that these discrete measurements are blurred with white noise. Hence, in order to estimate the latent curves, some preprocessing is needed. Common techniques are kernel smoothing, non-linear regression, spline fitting, etc. Although such methods are massively applied in real data implementations, we seldom see those empirical analyses accompanied by diagnostic checks, evaluating the quality or suitability of the chosen preprocessing method. We consider functional data which are recorded on a regular time grid. Hence, the residuals related to each such functional datum may be viewed as a time series. If the preprocessing is accurately executed, these residuals should roughly behave like a white noise sequence. We illustrate, on the basis of some toy data, that for standard preprocessing techniques, this is often not the case. Rather we observe spurious periodicity in the residuals' autocovariance function. We will discuss this phenomenon and establish a suitable inferential framework for testing the white noise hypothesis.

E1186: **Graphical models for nonstationary time series**
Presenter: **Suhasini Subbarao**, Texas A&M, United States
Co-authors: Sumanta Basu

NonStGM, a general nonparametric graphical modeling framework, is proposed for studying dynamic associations among the components of a nonstationary multivariate time series. It builds on the framework of Gaussian Graphical Models (GGM) and stationary time series Graphical models (StGM), and complements existing works on parametric graphical models based on change point vector autoregressions (VAR). Analogous to StGM, the proposed framework captures conditional noncorrelations (both intertemporal and contemporaneous) in the form of an undirected graph. In addition, to describe the more nuanced nonstationary relationships among the components of the time series, we introduce the new notion of conditional nonstationarity/stationarity and incorporate it within the graph architecture. This allows one to distinguish between direct and indirect nonstationary relationships among system components and can be used to search for small subnetworks that serve as the "source" of nonstationarity in a large system. Together, the two concepts of conditional noncorrelation and nonstationarity/stationarity provide a parsimonious description of the dependence structure of the time series.

E1412: **Drift vs Shift: Decoupling trends and changepoint analysis**
Presenter: **David Matteson**, Cornell University, United States
Co-authors: Haoxuan Wu, Sean Ryan

A new approach is introduced for decoupling trends (drift) and changepoints (shifts) in time series. Our locally adaptive model-based approach for robust decoupling combines Bayesian trend filtering and machine learning-based regularization. An over-parameterized Bayesian dynamic linear model (DLM) is first applied to characterize drift. Then a weighted penalized likelihood estimator is paired with the estimated DLM posterior distribution to identify shifts. We show how Bayesian DLMs specified with so-called shrinkage priors can provide smooth estimates of underlying trends in the presence of complex noise components. However, their inability to shrink exactly to zero inhibits direct changepoint detection. In contrast, penalized likelihood methods are highly effective in locating changepoints. However, they require data with simple patterns in both signal and noise. The proposed decoupling approach combines the strengths of both, i.e. the flexibility of Bayesian DLMs with the hard thresholding property of penalized likelihood estimators, to provide changepoint analysis in complex, modern settings. The proposed framework is outlier robust and can identify a variety of changes, including in mean and slope. We illustrate the flexibility and contrast the performance and robustness of our approach with several alternative methods across a wide range of simulations and application examples.

EO612 Room S-2.23 PLATFORM TRIALS FOR COVID-19 INTERVENTIONS Chair: Jonathan Schildcrout

E1333: **Picking outcomes for outpatient platform trials in a pandemic**
Presenter: **Christopher Lindsell**, Vanderbilt University Medical Center, United States

One of the most important decisions in a clinical trial is which outcome to use. The preferred outcome for the same trial may depend on the viewpoint of the stakeholder consuming the results, and the intended policy or regulatory impacts. During the Covid-19 pandemic, we were asked to design a platform trial to test whether existing therapies with a known safety profile could be repurposed to help mild to moderately ill outpatients with Covid-19 get better faster, and stay out of the hospital. In this context, there are many different viewpoints as to the most important question, and the relevant question might change over time. The balance between generating evidence is reviewerd quickly using sensitive outcomes and generating sufficient evidence to support regulatory claims, and between a decentralized pragmatic trial deployed nationwide and a controlled clinical trial intended for supporting registrational claims.

E1386: **Designing an adaptive platform for an evolving pandemic: The PRINCIPLE trial**
Presenter: **Joe Marion**, Berry Consultants, United States

The PRINCIPLE trial is a randomized, adaptive, perpetual platform trial enrolling an outpatient population with COVID-19. As one of the UK's national priority trials, the PRINCIPLE trial was established early in the pandemic to assess the effectiveness of repurposed drugs for the treatment of COVID-19. Designing this clinical trial in a rapidly changing environment for a then-unknown disease presented substantial challenges. The aim is to discuss the adaptive and statistical elements of the design that were chosen to facilitate the rapid evaluation of drugs, while also providing the flexibility necessary to accommodate the uncertain and evolving pandemic landscape.

E1178: **Models for longitudinal ordinal outcomes**
Presenter: **Thomas Stewart**, University of Virginia, United States

Many COVID-19 studies captured patient-reported symptom data over several days, including our trial of repurposed medications. We will discuss the estimands, models, and missing data strategy for the longitudinal ordinal outcome we proposed for our trial. We will discuss computational challenges and solutions for implementing the model and missing data strategy, including generating power and sample size calculations for trial planning.

E1170: **Summary outcomes in longitudinal clinical trials: Robustness to MNAR and Modeling Assumptions**
Presenter: **Matt Shotwell**, Vanderbilt University Medical Center, United States

Longitudinal designs are common in randomized clinical trials, but longitudinal assessments are often summarized prior to statistical analysis (e.g., symptom-free days). An alternative longitudinal analysis may be more efficient but also more sensitive to uncertainties regarding missing data and data-generating mechanisms. The efficiency and robustness of the two approaches in estimating a common estimand (e.g., difference in mean symptom-free days) are compared under missing-not-at-random (MNAR) and alternative correlation structures. The context and motivation are two large platform studies in hospitalized patients and outpatients with COVID-19: ACTIV-4 Host Tissue and ACTIV 6. These findings may be useful in selecting an analysis approach for studies that prioritize robustness to such uncertainties.

EO340 Room S-2.25 STATISTICAL LEARNING IN PRACTICE Chair: Alejandro Murua

E1117: **Statistical learning of cyclic autocorrelation functions with application to streamflow data modelling**
Presenter: **Samuel Perreault**, University of Toronto, Canada

A cyclostationnary process is considered and the usual Kendall autocorrelation is modified to account for the dynamic, yet periodic nature of the process under study. This is done by letting the autocorrelation be a function of the usual lag parameter, as well as a periodic time index indicating, e.g., the time of the year. We then describe a learning algorithm whose purpose is to smooth the empirical autocorrelation along both the temporal (cyclic time index) and lag dimensions. We apply it to Canadian daily streamflow data and show how it can be used for testing certain hypotheses, such as the presence of negative dependence within a seasonal cycle.

E0655: **Functional trait locus mapping by functional data clustering**
Presenter: **Marie-Helene Descary**, University of Quebec in Montreal, Canada

The main goal of trait locus mapping (gene mapping) is to identify the locus (gene), i.e. a region of the DNA, that affects a trait of interest. Traditionally, traits of interest were binary (e.g. case vs control) or quantitative (e.g. blood pressure), but it is more and more common to work with functional traits (e.g. growth curve). We will present a new methodology of gene mapping for a functional trait that uses tools developed to analyze functional data, i.e. data that can be seen as realizations of a random function. The idea behind the new method is to translate the problem of identifying a gene associated with a functional trait into the problem of finding the "best" clustering of a set of functions. We consider different measures of functional dissimilarity leading to a flexible gene mapping method, in the sense that it can detect a wide range of effects of a gene on a functional trait. The performance of the new method is assessed with a simulation study and an application to real data.

E1076: **Decision surface Markov chain**
Presenter: **Nicolas Wicker**, University of Lille, France
Co-authors: Amael Broustet

The purpose is to provide a diagnostic tool for supervised learning algorithms once a predictor is obtained. More specifically, we want to assess how complicated a decision surface is. A lot of theory has been developed to assess the complexity of a class of functions (Rademacher complexity, VC dimension, covering numbers), but to our knowledge, tools are lacking to assess the complexity of a particular decision function. The main idea is to use a Markov chain, sampling asymptotically a uniform distribution and hindering its progress whenever it encounters the decision surface, so that the slower it converges, the more complex the surface is supposed to be. This approach is shown in a theoretical example as well as in real data sets.

E1592: **A Bayesian group selection with compositional responses for analysis of tumor proportions and their genomic determinants**
Presenter: **Thierry Chekouo**, University of Minnesota, United States

Volumetric imaging features are used in cancer research to determine the size and the composition of a tumor, and have been shown to be prognostic of overall survival. We focus on the analysis of tumor component proportions of brain cancer patients collected through The Cancer Genome Atlas (TCGA) project. The main goal is to identify pathways and corresponding genes that can explain the heterogeneity of the composition of a brain tumor. In particular, we focus on the glioblastoma multiform (GBM), as it is the most common malignant brain neoplasm.

EO641 Room S-1.04 STATISTICAL MODELS FOR COMPLEX DEPENDENT DATA Chair: Eardi Lila

E0487: **Latent multimodal functional graphical model estimation**
Presenter: **Mladen Kolar**, University of Chicago, United States

Joint multimodal functional data acquisition, where data multiple modes of functional data are measured from the same subject simultaneously, has emerged as an exciting modern approach enabled by recent engineering breakthroughs in the neurological and biological sciences. One prominent motivation for acquiring such data is to enable new discoveries of the underlying connectivity by combining signals from multiple modalities. Yet, despite scientific interest in this problem, there remains a gap in principled statistical methods for estimating the graph underlying joint multimodal functional data. To this end, we propose a new integrative framework to estimate a single latent graphical model from multimodal functional data. We take the generative perspective by modeling the data generation process, from which we identify the inverse mapping from the observation space to the latent space as transformation operators. We then develop an estimator that simultaneously estimates the transformation operators and the latent graph via functional neighborhood regression. The approach is motivated and applied to analyzing simultaneously acquired multimodal brain imaging data where the graph indicates the underlying brain functional connectivity.

E0585: **CLEAN: Leveraging spatial autocorrelation in neuroimaging data in clusterwise inference**
Presenter: **Jun Young Park**, University of Toronto, Canada
Co-authors: Mark Fiecas

While clusterwise inference is a popular approach in neuroimaging that improves sensitivity, current methods do not account for explicit spatial autocorrelations because most use univariate test statistics to construct cluster-extent statistics. Failure to account for such dependencies could result in decreased reproducibility. To address methodological and computational challenges, we propose a new powerful and fast statistical method called CLEAN (Clusterwise inference Leveraging spatial Autocorrelations in Neuroimaging). CLEAN computes multivariate test statistics by modelling brain-wise spatial autocorrelations, constructs cluster-extent test statistics, and applies a refitting-free resampling approach to control false positives. We validate CLEAN using simulations and applications to the Human Connectome Project. This novel method provides a new direction in neuroimaging that paces with advances in high-resolution MRI data, which contains a substantial amount of spatial autocorrelation.

E0746: **Estimation and inference for networks of multi-experiment point processes**
Presenter: **Ali Shojaie**, University of Washington, United States

Modern high-dimensional point process data, especially those from neuroscience experiments, often involve observations from multiple conditions and/or experiments. Networks of interactions corresponding to these conditions are expected to share many edges, but also exhibit unique, condition-specific ones. However, the degree of similarity among the networks from different conditions is generally unknown. To address these needs, we propose a joint estimation procedure for networks of high-dimensional point processes that incorporates easy-to-compute weights in order to data-adaptively encourage similarity between the estimated networks. We also propose a powerful hierarchical multiple testing procedure for edges of all estimated networks, which takes into account the data-driven similarity structure of the multi-experiment networks. Compared to conventional multiple testing procedures, our proposed procedure greatly reduces the number of tests and results in improved power, while tightly controlling the family-wise error rate. Unlike existing procedures, our method is also free of assumptions on the dependency between tests, offers flexibility on p-values calculated along the hierarchy, and is robust to misspecification of the hierarchical structure.

E1401: **Graphical models for stationary time series**
Presenter: **Sumanta Basu**, Cornell University, United States

A spectral precision matrix, the inverse of a spectral density matrix, is an object of central interest in frequency-domain multivariate time series analysis. The estimation of the spectral precision matrix is a key step in calculating partial coherency and graphical model selection of stationary time series. When the dimension of a multivariate time series is moderate to large, traditional estimators of spectral density matrices such as averaged periodograms tend to be severely ill-conditioned, and one needs to resort to suitable regularization strategies. We propose a complex

graphical lasso (cglasso), an l1-penalized estimator of spectral precision matrix based on local Whittle likelihood maximization. We develop fast pathwise coordinate descent algorithms to implement cglasso for large dimensional time series. We also present a complete non-asymptotic theory of our proposed estimator which shows that consistent estimation is possible in a high-dimensional regime as long as the underlying spectral precision matrix is suitably sparse. We illustrate the advantage of cglasso over competing alternatives using extensive numerical experiments on simulated data sets.

EO713 Room S-1.06 RECENT ADVANCES IN IMAGING STATISTICS Chair: Shuo Chen

E0830: **Mediating role of neuroimaging data image-related cognitive decline**
Presenter: **Shuo Chen**, University of Maryland, School of Medicine, United States
Co-authors: Hwiyoung Lee

Aging changes brain functions and structures in a downward trajectory and consequently leads to a decline in neurocognitive performance. Neuroimaging plays a central role in understanding the neurophysiology of brain aging. We consider aging as an independent variable while treating neuroimaging data and cognitive function as the multivariate mediators and outcome, respectively. We aim to investigate the mediation role of multivariate neuroimaging variables in age-related cognitive decline. We present a new multivariate mediation model that maximizes the mediation proportion to reflect that the brain is the primary organ determining cognitive function. Specifically, we consider the aggregating effect of selected neuroimaging variables as a functional brain age score, which can maximally explain age-related cognitive decline. We propose a new objective function that transforms mediation proportion maximization into a quadratic form with an l_1 penalty and l_2 constraint. We develop a computationally efficient algorithm to handle the nonconvexity of the objective function based on the alternating direction method of multipliers with semidefinite relaxation. We apply our method to 37,441 UK Biobank participants with whole-brain cortical thickness and white-matter integrity measures and cognitive performance scores. Our results show that the mediation effect of brain-imaging variables can explain more than 99% of age-related cognitive decline.

E0920: **Semi-parametric independent component analysis in the time-frequency domain**
Presenter: **Seonjoo Lee**, Columbia University/New York State Psychiatric Institute, United States

Independent component analysis (ICA) is a blind source separation method to recover source signals of interest from their mixtures. Most existing ICA procedures are for independent sampling assumptions and are carried out by estimating the marginal density functions. Second-order statistics-based source separation methods have been developed based on parametric time series models for the mixtures from autocorrelated sources. However, when the sources have temporal autocorrelations with mixed spectra, the second-order statistics-based methods cannot separate the sources. To address this issue, we propose a new ICA method by estimating spectral density functions and line spectra of the source signals using cubic splines and indicator functions, respectively. The mixed spectra and the mixing matrix are estimated via maximizing the Whittle likelihood function. Then, we extended this method for the non-stationary sources using time-frequency domain analysis. We illustrate the performance of the proposed method through simulation experiments and a resting stage EEG data application. The numerical results indicate that our approach outperforms existing ICA methods, including SOBI algorithms.

E1566: **Time-varying functional connectivity features are strong predictors of Alzheimers disease**
Presenter: **Fei Jiang**, The University of California, San Francisco, United States

Dynamic resting state functional connectivity (RSFC) characterizes time-varying fluctuations of functional brain network activity. Considered superior to static functional connectivity, it has been unclear whether features of dynamic functional connectivity are associated with neurodegenerative diseases. Popular sliding-window and clustering methods for extracting dynamic RSFC have various limitations preventing them from extracting reliable features to address this question. We use a novel and robust time-varying dynamic network (TVDN) approach to extract the dynamic RSFC features from high-resolution magnetoencephalography data. This algorithm automatically and adaptively learns the low-dimensional manifold of dynamic RSFC and detects dynamic state transitions in data. We show that both the number of transitions, dwell times, and the number of brain states are strong predictors of Alzheimer's disease (AD). Furthermore, these dynamic features from TVDN have high sensitivity and specificity in distinguishing AD and healthy subjects. These results indicate that robust dynamic resting-state functional connectivity features are impacted in dementias like Alzheimer's disease, and may be crucial to understanding the neuropathological disease impact and trajectory.

E1586: **Bayesian Bootstrap uncertainty quantification for spatial lesion regression modelling**
Presenter: **Anna Menacher**, University of Oxford, United Kingdom
Co-authors: Thomas Nichols, Chris Holmes, Habib Ganjgahi

Neural demyelination and brain damage accumulated in white matter appear as hyperintense areas on MRI scans in the form of lesions. Modeling binary images at the population level, where each voxel represents the existence of a lesion, plays an important role in understanding aging and inflammatory diseases. We propose a scalable hierarchical Bayesian spatial model, called BLESS, capable of handling binary responses by placing continuous spike-and-slab mixture priors on spatially-varying parameters and enforcing spatial dependency on the parameter dictating the amount of sparsity within the probability of inclusion. The use of mean-field variational inference with dynamic posterior exploration, which is an annealing-like strategy that improves optimization, allows our method to scale to large sample sizes. The method also accounts for underestimation of posterior variance due to variational inference by providing an approximate posterior sampling approach based on Bayesian bootstrap ideas and spike-and-slab priors with random shrinkage targets. Besides accurate uncertainty quantification, this approach is capable of producing novel cluster size based imaging statistics, such as credible intervals of cluster size, and measures of reliability of cluster occurrence. Lastly, we validate our results via simulation studies and an application to the UK Biobank, a large-scale lesion mapping study with a sample size of 40,000 subjects.

EO450 Room S-1.27 STATISTICAL METHODS FOR WEARABLE DEVICES Chair: Jaroslaw Harezlak

E0761: **A Riemann manifold model framework for longitudinal changes in physical activity patterns**
Presenter: **Jingjing Zou**, University of California, San Diego, United States
Co-authors: Chongzhi Di, Loki Natarajan

The wide usage of wearable accelerometer-based activity trackers in recent years has provided a unique opportunity for in-depth research on physical activity (PA) and its relationship with health outcomes and interventions. Past analysis of activity tracker data relies heavily on aggregating minute-level PA records into day-level summary statistics, in which important information on diurnal PA patterns is lost. We propose a novel functional data analysis approach based on the theory of Riemann manifolds for modeling PA records and longitudinal changes in PA temporal patterns. We model smoothed minute-level PA of a day as one-dimensional Riemann manifolds and longitudinal changes in PA in different visits as deformations between manifolds. With the proposed approach, we conduct comprehensive analyses of data from two clinical trials, focusing on the effect of interventions on longitudinal changes in PA patterns and how different patterns of changes in PA influence weight loss, respectively. For both studies, important modes of variation in PA were identified to be significantly associated with lifestyle interventions/health outcomes.

E0862: **Quantile regression with a mixture of function-valued and a scalar-valued covariate prone to classical measurement error**
Presenter: **Carmen Tekwe**, Indiana University - Bloomington, United States

Current recommendations for dietary intake (DI) and physical activity (PA) to minimize risks for chronic health conditions are based on statistical analyses of data prone to measurement error, including those collected from self-reported questionnaires and wearable devices. Self-reported measures based on food frequency questionnaires are often used in DI assessments; however, they are prone to recall bias. Wearable devices

enable the continuous monitoring of PA but generate complex functional data with poorly characterized systematic errors. We propose the quantile regression model with function- and scalar- valued covariates prone to measurement errors. We develop semiparametric and parametric approaches to correct measurement errors associated with the mixture of functional and scalar covariates prone to errors in quantile regression settings. Simulations are performed to assess the finite sample properties. The developed methods are applied to investigate the influence of wearable-device-based PA and self-reported measures of total caloric intake on the quantile function of body mass index (BMI). The device-based measures of PA are assumed to be prone to functional covariates prone to complex arbitrary heteroscedastic errors. In contrast, DI is assumed to be a scalar-valued covariate prone to error. The developed methods are applied to assess the relationship between PA and DI with quantile functions of BMI among community-dwelling adults living in the US.

E1152: **Mediation analysis with densities as mediators with an application to iCOMPARE trial**
Presenter: **Jingru Zhang**, University of Pennsylvania, United States
Co-authors: Haochang Shou, Hongzhe Li

Physical activity has long been shown to be associated with biological and physiological performance and the risk of diseases. It is of great interest to assess whether the effect of an exposure or intervention on an outcome is mediated through physical activity measured by modern wearable devices such as actigraphy. However, existing methods for mediation analysis focus almost exclusively on mediation variable that is in the Euclidean space, which cannot be applied directly to the actigraphy data of physical activity. Such data is best summarized in the form of a random histogram or random density. We develop the structural equation models (SEMs) to the settings where a random density is treated as the mediator to study the indirect mediation effect of physical activity on an outcome. We provide sufficient conditions for identifying the average causal effects of a density mediator and present methods for estimating the direct and mediating effects of a density on an outcome. We apply our method to the data set from the iCOMPARE trial that compares flexible duty-hour policies and standard duty-hour policies on interns' sleep-related outcomes to explore the mediation effect of physical activity on the causal path between flexible duty-hour policies and sleep-related outcomes.

E1777: **Robust functional principal components analysis with application to accelerometry data**
Presenter: **Chongzhi Di**, Fred Hutchinson Cancer Center, United States
Co-authors: Guangxing Wang, Fang Han

Accelerometers are widely used to measure physical activity in biomedical studies objectively. They collect high-resolution functional data, which are often highly skewed and have outliers. Standard functional principal component analysis (FPCA) is based on empirical covariance operators and might not work well in these settings. To address these challenges, we propose a new robust approach for FPCA, based on a functional pairwise spatial sign operator (PASS). Theoretical properties of the proposed method are established. In particular, it is shown that the PASS has the same set of eigenfunctions as the standard covariance operator and that their corresponding eigenvalues are in the same order. Through extensive simulation studies, the proposed robust FPCA is shown to perform well under various types of functional data. We applied the method to an ancillary study of the Women Health Initiative that recorded 7-day accelerometry data on 6500 women.

EO536 Room K0.16 MODELING VARIOUS DATA TYPES WITH NETWORK STRUCTURES Chair: Philip White

E0436: **Mutually exciting point process graphs for modelling dynamic networks**
Presenter: **Francesco Sanna Passino**, Imperial College London, United Kingdom
Co-authors: Nick Heard

A new class of models for dynamic networks is proposed, called mutually exciting point process graphs (MEG). MEG is a scalable network-wide statistical model for point processes with dyadic marks, which can be used for anomaly detection when assessing the significance of future events, including previously unobserved connections between nodes. The model combines mutually exciting point processes to estimate dependencies between events and latent space models to infer relationships between the nodes. The intensity functions for each network edge are characterised exclusively by node-specific parameters, which allow information to be shared across the network. This construction enables estimation of intensities even for unobserved edges, which is particularly important in real-world applications, such as computer networks arising in cyber-security. A recursive form of the log-likelihood function for MEG is obtained, which is used to derive fast inferential procedures via modern gradient ascent algorithms. An alternative EM algorithm is also derived. The model and algorithms are tested on simulated graphs and real-world datasets, demonstrating excellent performance.

E0497: **Space-time covariance models on networks**
Presenter: **Jun Tang**, University of Iowa, United States
Co-authors: Dale Zimmerman

The second-order, small-scale dependence structure of a stochastic process defined in the space-time domain is key to prediction (or kriging). While great efforts have been dedicated to developing models for cases in which the spatial domain is either a finite-dimensional Euclidean space or a sphere, counterpart developments on a generalized linear network are practically non-existent. To fill this gap, we develop a broad range of parametric, non-separable space-time covariance models on generalized linear networks. For the important subgroup of Euclidean trees, we develop models by the space embedding technique, in concert with the generalized Gneiting class of models and 1-symmetric characteristic functions, and by the convex cone and scale mixture approaches. We give examples from each class of models and investigate the geometric features of these covariance functions near the origin and at infinity. We also reveal connections between different classes of space-time covariance models on Euclidean trees. We conclude by investigating the performance of maximum likelihood estimators of certain proposed models in a simulation study.

E0503: **Stochastic epidemic models on dynamic contact networks**
Presenter: **Fan Bu**, UCLA, United States
Co-authors: Allison Aiello, Jason Xu, Alexander Volfovsky

Infectious disease transmission relies on interpersonal contact networks, but traditional epidemic models often assume a random-mixing population where all individuals are equally likely to get infected. We seek to develop a more realistic and generalized stochastic epidemic model that considers transmission over dynamic networks. We propose a joint epidemic-network model through a continuous-time Markov chain such that disease transmission is constrained by the contact network structure, and network evolution is, in turn, influenced by individual disease statuses. To accommodate partial epidemic observations commonly seen in real-world data, we design efficient inference algorithms through data augmentation that leverages dynamic network features and infection mechanisms. At the same time, our approach can account for individual heterogeneity and explore intervention effects on disease transmission. Experiments on both synthetic and real datasets demonstrate that our inference method can accurately and efficiently recover model parameters and provide valuable insight into epidemic data in the presence of unobserved disease episodes.

E0613: **Scalable warped directional traffic network models for traffic accident data**
Presenter: **Philip White**, Brigham Young University, United States

Traffic accident data from the Utah Department of Transportation are considered. Specifically, we consider six years of traffic accidents from I-15, the most heavily trafficked highway in Utah, USA. In total, there are 48,704 traffic accidents. For these data, we propose a scalable model to identify roadway characteristics associated with more traffic accidents, learn dynamics in accident patterns, capture nonstationary patterns present in the data, and forecast future accident patterns. We present a scalable framework for approximating log-Gaussian Cox processes. In addition, we use spatial warping to capture non-stationary patterns in traffic accident data. We also include dynamics and direction in the regression coefficients

and the spatial model to capture year-to-year and directional variation in accident patterns. We compare various model specifications. For the best model, discuss the results of this model in this dataset.

EO326 Room K0.19 MODELING AND COMPUTING FOR HETEROGENEOUS AND CLUSTERED DATA Chair: Weixin Yao

E0538: **Marginal accelerated failure time mixture cure model for clustered survival data**
Presenter: **Yi Niu**, Dalian University of Technology, China
Co-authors: Duze Fan, Yingwei Peng

The semiparametric accelerated failure time (AFT) mixture cure model is an appealing alternative to the semiparametric proportional hazards mixture cure model in analyzing multivariate failure time data with long-term survivors. However, the former received less attention than the latter due to the complexity of the estimation method for the former, and the model was not proposed for clustered survival data. We consider a marginal semiparametric AFT mixture cure model for clustered failure time data with a potential cure fraction. We propose a generalized estimating equations (GEE) approach based on the Expectation-Solution (ES) algorithm to estimate the regression parameters in the model. The correlation structures within clusters are modeled by working correlation matrices in the GEE. We use a bootstrap method to obtain the variances of the estimators. Numerical studies demonstrate that the efficiency gain of the regression coefficient estimators is robust to the misspecification of working matrices, and the efficiency is higher when the working correlation structure is closer to the truth. Finally, we apply the model and the proposed method to analyze the data from a smoking cessation study and a tonsil cancer study for illustration.

E0688: **Bayesian model-based clustering with the telescoping sampler**
Presenter: **Bettina Gruen**, WU (Vienna University of Economics and Business), Austria
Co-authors: Sylvia Fruehwirth-Schnatter, Gertraud Malsiner-Walli

Mixtures of finite mixtures (MFM) are a suitable model class for model-based clustering in a Bayesian framework. In MFMs, a prior is also specified for the number of components in the finite mixture model to account for the uncertainty regarding the number of clusters. We discuss suitable prior specifications for the MFM model in model-based clustering applications as well as possible methods for inspecting implicitly induced prior distributions on the number of data clusters and partitions of the observations. Assuming that only a small number of data clusters are observed in the data set suggests using a dynamic specification of the weight prior where the gap between the number of components and the number of data clusters a-priori increases with an increasing number of components. For inference of the dynamic MFM, we propose to use the telescoping sampler, which extends the Markov chain Monte Carlo sampling scheme with data augmentation of the finite mixture model with a fixed number of components by sampling also from the posterior of the number of components. We will demonstrate the general applicability and performance of the telescoping sampler on mixture models with different component models.

E0819: **Modelling heterogeneity in human gut microbiome: A clustering-based perspective**
Presenter: **Angela Montanari**, Universita di Bologna, Italy
Co-authors: Laura Anderlucci, Silvia Dallari

Microbiota are largely recognized as being central players in human health and in that of all organisms and ecosystems, and have been the subject of intense study. Next-generation sequencing techniques proved very effective for characterizing microbial communities by sequencing suitable molecular targets such as 16S ribosomal RNA gene amplicons for bacteria. However, the analysis and translation of microbiome data into meaningful biological insights remains very challenging. Firstly, microbiome data are compositional, i.e. microbial counts represent proportions instead of absolute abundances. Secondly, sparsity in the dataset can lead to false associations of microorganisms; a zero indicates either the absence of a microorganism, or an insufficient sequencing depth. Thirdly, it is challenging to differentiate between direct and indirect associations, in particular, if these are related to environmental factors. Different methods designed to account for unobserved heterogeneity are studied and compared on a real data set example.

E1984: **Adaptive distributed inference for multi-source massive heterogeneous data**
Presenter: **Mixia Wu**, Beijing University of Technology, China

Distributed inference for a heterogeneous linear model with massive datasets is considered. The goal is to extract common features across all subpopulations and explore the heterogeneity of each subpopulation. Noticing that heterogeneity may exist not only in expectations of subpopulations but also in their variances, we propose the heterogeneity-adaptive distributed aggregation (HADA) estimation, which is shown to be communication-efficient and asymptotically optimal, irrespective of homoscedasticity or heteroscedasticity. Furthermore, a distributed test for parameter heterogeneity across sub-populations is constructed based on the HADA estimator. The finite-sample performance of the proposed methods is evaluated via simulation studies and bike-share data.

EO619 Room K0.20 ROC METHODS FOR THE EVALUATION OF BIOMARKERS Chair: Leonidas Bantis

E0342: **Covariate adjustment methods for the evaluation of biomarkers in multi-class setting**
Presenter: **Duc Khanh To**, University of Padova, Italy
Co-authors: Gianfranco Adimari, Monica Chiogna

The statistical evaluation of a biomarker plays an important role in medical research, but the evaluating process is often done marginally, i.e., by using the biomarkers values only. In some cases, however, there are covariates, for instance, age, gender, and smoking status, that can influence the biomarker behavior, and, therefore, also impact its accuracy. Thus, in practice, the evaluation of such possible effects is needed. Recently, various methods have been developed to address possible covariates' effects on the evaluation of biomarkers. Most proposed methods focus on a two-class setting, whereas a multi-class setting is very scarcely considered in the statistical literature. We will introduce our new proposed methods to evaluate the accuracy of biomarkers in the presence of covariates, and will provide some discussion on the future development of this topic.

E0983: **Estimation and inference on the partial volume under the ROC surface with applications to pancreatic cancer**
Presenter: **Katherine Young**, University of Kansas Medical Center, United States
Co-authors: Leonidas Bantis

Summary measures of biomarker accuracy that employ the receiver operating characteristic (ROC) surface have been proposed for biomarkers that classify patients into one of three groups: healthy, benign, or aggressive disease. The volume under the ROC surface (VUS) summarizes the overall discriminatory ability of a biomarker in such configurations, but includes cutoffs associated with clinically irrelevant true classification rates (TCR's). Due to the lethal nature of pancreatic cancer, cutoffs associated with a low TCR for identifying patients with pancreatic cancer may be undesirable and not appropriate for use in a clinical setting. In this project, we study the properties of a more focused criterion, the partial VUS (pVUS), that summarizes the diagnostic accuracy of a marker in the three-class setting for regions restricted to only those of clinical interest. We propose methods for estimation and inference on the pVUS under parametric and non-parametric frameworks and apply these methods to the evaluation of potential biomarkers for the diagnosis of pancreatic cancer.

E1086: **Nonparametric methods for the comparison of ROC curves with covariate information**
Presenter: **Aris Fanjul Hevia**, Universidad de Oviedo, Spain
Co-authors: Juan-Carlos Pardo-Fernandez, Wenceslao Gonzalez-Manteiga

The Receiver Operating Characteristic (ROC) curve is a statistical tool that combines the notions of sensitivity and specificity to evaluate the

discriminatory capability of a classification problem. Whenever more than one classification procedure is available, the ROC curves may be used for comparing their behaviour. Furthermore, these methods may be affected by the information provided by some covariates, and thus they should be taken into account when comparing the corresponding ROC curves. There are several ways to incorporate the covariates into a ROC curve analysis, the main ones being the conditional ROC curves and the covariate-adjusted ROC curves. We use nonparametric techniques to study these curves with the aim of comparing methods of classification and, at the same time deciding whether the covariate information should be taken into account or not.

E1140: **Multiple comparisons of areas under the ROC curve**
Presenter: **Jeremie Riou**, Universite de Angers, France
Co-authors: Paul Blanche

Comparing areas under the ROC curve is a popular approach to comparing prognostic biomarkers. The aim is to present an efficient method to control the family-wise error rate when multiple comparisons are performed. We suggest combining the max-t test and the closed testing procedures. We build on previous work on asymptotic results for ROC curves and on general multiple testing methods to efficiently take into account both the correlations between the test statistics and the logical constraints between the null hypotheses. The proposed method results in a uniformly more powerful procedure than both the single-step max-t test procedure and popular stepwise extensions of the Bonferroni procedure, such as BonferroniHolm. As demonstrated, the method can be applied in most usual contexts, including the time-dependent context with right censored data. We show how the method works in practice through a motivating example where we compare several psychometric scores to predict the t-year risk of Alzheimer's disease. The example illustrates several multiple testing settings and demonstrates the advantage of using the proposed methods over common alternatives. R code has been made available to facilitate the use of the methods by others.

EO028 Room K0.50 ADVANCES IN DESIGN OF EXPERIMENTS Chair: Tim Waite

E0523: **General Stratum Orthogonal Arrays (GSOAs): Constructions and properties**
Presenter: **Ulrike Groemping**, Berliner Hochschule fuer Technik, Germany

Stratum Orthogonal Arrays (SOAs) and their generalization to General Stratum Orthogonal Arrays (GSOAs) were proposed for experimenting with quantitative variables in computer experiments. A review of SOA constructions, using a unifying notation in the form of simple equations, has been recently given. (G)SOAs can be latin hypercube designs or have much fewer distinct values for each experimental variable. Designs for computer experiments are commonly requested to have good space-filling properties. Many different metrics for assessing space-filling can be found in the literature. An entirely different approach has been presented recently: a stratification pattern derived from a generalized word length pattern captures the stratification properties of (G)SOAs, which can also be considered an aspect of space-filling. The purpose is to present general ideas behind (G)SOAs and their constructions and explain a recent stratification pattern. The R package SOAs allows us to apply all the constructions and assessments. Note that SOAs were initially named "strong orthogonal arrays" although they usually have very low orthogonal array strength; this contradiction is the reason why this author has repurposed the acronym to "stratum orthogonal arrays".

E0842: **Locally optimal design for A/B tests in the presence of covariates and network dependence**
Presenter: **Lulu Kang**, Illinois Institute of Technology, United States

A/B test, a simple type of controlled experiment, refers to the statistical procedure of conducting an experiment to compare two treatments applied to test subjects. For example, many IT companies frequently conduct A/B tests on their users who are connected and form social networks. Often, the users' responses could be related to the network connection. We assume that the users, or the test subjects of the experiments, are connected on an undirected network, and the responses of two connected users are correlated. We include the treatment assignment, covariate features, and network connection in a conditional autoregressive model. Based on this model, we propose a design criterion that measures the variance of the estimated treatment effect and allocate the treatment settings to the test subjects by minimizing the criterion. Since the design criterion depends on an unknown network correlation parameter, we adopt the locally optimal design method and develop a hybrid optimization approach to obtain the optimal design. Through synthetic and real social network examples, we demonstrate the value of including network dependence in designing A/B experiments and validate that the proposed locally optimal design is robust to the choices of parameters.

E0796: **Gibbs optimal design of experiments**
Presenter: **Antony Overstall**, University of Southampton, United Kingdom

Gibbs (or generalised Bayesian) inference is a generalisation of Bayesian inference made by replacing the log-likelihood in Bayes' theorem by a (negative) loss function. The loss function identifies desirable parameter values for given responses. The advantage of Gibbs inference over traditional Bayesian inference is that it does not require the specification of a probabilistic data-generating process and, therefore, should be less sensitive to this process. Gibbs optimal design of experiments are proposed for this inferential framework, extended decision-theoretic Bayesian optimal design. The challenge is that the decision-theoretic approach relies on a probabilistic data-generating process that is notably absent from Gibbs inference. This is circumvented by assuming a designer model: a probabilistic data-generating process which is only used to find a design rather than in the ensuing inference. Because of this, the designer model can encapsulate very general data-generating processes with the aim of introducing robustness into the design procedure. The proposed Gibbs optimal design framework is demonstrated in several illustrative examples.

E1265: **Optimal designs for the detection and characterisation of hormesis in toxicological tests**
Presenter: **Sergio Pozuelo Campos**, University of Castilla-La Mancha, Spain
Co-authors: Victor Casero-Alonso, Mariano Amo-Salas

Toxicological tests are experiments that allow the effect of a toxicant on organisms, ecosystems, etc., to be determined. The focus is on tests in the aquatic environment, in which the Ceriodaphnia Dubia test has a particular interest. It has been found that hormesis occurs in two out of three of the experiments carried out with this organism. We apply the theory of optimal design of experiments to a linear quadratic model with Poisson distribution to obtain designs that allow us to detect and characterise hormesis efficiently. For this purpose, different utility functions, such as the dose at which the maximum concentration is reached, the AUC, the ZEP, the RIp or the RIMp are used.

EO638 Room S0.03 SPATIAL STATISTICS AND STOCHASTIC PDES Chair: David Bolin

E0329: **Non-separable diffusion-based spatio-temporal Gaussian fields**
Presenter: **Finn Lindgren**, University of Edinburgh, United Kingdom
Co-authors: David Bolin, Elias Krainski, Haavard Rue, Haakon Bakka

Gaussian random fields with Matern covariance functions are popular models in spatial statistics and machine learning. The easiest approach to constructing space-time models is by taking the product between a spatial covariance kernel and a temporal covariance kernel. However, these space-time separable models have both theoretical and practical drawbacks. An alternative is to take advantage of temporal extensions of the spatial Whittle-Matern model to define non-separable models as solutions to stochastic partial differential equations. Such models can keep the marginal Matern covariance in space, but besides the parameters of the spatial covariance (variance, smoothness, and practical correlation range), they include parameters controlling the practical correlation range in time, the smoothness in time, and the degree of non-separability of the spatio-temporal covariance. A sparse representation based on a finite element approximation can be constructed in closed form for the Markovian subset of models, which is well suited for statistical inference on flat domains, spheres, as well as other manifolds. This has been implemented in the

R-INLA software. The full range of models can be handled either through spectral representations or by using new methods for fractional operators. The flexibility of the model is illustrated in an application to spatio-temporal modelling of global temperature data.

E0874: **Finite element and graphical representations of Gaussian processes**
Presenter: **Daniel Sanz-Alonso**, University of Chicago, United States
Co-authors: Ruiyi Yang

Gaussian processes (GPs) are popular models for random functions in computational and applied mathematics, statistics, machine learning and data science. However, GP methodology scales poorly to large data sets due to the need to factorize a dense covariance matrix. In spatial statistics, a standard approach to surmount this challenge is to represent Matern GPs using finite elements, obtaining an approximation with a sparse precision matrix. A new understanding of this approach will be given for regression and classification with large data sets, showing that under mild smoothness assumptions, the dimension of the matrices that need to be factorized can be reduced without hindering the estimation accuracy. The analysis balances finite element and statistical errors to show that there is a threshold beyond which further refining of the discretization increases the computational cost without improving the estimation accuracy. We will also introduce graphical representations of GPs to model random functions on high-dimensional point clouds, greatly expanding the important but limited scope of the finite element approach. We will show error bounds on the graphical representations, and study the associated posterior contraction in a semi-supervised learning problem.

E0993: **Gaussian Whittle Matern fields on metric graphs**
Presenter: **Jonas Wallin**, Lund University, Sweden
Co-authors: David Bolin, Alexandre de Bustamante Simas

A new class of Gaussian processes on compact metric graphs such as street or river networks is defined. The proposed models, the Whittle Matern fields, are defined via a fractional stochastic partial differential equation on the compact metric graph and are a natural extension of Gaussian fields with Matern covariance functions on Euclidean domains to the non-Euclidean metric graph setting. The existence of the processes, as well as their sample path regularity properties, are derived. The processes are stable in the sense that they do not change when vertices are added to existing edges of the graph, or when vertices of degree two are removed. This property is important for applications and is lacking for Gaussian processes based on the graph Laplacian on non-metric graphs. The model class contains, as particular cases, differentiable Gaussian processes. This is the first construction of a valid differentiable Gaussian field on general compact metric graphs.

E1556: **The SPDE approach for spatio-temporal datasets with advection and diffusion: A matrix-free approach**
Presenter: **Lucia Clarotto**, Mines Paris, PSL University, France
Co-authors: Denis Allard, Thomas Romary, Nicolas Desassis

In the task of predicting spatio-temporal fields in environmental science, introducing models inspired by the physics of the underlying phenomena that are numerically efficient is of growing interest in spatial statistics. The size of space-time datasets calls for new numerical methods to process them efficiently. The SPDE (Stochastic Partial Differential Equation) approach has proven to be effective for the estimation and prediction in a spatial context. We present the advection-diffusion SPDE with first-order derivative in time to enlarge the SPDE family to the space-time context. By varying the coefficients of the differential operators, the approach allows us to define a large class of non-separable spatio-temporal models. A Gaussian Markov random field approximation of the solution of the SPDE is built by discretizing the temporal derivative with a finite difference method (implicit Euler) and by solving the purely spatial SPDE with a finite element method (continuous Galerkin) at each time step. The Streamline Diffusion stabilization technique is introduced when the advection term dominates the diffusion term. Computationally efficient methods are proposed to estimate the parameters of the SPDE and to predict the spatio-temporal field by kriging. The approach is applied to a solar radiation dataset.

EO444 Room S0.11 BRANCHING PROCESSES: THEORY, COMPUTATION AND APPLICATIONS I Chair: Miguel Gonzalez Velasco

E1367: **Algorithmic methods for parameter estimation in two-type branching processes**
Presenter: **Vessela Stoimenova**, Sofia University, Bulgaria
Co-authors: Ana Staneva, Dimitar Atanasov

Different approaches to parameter estimation in the multitype branching setting are discussed. We consider and compare various approximation methods which allow for an algorithmic estimation in the presence of hidden data. We provide examples of two-type discrete time branching processes, their simulations and numerical results.

E1421: **Simulative study of Markov branching processes derived from geometric reproduction of particles**
Presenter: **Assen Tchorbadjieff**, Bulgarian Academy of Sciencies, Bulgaria

The evolution of any branching particle system depends on the reproduction law and the lifetime of particles. Usually, for many processes, these parameters of reproductions follow different natural rules and stochastic conditions. A possible approach for their study is to generalise reproduction by any probabilistic distribution. The primary case is to explain reproduction laws using Geometric distribution. The analytical results for sub-critical and critical for this case are yielded using special functions. However, the solution is unclear when more complicated distributions are considered. In these cases, we implement a computational approach using stochastic simulation methods. Using the supercomputer facility, we are studying reproduction mechanisms based on Negative Binomial and Poisson distributions. The primary results are presented.

E1460: **Ancestral inference for age-dependent branching processes with immigration**
Presenter: **Anand Vidyashankar**, George Mason University, United States

Age-dependent branching processes are routinely used to model various dynamical models arising in biological and physical systems. In some of these applications, the initial number of ancestors initiating the process is unknown, and it is of critical importance to estimate the parameters associated with the ancestral distribution for addressing meaningful scientific questions. We describe a new estimated martingale technique to develop an estimator for the parameters of the ancestral distribution. We establish the asymptotic properties of the estimators under both the true model and under model misspecification. Extensions to birth-death processes will also be indicated.

E1886: **Large deviation results for controlled branching processes with immigration**
Presenter: **Ines M del Puerto**, University of Extremadura, Spain
Co-authors: Miguel Gonzalez Velasco, Carmen Minuesa Abril, Anand Vidyashankar

A control branching process (CBP) is a generalization of Byenaime-Galton-Watson processes where at each generation, the number of progenitors is randomly chosen through a random control function. We modify a CBP, including the possibility of immigration of individuals at each generation. The aim is to give several large deviation results for the CBP with immigration. We consider the supercritical case and establish the rate of convergence of the process normalized by the number of progenitors to the offspring mean and of the ratio of successive generations to the growth rate of the process. We analyze the large deviations under an assumption on the exponential moments of the offspring and immigration distributions and also based on the asymptotic behaviour of the harmonic moments of the generation and control sizes.

EO148 Room S0.13 STATISTICAL SUMMITS: METHODOLOGY AND COMPUTING I Chair: Andriette Bekker

E1033: **Mixture modeling of mixed type data for clustering**
Presenter: **Michael Gallaugher**, Baylor University, United States

Co-authors: Eman Alamer, Paul McNicholas

Clustering, also known as unsupervised classification, forms the foundation of machine learning techniques and is used to find underlying group structures in data. There are many well-established model-based techniques to analyze either categorical or continuous data in the clustering paradigm; however, there is a relative paucity of work for mixed-type data, especially mixed data where the continuous variables exhibit skewness and/or heavy tails. We consider two different avenues. This first is to consider the case where the continuous variables exhibit skewness and heavy tails. In this case, we consider combining a latent variable model with the skew-t distribution for modelling the continuous variables. The second is to consider outlier detection for mixed-type data by combining a latent variable model with the contaminated normal distribution. Both simulated and real data will be used for illustration.

E1072: **Improved inference for LGM's using INLA**

Presenter: **Janet Van Niekerk**, King Abdullah University of Science and Technology, Saudi Arabia

Co-authors: Denis Rustand, Elias Krainski, Haavard Rue

INLA has gained popularity as an approximate inferential method due to its accuracy and efficiency for the class of latent Gaussian models. We present recent advancements in the methodology, implemented in the R-INLA library, that performs approximate Bayesian inference even faster while achieving the same accuracy. This advancement avails INLA as an inferential framework for huge data such as fMRI analysis and complex joint survival models.

E1133: **Some remarks about maximization of the likelihood function**

Presenter: **Adelchi Azzalini**, Universita di Padova, Italy

In most practical cases, maximum likelihood estimation is performed via some numerical method. Within this frame, there exists a range of alternatives in current use, and the question of choosing the more convenient option for a given problem does not usually have a clear-cut answer. We discuss some issues and points to consider in this context, with special emphasis on the contrast between EM-type algorithms and those based on direct numerical maximization.

E1402: **Conditional mixture modeling**

Presenter: **Volodymyr Melnykov**, The University of Alabama, United States

Modern mixture models are complex and often under risk of overparameterization. To address this concern, one popular approach is to consider various parsimonious models obtained by introducing constraints on covariance matrix structures. We propose an alternative approach to parsimonious mixture modeling that is based on modeling location rather than dispersion parameters. The developed model proves to be flexible, especially in the presence of non-compact clusters. Fast parameter estimation is possible due to the availability of analytical expressions. Conducted simulation studies, as well as applications to well-known classification data sets, demonstrate the promise and competitiveness of the proposed model.

EO404 Room Virtual R01 RECENT ADVANCES IN BAYESIAN MODELING AND COMPUTATION Chair: Marco Ferreira

E0734: **Bayesian analysis of GLMMs with nonlocal priors for genome-wide association studies**

Presenter: **Marco Ferreira**, Virginia Tech, United States

Co-authors: Shuangshuang Xu, Jacob Williams

A novel Bayesian method is presented to find single nucleotide polymorphisms (SNPs) associated with particular phenotypes measured as discrete data from genome-wide association studies (GWAS). This is a regression problem with p two to three orders of magnitude larger than n, the subjects are correlated, and the SNPs regressors are highly correlated. To deal with these challenges, we propose nonlocal priors specifically tailored to GLMMs and develop related fast approximate computations for Bayesian model selection. To search through hundreds of thousands of possible SNPs, we use a two-step procedure: first, we screen for candidate SNPs; second, we perform model search that considers all screened candidate SNPs as possible regressors. A simulation study shows favorable performance of our Bayesian method when compared to other methods widely used in the GWAS literature. We illustrate our method with applications to the analysis of real GWAS datasets from plant science and human health.

E0876: **Objective Bayesian model selection for generalized linear mixed models**

Presenter: **Shuangshuang Xu**, Virginia Tech, United States

Co-authors: Marco Ferreira, Erica Porter, Christopher Franck

An objective Bayesian model selection approach is proposed for generalized linear mixed models. Since random effects cannot be integrated out of GLMMs analytically, we approximate the integrated likelihood function using a pseudo-likelihood approach. We study performance assuming an approximate reference prior for the parameters of the model. In addition to the approximate reference prior, we also consider the half-Cauchy prior for the square root of variance components of the random effects. Since the approximate reference prior is improper, we develop a fractional Bayes factor approach with a minimum training fraction. We then perform model selection based on the resulting posterior probabilities of the several competing models. Simulation studies with Poisson generalized linear mixed models with spatial random effects and overdispersion random effects show that our approach performs favorably when compared to widely used competing Bayesian methods, including DIC and WAIC. We illustrate the usefulness and flexibility of our approach with three case studies, including a Poisson longitudinal model, a Poisson spatial model, and a logistic mixed model.

E0992: **BGWAS: Bayesian variable selection in linear mixed models with nonlocal priors for genome-wide association studies**

Presenter: **Jacob Williams**, Virginia Tech, United States

Co-authors: Shuangshuang Xu, Marco Ferreira

We propose BGWAS, a novel Bayesian variable selection method based on nonlocal priors for linear mixed models specifically tailored for genome-wide association studies. Genome-wide association studies (GWAS) seek to identify single nucleotide polymorphisms (SNPs) that cause observed phenotypes. However, with highly correlated SNPs, correlated observations, and the number of SNPs being two orders of magnitude larger than the number of observations, GWAS procedures often suffer from high false positive rates. To correct for this, our proposed method BGWAS uses a novel nonlocal prior for linear mixed models (LMMs). BGWAS has two steps: screening and model selection. The screening step scans through all the SNPs fitting one LMM for each SNP and then uses Bayesian false discovery control to select a set of candidate SNPs. After that, a model selection step searches through the space of LMMs that may have any number of SNPs from the candidate set. A simulation study shows that, when compared to popular GWAS procedures, BGWAS greatly reduces false positives while maintaining the same ability to detect true positive SNPs. We show the utility and flexibility of BGWAS with two case studies: a case study on salt stress in plants, and a case study on alcohol use disorder.

E1172: **Smooth covariance functions for multiscale spatiotemporal models**

Presenter: **Thais C O Fonseca**, Universidade Federal do Rio de Janeiro, Brazil

Co-authors: Marco Ferreira

Spatial covariance functions resulting from multiscale models are step functions given the discrete partition of the spatial domain of interest. However, the original process might be smooth, and step functions might not be a good characterization of spatial dependence. In that context, we derive smooth spatiotemporal covariance structures, which are obtained by averaging spatially shifted versions of the original multiscale model. To illustrate the proposal, we consider a dyadic tree data structure with L levels of resolution and present the smooth original covariance functions and

the suggested mixture approximation. In particular, we consider several specifications of the model parameters in order to illustrate the possible behaviours of the implied spatiotemporal covariance functions. Posterior independence of model parameters for different tree guarantee scalability for large data analysis.

EO230 Room Virtual R02 ADVANCES IN HIGH-DIMENSIONAL STATISTICS Chair: Yang Ning

E0831: **Optimal and safe estimation for high-dimensional semi-supervised learning**
Presenter: **Yang Ning**, Cornell University, United States

There are many scenarios, such as electronic health records, where the outcome is much more difficult to collect than the covariates. We consider the linear regression problem with such a data structure under high dimensionality. Our goal is to investigate when and how unlabeled data can be exploited to improve the estimation and inference of the regression parameters in linear models, especially in light of the fact that such linear models may be misspecified in data analysis. In particular, we address the following two important questions. (1) Can we use the labeled data as well as the unlabeled data to construct a semi-supervised estimator such that its convergence rate is faster than the supervised estimators? (2) Can we construct confidence intervals or hypothesis tests that are guaranteed to be more efficient or powerful than the supervised estimators?

E1157: **Optimal tuning-free convex relaxation for noisy matrix completion**
Presenter: **Cong Ma**, University of Chicago, United States
Co-authors: Yuepeng Yang

The focus is on noisy matrix completion, the problem of recovering a low-rank matrix from partial and noisy entries. Under uniform sampling and incoherence assumptions, we prove that a tuning-free square-root matrix completion estimator (square-root MC) achieves optimal statistical performance for solving the noisy matrix completion problem. Similar to the square-root Lasso estimator in high-dimensional linear regression, square-root MC does not rely on the knowledge of the size of the noise. While solving square-root MC is a convex program, our statistical analysis of square-root MC hinges on its intimate connections to a nonconvex rank-constrained estimator.

E1190: **Functional individualized treatment regimes with imaging features**
Presenter: **Xinyi Li**, Clemson University, United States
Co-authors: Michael Kosorok

Precision medicine seeks to discover an optimal personalized treatment plan and thereby provide informed and principled decision support, based on the characteristics of individual patients. With recent advancements in medical imaging, it is crucial to incorporate patient-specific imaging features in the study of individualized treatment regimes. We propose a novel, data-driven method to construct interpretable image features which can be incorporated, along with other features, to guide optimal treatment regimes. The proposed method treats imaging information as a realization of a stochastic process, and employs smoothing techniques in estimation. We show that the proposed estimators are consistent under mild conditions. The proposed method is applied to a dataset provided by the Alzheimer's Disease Neuroimaging Initiative.

E1228: **Estimation and inference for partially linear models with estimated outcomes using high-dimensional data**
Presenter: **Jing Tao**, University of Washington, United States

Methods are given for estimating and conducting inference on partially linear models with estimated outcomes using high-dimensional data. Our new estimator allows for but does not require many more regressors than the number of observations for the parametric part. The first stage allows a general set of machine learning methods to be used to form the estimated outcomes. In the second stage, we derive the convergence rates for the linear parameters and the nonparametric function under a partially linear specification for the outcome equation, respectively. We also provide bias correction procedures to allow for valid pointwise and uniform inference for both the linear parameters and the nonparametric function. We evaluate the finite sample performance with simulation studies. Additionally, a real data analysis of the effect of the Fair Minimum Wage Act on the unemployment rate is performed as an illustration of our methods.

EO664 Room Virtual R03 ADVANCE IN APPLICATION OF FUNCTIONAL DATA ANALYSIS Chair: Haolun Shi

E0474: **Deep neural network with a smooth monotonic output layer for dynamic risk prediction**
Presenter: **Zhiyang Zhou**, University of Manitoba, Canada

The fundamental concern of risk prediction is the relationship between predictors and the survival function. The recent success of survival analysis has already been extended to dynamic risk prediction, where the model considers repeated measurements of time-varying predictors. However, existing approaches usually involve strong model assumptions (e.g., additive effects and/or proportional hazard) or discretize the time domain and approximate the survival function by a step function, which may lead to biased prediction. To tackle these issues, we present a deep neural network with a novel output layer termed the Smooth Monotonic Output Layer (SMOL). The resulting network involves no discretization and specifies no parametric structure for the underlying relationship between predictors and the time to event. Attaching SMOL to a neural network, one may infer/learn the cumulative distribution function for a continuous random variable, directly and nonparametrically. We conduct experiments on datasets from the Lifetime Risk Pooling Project (LRPP). LRPP pools together individual data from twenty community-based studies on cardiovascular disease and involves around three hundred thousand participants with long-term follow-ups of longitudinal risk factors (e.g., blood pressure and cholesterol). Extensive results show that our proposal achieves state-of-the-art accuracy in predicting the individual-level risk of atherosclerotic cardiovascular disease.

E0588: **Exploring pre-launch movie electronic word-of-mouth time series by functional data analysis**
Presenter: **Tianyu Guan**, Brock University, Canada

the aim is to explore the dynamic patterns of the pre-launch online movie reviews, or movie electronic word-of-mouth (eWOM), over time and investigate the impact of pre-launch eWOM on explaining the box office revenues. One focus is to use the pre-launch eWOM evolution data as an early indicator of strong or weak box office, which would be helpful to business decision-makers in the movie industry. The eWOM data evolve in time and are treated as functional data. We observe that most eWOM data exhibit a positivity bias and extremity; therefore, we apply the functional principal component analysis, a dimension reduction technique in functional data analysis, to explore the dynamic patterns of various quantile trajectories of the movie eWOM, instead of directly studying the eWOM trajectories. The functional principal component (FPC) scores of quantile trajectories at various quantile levels are used to explain the box office revenues. We use the sparse group lasso method to select the quantile levels and individual FPC scores that make significant contributions to the prediction of box office revenues. The results show that compared with other measures such as valence and variance, the top-end quantiles would be a better measure in capturing the relations between the pre-launch product ratings time pattern and launch sales.

E2009: **Long-term risk prediction utilizing mammography data**
Presenter: **Shu Jiang**, Washington university, United States

Screening mammography aims to identify breast cancer early and secondarily measures breast density to classify women at higher or lower than average risk for future breast cancer in the general population. Our primary goal is to extract mammogram-based features that augment the well-established breast cancer risk factors to improve prediction accuracy. We will present a novel supervised functional principal component analysis to extract image-based features that are ordered by association with the failure times. A closed-form solution is provided through the proposed eigenvalue decomposition algorithm. Empirical comparisons are made to the conventional functional principal component analysis and

the functional partial least squares method. The proposed method is applied to the motivating Joanne Knight Breast Health cohort at Siteman Cancer Center. Our study demonstrates superior prediction performance compared to the benchmark models and reveals insights into risk patterns within mammograms.

E0198: **Robust regression-based functional principal component analysis**
Presenter: **Haolun Shi**, Simon Fraser University, Canada

It is of great interest to conduct robust functional principal component analysis (FPCA) that can identify the major modes of variation in the stochastic process with the presence of outliers. A new robust FPCA method is proposed in a new regression framework. An M-estimator for the functional principal components is developed based on the Hubers loss by iteratively fitting the residuals from the KarhunenLovve expansion for the stochastic process under the robust regression framework. Our method can naturally accommodate sparse and irregularly-sampled data. When the functional data have outliers, our method is shown to render stable and robust estimates of the functional principal components; when the functional data have no outliers, we show via simulation studies that the performance of our approach is similar to that of the conventional FPCA method. The proposed robust FPCA method is demonstrated by analyzing the Hawaii ocean oxygen data and the kidney glomerular filtration rates for patients after renal transplantation.

EO512 Room Virtual R04 EMERGING STATISTICAL ISSUES IN OVERPARAMETERIZED MODELING Chair: Yoonkyung Lee

E0451: **A universal trade-off between the model size, test loss, and training loss of linear predictors**
Presenter: **Nikhil Ghosh**, University of California, Berkeley, United States
Co-authors: Mikhail Belkin

The aim is to establish an algorithm and distribution-independent non-asymptotic trade-off between the model size, excess test loss, and training loss of linear predictors. Specifically, we show that models that perform well on the test data (have a low excess loss) are either "classical" – have training loss close to the noise level, or are "modern" – have a much larger number of parameters compared to the minimum needed to fit the training data exactly. We also provide a more precise asymptotic analysis when the limiting spectral distribution of the whitened features is Marchenko-Pastur. Remarkably, while the Marchenko-Pastur analysis is far more precise near the interpolation peak, where the number of parameters is just enough to fit the training data, in settings of most practical interest, it differs from the distribution independent bound by only a modest multiplicative constant.

E0414: **Predictive model degrees of freedom in linear regression**
Presenter: **Bo Luan**, Google, United States
Co-authors: Yoonkyung Lee, Yunzhang Zhu

Overparametrized interpolating models have drawn increasing attention from machine learning. Some recent studies suggest that regularized interpolating models can generalize well. This phenomenon seemingly contradicts the conventional wisdom that interpolation tends to overfit the data and performs poorly on test data. Further, it appears to defy the bias-variance trade-off. As one of the shortcomings of the existing theory, the classical notion of model degrees of freedom fails to explain the intrinsic difference among the interpolating models since it focuses on the estimation of in-sample prediction error. This motivates an alternative measure of model complexity which can differentiate those interpolating models and take different test points into account. In particular, we propose a measure with a proper adjustment based on the squared covariance between the predictions and observations. Our analysis with the least squares method reveals some interesting properties of the measure, which can reconcile the double descent phenomenon with the classical theory. This opens doors to an extended definition of model degrees of freedom in modern predictive settings.

E0552: **Benefit of interpolation in nearest neighbor algorithms**
Presenter: **Yue Xing**, Purdue University, United States
Co-authors: Qifan Song, Guang Cheng

In some studies of deep learning, it is observed that over-parametrized deep neural networks achieve a small testing error even when the training error is almost zero. Despite numerous works towards understanding this so-called double descent phenomenon, we turn to another way to enforce zero training error (without over-parametrization) through a data interpolation mechanism. Specifically, we consider a class of interpolated weighting schemes in the nearest neighbors (NN) algorithms. By carefully characterizing the multiplicative constant in the statistical risk, we reveal a U-shaped performance curve for the level of data interpolation in both classification and regression setups. This sharpens the existing result that zero training error does not necessarily jeopardize predictive performances and claims a counter-intuitive result that a mild degree of data interpolation actually strictly improves the prediction performance and statistical stability over those of the (un-interpolated) k-NN algorithm. In the end, the universality of our results, such as the change of distance measure and corrupted testing data, will also be discussed.

E1676: **On the Generalization Power of the Overfitted Three-Layer Neural Tangent Kernel Model**
Presenter: **Peizhong Ju**, The Ohio State University, United States
Co-authors: Xiaojun Lin, Ness Shroff

The generalization performance of overparameterized 3-layer NTK models is studied. We show that, for a specific set of ground-truth functions (which we refer to as the "learnable set"), the test error of the overfitted 3-layer NTK is upper bounded by an expression that decreases with the number of neurons of the two hidden layers. Different from 2-layer NTK, where only one hidden layer exists, the 3-layer NTK involves interactions between two hidden layers. Our upper bound reveals that, between the two hidden layers, the test error descends faster with respect to the number of neurons in the second hidden layer (the one closer to the output) than with respect to that in the first hidden layer (the one closer to the input). We also show that the learnable set of 3-layer NTK without bias is no smaller than that of 2-layer NTK models with various choices of bias in the neurons. However, in terms of the actual generalization performance, our results suggest that 3-layer NTK is much less sensitive to the choices of bias than 2-layer NTK, especially when the input dimension is large.

EO052 Room Virtual R05 NEW APPROACHES IN HIGH-DIMENSIONAL TIME SERIES MODELING Chair: Abolfazl Safikhani

E1838: **Dimension reduction in multivariate time series**
Presenter: **Seyed Yaser Samadi**, Southern Illinois University Carbondale, United States
Co-authors: Wiranthe Herath

Dimensionality reduction is very important in multivariate time series analysis because the number of parameters grows rapidly with the time series dimension. There are many dimensionality reduction techniques for time series; however, the achieved dimensionality reductions by these methods are not substantial and they usually fail to extract relevant information from a complex body of data because they fail to distinguish between information that is important to the scientific goals. We introduce a new parsimonious multivariate time series model that achieves efficient estimation by linking the mean function and covariance matrix and using the minimal reducing subspace. The results of simulation studies and real data analysis that compare the performance of the proposed model with that of the existing models in the literature will be presented.

E1818: **Theoretical analysis of deep neural networks for temporally dependent observations**
Presenter: **Abolfazl Safikhani**, University of Florida, United States

Deep neural networks are powerful tools to model observations over time with non-linear patterns. Despite the widespread use of neural networks

in such settings, most theoretical developments of deep neural networks are under the assumption of independent observations, and theoretical results for temporally dependent observations are scarce. To bridge this gap, we study theoretical properties of deep neural networks on modeling non-linear time series data. Specifically, non-asymptotic bounds for prediction error of (sparse) feed-forward neural network with ReLU activation function are established under mixing-type assumptions. These assumptions are mild such that they include a wide range of time series models, including auto-regressive models. Compared to independent observations, established convergence rates have additional logarithmic factors to compensate for additional complexity due to dependence among data points. The theoretical results are supported via various numerical simulation settings as well as an application to a macroeconomic data set.

E1732: **An interpretable and efficient infinite-order vector autoregressive model for high-dimensional time series**
Presenter: **Yao Zheng**, University of Connecticut, United States

As a special infinite-order vector autoregressive (VAR) model, the vector autoregressive moving average (VARMA) model can capture much richer temporal patterns than the widely used finite-order VAR model. However, its practicality has long been hindered by its non-identifiability, computational intractability, and the relative difficulty of interpretation. A novel infinite-order VAR model is introduced, which, with only a little sacrifice of generality, inherits the essential temporal patterns of the VARMA model but avoids all of the above drawbacks. As another attractive feature, the temporal and cross-sectional dependence structures of this model can be interpreted separately, since they are characterized by different sets of parameters. For high-dimensional time series, this separation motivates us to impose sparsity on the parameters determining the cross-sectional dependence. As a result, greater statistical efficiency and interpretability can be achieved, while no loss of temporal information is incurred by the imposed sparsity. We introduce an ℓ_1-regularized estimator for the proposed model and derive the corresponding nonasymptotic error bounds. An efficient block coordinate descent algorithm and a consistent model order selection method are developed. The merit of the proposed approach is supported by simulation studies and real-world macroeconomic data analysis.

E1778: **Optimal change point testing in high-dimensional regression time series**
Presenter: **Daren Wang**, University of Notre Dame, United States

Detecting changes in regression time series is a fundamental problem arising in a broad spectrum of applications such as dynamic pricing, predictive maintenance, signal processing, and many more. We focus on multiple change-point testing in the high-dimensional linear regression setting. Specifically, we assume that the unobserved high-dimensional regression coefficients can potentially change over time in a piecewise constant manner. We propose a new statistic named the Covariance-based Quadratic CUSUM statistic (CQC) to test the existence of change points. We characterize the null and alternative limiting distributions of CQC. We show that CQC can not only consistently test the existence of change points but also achieve the optimal detection boundary. Furthermore, the proposed methodology enjoys broad applicability as it allows temporal dependence among the regression time series, even when the dimensionality of the regression coefficients grows exponentially with the sample size.

EO542 Room Virtual R06 ROBUSTNESS AND RELATED TOPICS II Chair: Ana Maria Bianco

E0374: **Robust estimation based on B-splines with simultaneous variable selection for partially linear additive models**
Presenter: **Alejandra Martinez**, Universidad de Buenos Aires, Argentina
Co-authors: Nicolas Murrone

To deal with the curse of dimensionality, partially linear additive models provide a flexible and interpretable approach to building predictive models. Under these models, the response variable depends linearly on some covariates, while the others enter the model in a fully nonparametric way, as a sum of univariate functions of each variable, that is, as an additive model. In practice, a large number of covariates may be collected, and the non-significative ones should be excluded from the model. For that reason, it is important to automatically select variables either in the parametric or in the nonparametric components. In order to obtain simultaneously robust estimators and select the active covariates, we introduce a family of robust estimators that combines B-splines and robust regression estimators with a regularization procedure based on a SCAD penalty which penalizes both the coefficients of the linear and additive components. Through a Monte Carlo study, we will show the advantage of the proposed methodology with respect to that based on least squares.

E0564: **Robust inference for non-inferiority studies**
Presenter: **Laura Ventura**, University of Padova, Italy
Co-authors: Elena Bortolato

Nowadays, the goal of many studies is to determine if new therapies have equivalent or non-inferior efficacies to the ones currently in use. These studies are called equivalence and non-inferiority studies, and the statistical methods for their analysis require simple modifications to the traditional hypotheses testing framework. The simplest and most widely used approach to test equivalence or non-inferiority is the two one-sided test (TOST) procedure, which evolves around the use of likelihood methods for testing the comparison of parameters like means, odds ratios, hazard ratios, etc. However, it is well-known that for model misspecifications or in the presence of influential observations, likelihood methods are highly unstable in many applications and to overcome this drawback, the theory of robust unbiased estimating equations may be usefully considered. The aim of this contribution is to discuss the use of robust unbiased estimating equations in the context of non-inferiority studies. In particular, we will resort on robust confidence distributions for the scalar parameter of interest, which allow deriving not only confidence intervals and p-values, but also suitable robust measures of evidence and performing robust sensitivity analysis to the preliminary choice of inferiority and equivalence margins.

E0825: **Robust estimators for functional logistic regression**
Presenter: **Marina Valdora**, Universidad de Buenos Aires, Argentina
Co-authors: Graciela Boente

The logistic regression model is widely used in data analysis. In many applied problems, the data are originated from phenomena that are better modeled by continuous functions than by finite-dimensional vectors. Functional logistic regression is a generalization of the logistic regression model that assumes that the covariates are functions while the responses are 0 or 1. We will present a robust proposal to estimate the slope under a functional logistic regression model which combines a dimension reduction with M-estimators. Through the results of a numerical study, we will illustrate the sensitivity of the classical estimator and the stability of the proposed method.

E1328: **On depths for noisy functional data: The quantile Integrated depth**
Presenter: **Sara Lopez Pintado**, Northeastern University, United States

Functional data analysis is an exciting field of statistics where the basic unit of observation is a function or image. The development of robust exploratory tools and inferential methods for functional data is very much needed. Data depth is a well-known and useful non-parametric tool for analyzing functional data. It provides a way of ranking a sample of curves from the centre outwards and of defining robust statistics. Several notions of depth for functional data were introduced in the literature in the last few decades. These functional depths usually satisfy desirable properties, such as some type of invariance, maximality at the center and monotonicity with respect to the deepest point. We develop a new family of depths denoted by quantile integrated depth (QID) based on integrating up to the K-th percentile/quantile of the univariate depths. The theoretical properties of this new family of depths are studied. In practice, functional data are often observed with noise. We explore the effect of noise on different notions of depth and propose the SAD (Spearman Agreement Depth) plot to compare the performance of these functional depths in the presence of noise. The proposed quantile integrated depths are shown to be robust to noisy functional data outperforming alternative functional depths. Procedures for choosing optimal tuning parameter K in QID based on the SAD plot are discussed.

EO720 Room Virtual R07 BAYESIAN NONPARAMETRICS FOR CAUSAL INFERENCE: PART I Chair: Chanmin Kim

E0261: **Bayesian semiparametric model for sequential AML treatment decisions with informative timing**
Presenter: **Arman Oganisian**, Brown University, United States
Co-authors: Jason Roy
A Bayesian semiparametric model is developed for the impact of dynamic treatment rules on survival among patients diagnosed with pediatric acute myeloid leukemia (AML). The data are from a phase III clinical trial in which patients move through a sequence of four treatment courses where they are treated with either anthracycline-based chemotherapy (ACT) agents or non-anthracycline-based agents only. While ACT is thought to suppress AML aggressively, it is also cardiotoxic, so that treating overzealously with either may reduce survival. Our task is to estimate the potential survival probability under hypothetical treatment rules, but there are several impediments. First, since ACT is not randomized, its effect on survival is confounded over time. Second, subjects initiate the next course at varying times depending on when they recover from the previous course - making timing potentially informative of subsequent ACT decisions and survival. Third, patients may die or drop out before completing the full treatment sequence. To address these issues, we develop a generative Bayesian semi-parametric model based on Gamma Process priors that capture subjects' transition to subsequent treatment or death in continuous time. A g-computation procedure is used to compute posterior potential survival probabilities while adjusting for time-varying confounding.

E0666: **Ordinal causal discovery**
Presenter: **Yang Ni**, Texas AM University, United States
Causal discovery for purely observational, categorical data is a long-standing challenging problem. Unlike continuous data, the vast majority of existing methods for categorical data focus on inferring the Markov equivalence class only, which leaves the direction of some causal relationships undetermined. An identifiable ordinal causal discovery method is proposed that exploits the ordinal information contained in many real-world applications to identify the causal structure uniquely. The proposed method is applicable beyond ordinal data via data discretization. Through real-world and synthetic experiments, we demonstrate that the proposed ordinal causal discovery method combined with simple score-and-search algorithms has favorable and robust performance compared to state-of-the-art alternative methods in both ordinal categorical and non-categorical data.

E0735: **Avoiding highly-informative "nonparametric" priors**
Presenter: **Antonio Linero**, University of Texas at Austin, United States
The problem of specifying nonparametric priors when the goal is to estimate (conditional) average causal effects is considered. Ironically, the flexibility aimed for in specifying a nonparametric prior can sometimes accomplish the opposite of what we want. Rather than flexibly modeling causal effects, we may inadvertently (i) encode information that highly constrains them or (ii) encode unrealistically large amounts of heterogeneity. We illustrate the problem through simple examples using Gaussian processes and ridge regression and present solutions. The proposed corrections take the form of propensity score adjustments, giving a Bayesian nonparametric take on the wisdom of controlling for elements of the design in constructing estimates of causal effects.

E0873: **TEA prior: A Bayesian approach with application for adaptive platform trials having temporal changes**
Presenter: **Chenguang Wang**, Regeneron Pharmaceuticals, United States
Temporal changes exist in clinical trials. Over time, shifts in patient characteristics, trial conduct, and other features of a clinical trial may occur. In typical randomized clinical trials, temporal effects, i.e., the impact of temporal changes on clinical outcomes and study analysis, are largely mitigated by randomization and usually need not be explicitly addressed. However, temporal effects can be a serious obstacle to conducting clinical trials with complex designs, including the adaptive platform trials that are gaining popularity in recent medical product development. We introduce a Bayesian robust prior for mitigating temporal effects based on a hidden Markov model and propose a particle filtering algorithm for computation. We conduct simulation studies to evaluate the performance of the proposed method and provide illustrative examples based on trials of Ebola virus disease therapeutics and hemostat in vascular surgery.

EO670 Room Virtual R08 STATISTICAL METHODS FOR MASSIVE OR HIGH-DIMENSIONAL DATA Chair: Alexander Munteanu

E0701: **Algorithmic Gaussianization through sketching: Converting data into sub-Gaussian random designs**
Presenter: **Michal Derezinski**, University of Michigan, United States
Algorithmic Gaussianization is a phenomenon that can arise when using randomized sketching or sampling methods to produce smaller representations of large datasets: For certain tasks, these sketched representations have been observed to exhibit many robust performance characteristics that are known to occur when a data sample comes from a sub-gaussian random design, which is a powerful statistical model of data distributions. However, this phenomenon has only been studied for specific tasks and metrics, or by relying on computationally expensive methods. We address this by providing an algorithmic framework for Gaussianizing data distributions via averaging, proving that it is possible to efficiently construct data sketches that are nearly indistinguishable (in terms of total variation distance) from sub-gaussian random designs. In particular, relying on a recently introduced sketching technique called Leverage Score Sparsified (LESS) embeddings, we show that one can construct an $n \times d$ sketch of an $N \times d$ matrix A, where $n << N$, that is nearly indistinguishable from a sub-Gaussian design, in time nearly linear in the size of A. As a consequence, strong statistical guarantees and precise asymptotics available for the estimators produced from sub-gaussian designs can be straightforwardly adapted to our sketching framework. We illustrate this with a new approximation guarantee for sketched least squares.

E0974: **High-dimensional sufficient dimension reduction through principal projections**
Presenter: **Eugen Pircalabelu**, Universita catholique de Louvain, Belgium
Co-authors: Andreas Artemiou
A new dimension reduction method is presented for high-dimensional settings. The proposed procedure is based on a principal support vector machine framework where principal projections are used in order to overcome the non-invertibility of the covariance matrix. We show that one can accurately recover the central subspace using a projection on a lower dimensional subspace and then applying an l1 penalization strategy to obtain sparse estimators of the sufficient directions. Based next on a desparsified estimator, we provide an inferential procedure for high-dimensional models that allows testing for the importance of variables in determining the sufficient direction. Theoretical properties of the methodology are illustrated and computational advantages are demonstrated with simulated and real data experiments.

E0792: **Scalable Bayesian p-generalized probit and logistic regression via coresets**
Presenter: **Zeyu Ding**, TU Dortmund, Germany
The logit and probit link functions are arguably the two most common choices for binary regression models. Many studies have extended the choice of link functions to avoid possible misspecification and improve the model fit to the data. We introduce the p-generalized normal distribution into binary regression in a Bayesian framework. The p-generalized normal distribution has received considerable attention due to its flexibility in modeling the tails while generalizing, for instance, over the standard normal distribution where $p = 2$ or the Laplace distribution where $p = 1$. A scalable maximum likelihood estimation (MLE) method for p-generalized probit regression has been developed recently. We extend the estimation from MLE to Bayesian posterior estimates using Markov Chain Monte Carlo (MCMC) sampling for the model parameter β and the link function parameter p. We use simulated and real-world data to verify the effect of different parameters p on the estimation results and how logistic regression and probit regression can be incorporated into a broader framework. To make our Bayesian methods scalable in the case of large data,

we also incorporate coresets as a means of reducing the data before performing the complex and time-consuming MCMC analysis. This allows us to perform very efficient calculations while retaining the original posterior parameter distributions up to little distortions in practice and with theoretical guarantees.

E1041: **Bounding the width of neural networks via coupled initialization**
Presenter: **Simon Omlor**, TU Dortmund, Germany

Two-layer ReLU neural networks with cross-entropy or squared loss can be seen as logistic resp. l_2 regression in the infinite-dimensional Neural Tangent Kernel (NTK) space when the number of neurons in the hidden layer is infinite. A common method in training such networks is to initialize all weights to be independent Gaussian vectors. We observe that by instead initializing the weights into independent pairs, where each pair consists of two identical Gaussian vectors, we can significantly improve the analysis for convergence to zero training error. Specifically, our technique allows reducing the number of hidden neurons required by the network, which corresponds to a dimensionality reduction for the NTK. In the under-parameterized setting with logistic loss, we improve previous width bounds from roughly γ^{-8} to γ^{-2}, where γ denotes the separation margin in the NTK space. We also present new lower bounds that corroborate the tightness of our analysis. Similar techniques also improve previous width bounds in the over-parameterized setting with squared loss from roughly n^4 to n^2.

EO116 Room BH (S) 2.05 BAYESIAN SEMI- AND NON-PARAMETRIC METHODS I Chair: Andrea Cremaschi

E0653: **Repulsion, chaos and equilibrium in mixture models**
Presenter: **Maria De Iorio**, National University of Singapore, Singapore

Mixture models are commonly used in applications presenting heterogeneity and overdispersion in the population. In the Bayesian framework, this entails the specification of suitable prior distributions for the weights and location parameters of the mixture. Indeed, the flexibility of these models and prior distributions often does not translate into interpretability of the identified clusters. To overcome this issue, clustering methods based on repulsive mixtures have been recently proposed. The basic idea is to include a repulsive term in the prior distribution of the atoms of the mixture, which favours mixture locations far apart. This approach leads to well-separated clusters, thus facilitating the interpretation of the results. However, the resulting models are usually not easy to handle due to the introduction of intractable normalising constants. Exploiting results from statistical mechanics, we propose a novel class of repulsive prior distributions based on Gibbs's measures. Specifically, we use Gibbs measures associated with joint distributions of eigenvalues of random matrices, which naturally possess a repulsive property. The proposed framework greatly simplifies computations, due to the availability of the normalising constant in closed form. We establish theoretical results that imply that the locations become independent as the number of components tends to infinity and illustrate the novel class of priors on benchmark datasets.

E1812: **Normalized latent measure factor models**
Presenter: **Jim Griffin**, University College London, United Kingdom
Co-authors: Mario Beraha

A methodology is proposed for modelling and comparing probability distributions within a Bayesian nonparametric framework. Building on dependent normalized random measures, we consider a prior distribution for a collection of discrete random measures where each measure is a linear combination of a set of latent measures, interpretable as characteristic traits shared by different distributions, with positive random weights. The model is non-identified, and a method for post-processing posterior samples to achieve identified inference is developed. This uses Riemannian optimization to solve a non-trivial optimization problem over a Lie group of matrices. The effectiveness of our approach will be illustrated in two applications to two real-world data sets: school student test scores and personal incomes in California. Our approach leads to interesting insights for populations and easily interpretable posterior inference

E0472: **A new BART prior for structured categorical inputs**
Presenter: **Sameer Deshpande**, University of Wisconsin–Madison, United States

Default implementations of Bayesian Additive Regression Trees (BART) represent categorical predictors using several binary indicators, one for each level of each categorical predictor. Regression trees with these indicators send one level of a categorical predictor to the left and all other levels to the right. Unfortunately, most partitions of the levels cannot be built with this "remove one at a time" strategy, meaning that default implementations of BART are limited in their ability to "borrow strength" across groups of levels. We overcome this limitation with a prior for a new class of regression trees that can send multiple levels of a categorical variable to each child of a decision node in a regression tree. Our prior corresponds to a partitioning process that respects a priori preferences to co-cluster certain levels of a structured categorical variable. In spatiotemporal applications, such variables are frequently used to encode membership in spatial units like census tracts or counties. In these applications, our new regression trees induce spatially-contiguous partitions of the spatial units. Our new prior often yields improved out-of-sample predictive performance without much additional computational burden. Despite its conceptual simplicity, our new prior opens the door for Bayesian treed regression over complex discrete spaces like networks.

E1581: **Exact Bayesian inference for a class of spatial generalized linear mixed effects models**
Presenter: **Jonathan Bradley**, Florida State University, United States

Markov chain Monte Carlo (MCMC) has become a standard in Bayesian statistics that allows one to generate dependent replicates from a posterior distribution for general Bayesian hierarchical models. However, convergence issues, tuning, and the effective sample size of the MCMC are nontrivial considerations that are often overlooked or can be difficult to assess. This motivates us to consider finding expressions of the posterior distribution that are computationally straightforward to sample from directly. We focus on a broad class of Bayesian generalized linear mixed-effects models (GLMM) that allows one to jointly model data of different types. We derive a class of distributions that allows one to specify the prior on fixed and random effects to be any conjugate multivariate distribution. The expression of the posterior distribution is given, and direct simulations have an efficient projection form. An analysis of a spatial dataset is presented.

EO386 Room K2.31 (Nash Lec. Theatre) WEIGHTING METHODS FOR CAUSAL INFERENCE AND SELECTION BIAS Chair: Shaun Seaman

E0828: **Sensitivity analysis for calibrated inverse probability of censoring weighted estimators under non-ignorable dropout**
Presenter: **Shaun Seaman**, University of Cambridge, United Kingdom
Co-authors: Li Su, Sean Yiu

Inverse probability of censoring weighting (IPCW) is a popular approach for dealing with dropout in longitudinal studies. The weights can be estimated by specifying a model for the probability of dropout and estimating its parameters using maximum likelihood. More recently, calibrated IPCW estimators have been proposed. These use weights that directly optimize covariate balance in the weighted sample, an approach which has been shown to reduce the mean-squared error of the IPCW estimator. Existing calibrated IPCW estimators are based on the unverifiable assumption of sequential ignorability, and sensitivity analysis strategies to violation of this assumption have been lacking. We shall describe an approach to sensitivity analysis for calibrated IPCW estimators under non-ignorable dropout. We shall illustrate its use on data from an international cohort study of systemic lupus erythematosus.

E0429: **Dynamic covariate balancing: Estimating treatment effects over time**
Presenter: **Davide Viviano**, Stanford University, United States

The problem of estimation and inference on the effects of time-varying treatment is discussed. We propose a method for inference on the treatment

effects histories, introducing a dynamic covariate balancing method combined with penalized regression. Our approach allows for (i) treatments to be assigned based on arbitrary past information, with the propensity score being unknown; (ii) outcomes and time-varying covariates to depend on treatment trajectories; (iii) high-dimensional covariates; (iv) heterogeneity of treatment effects. We study the asymptotic properties of the estimator, and we derive the parametric convergence rate of the proposed procedure. Simulations and an empirical application illustrate the advantage of the method over state-of-the-art competitors.

E1173: **Optimal weighting for estimating treatment effects**
Presenter: **Michele Santacatterina**, New York University, United States
Co-authors: Nathan Kallus

Weighted methods based on Inverse Probability Weights (IPW) have been widely used to estimate causal effects using observational data. Despite their wide use, IPW methods rely on the correct specification of the propensity score model, in which violations lead to biased estimates, and on the positivity assumption, in which practical violations lead to extreme weights and erroneous inferences. We will present novel approaches based on modern optimization techniques and machine learning methods that mitigate model misspecification while simultaneously controlling for precision. We will describe two methods that find weights that balance confounders to estimate the effect of binary and continuous treatments on continuous and time-to-event outcomes. We will also describe a method that finds weights that optimally balance time-dependent confounders for marginal structural models. We will present these approaches using HIV, spine surgery, and heart disease data.

E1819: **Approximate balancing weighting for treatment effects: Justifications, choices, and fundamental limitations**
Presenter: **Chad Hazlett**, UCLA, United States

Weighting approaches used in applied causal inference settings may rest on the estimation of the propensity score, on directly seeking to balance covariates, or on a combination of these goals. Accordingly, these approaches are provably unbiased under claims regarding the specification of the propensity score, the specification of the (conditional expectation of) outcome, or under a weaker combined assumption. We review these varying motivations for weighting approaches, emphasizing the (less widely-recognized) justification that requires only assumptions on the outcome model. An analysis of the bias in the estimated treatment effect illustrates the specification dependencies inherent in weighting approaches, aiding investigators in answering the practical question of what functions of the covariates should be balanced. This offers one motivation for a family of kernel-based weighting estimators proposed by authors on this panel. Finally, we turn to a central tradeoff at the heart of these methods: weighting approaches allow us to make weak specification assumptions, but attempt to obtain balance on features that may be irrelevant. Indeed, weights that produce worse balance, or even offsetting imbalances, can have better finite-sample performance. We discuss how this limits the practical performance of weighting-only approaches, and alternative or hybrid options.

EO182 Room K2.40 RECENT DEVELOPMENTS IN FUNCTIONAL DATA ANALYSIS Chair: Eliana Christou

E0367: **Deep neural network classifier for multi-dimensional functional data**
Presenter: **Guanqun Cao**, Auburn University, United States
Co-authors: Zuofeng Shang

A new approach is proposed, called functional deep neural network (FDNN), for classifying multi-dimensional functional data. Specifically, a deep neural network is trained based on the principle components of the training data, which shall be used to predict the class label of a future data function. Unlike the popular functional discriminant analysis approaches, which rely on Gaussian assumption, the proposed FDNN approach applies to general non-Gaussian multi-dimensional functional data. Moreover, when the log density ratio possesses a locally connected functional modular structure, we show that FDNN achieves minimax optimality. The superiority of our approach is demonstrated through both simulated and real-world datasets.

E0493: **Robust functional quantile regression**
Presenter: **Ufuk Beyaztas**, Marmara University, Turkey
Co-authors: Mujgan Tez, Han Lin Shang

Scalar-on-function quantile regression is a powerful regression model to characterize the entire conditional distribution of a scalar response variable for a given functional predictor. Compared with the conditional mean regression-based scalar-on-function regression model, the scalar-on-function quantile regression is robust to outliers in the response variable. However, it is susceptible to outliers in the functional predictor (called leverage points). The leverage points may alter the eigenstructure of the predictor matrix, leading to poor estimation and prediction results. A robust procedure is proposed to estimate the model parameters in the scalar-on-function quantile regression method and produce reliable predictions in the presence of both outliers and leverage points. The proposed method is based on a functional partial quantile regression procedure. The estimation and prediction performance of the proposed method is evaluated by a series of Monte-Carlo experiments and an empirical data example, diffusion tensor imaging data. The results are compared favorably with several existing methods. The method is implemented in an R package robfpqr.

E1075: **An agglomerative hierarchical local clustering algorithm for functional motif discovery**
Presenter: **Jacopo Di Iorio**, Penn State University, United States
Co-authors: Marzia Cremona, Francesca Chiaromonte

Two of the new issues that functional data analysis is recently dealing with are the identification of local clusters, i.e., clusters defined only on a portion of the domain, and the discovery of functional motifs, i.e. typical "shapes" that may be repeated - scaled or not along the y axis - multiple times within each curve, or across several curves belonging to the same set. We propose a new algorithm to solve these problems, leveraging ideas from multivariate and functional data analysis - especially curve alignment, functional clustering and biclustering. funBIalign is a multi-step algorithm based on agglomerative hierarchical clustering with complete linkage, which is able to discover local clusters and/or functional motifs both in a set of misaligned curves or in a single curve (e.g., time series). Differently from other alternatives, funBIalign is able to detect both shifting and scaling functional patterns thanks to the use of metrics based on functional and adjusted versions of widely used multivariate biclustering validation measures such as the mean squared residue score (or H-score) and the virtual error. Simulations and case studies results are shown to assess the goodness of the methodology proposed.

E1096: **Testing conditional mean independence for functional data**
Presenter: **Chung Eun Lee**, Baruch College, United States
Co-authors: Xianyang Zhang, Xiaofeng Shao

A new nonparametric conditional mean independence test for a response variable Y and a predictor variable X is proposed where either or both can be function-valued. Our test is built on a new metric, the so-called functional martingale difference divergence, which fully characterizes the conditional mean dependence of Y given X and extends the martingale difference divergence. We define an unbiased estimator of functional martingale difference divergence by using a U-centering approach, and obtain its limiting null distribution under mild assumptions. Since the limiting null distribution is not pivotal, we adopt the wild bootstrap method to estimate the critical value and show the consistency of the bootstrap test. The test does not require a finite-dimensional projection nor assume a linear model, and it does not involve any tuning parameters. Promising finite sample performance is demonstrated via simulations and a real data illustration in comparison with the existing tests.

EO563 Room K2.41 STATISTICAL METHODS FOR NEUROIMAGING DATA Chair: Elizabeth Sweeney

E0615: **Quantitative susceptibility maps in multiple sclerosis lesions**
Presenter: **Elizabeth Sweeney**, University of Pennsylvania, United States

Multiple sclerosis (MS) is an inflammatory disease of the central nervous system characterized by lesions in the brain and spinal cord. Magnetic resonance images (MRI) are sensitive to these lesions. A particular type of lesion, called a chronic active lesion, is characterized by a hyperintense rim of iron-enriched, activated microglia and macrophages, and has been linked to greater tissue damage. An MRI technique called quantitative susceptibility mapping (QSM) provides efficient in vivo quantification of susceptibility changes related to iron deposition. It identifies these chronic active lesions, called QSM rim-positive (rim+) lesions. QSM rim+ MS lesions and their longitudinal behavior have the potential to serve as a biomarker of chronic inflammation and to be utilized to monitor disease progression and evaluate disease-modifying therapies in MS. We will discuss the challenges of estimating treatment effects using the longitudinal behavior of QSM rim+ lesions. We will compare two disease-modifying treatments, Tecfidera and Copaxone, using linear mixed-effects regression models with inverse probability of censor weighting.

E0440: **Challenges and opportunities of using interleaved TMS/fMRI data to personalize treatments for depression**
Presenter: **Kristin Linn**, University of Pennsylvania, United States

Transcranial magnetic stimulation (TMS) is a form of non-invasive brain stimulation that can be safely delivered in trains of pulses (repetitive TMS; rTMS) to treat symptoms of depression in patients for whom standard treatments have failed. Recent advances in technology enable rTMS to be interleaved with functional magnetic resonance imaging (fMRI), producing real-time measures of blood oxygen level-dependent (BOLD) contrast in response to single pulses of TMS. We describe challenges and opportunities that arise when working with interleaved TMS/fMRI data to understand better and treat depression. We highlight the importance of study design for answering pertinent scientific questions. Based on findings from a small rTMS study that piloted a new individualized treatment protocol, we provide recommendations for the design and analysis of future rTMS studies.

E0739: **NeuroGen: A tool for activity modulation in human visual networks**
Presenter: **Amy Kuceyeski**, Weill Cornell Medicine, United States

Recording human visual system activity in response to image viewing has classically been a way to understand how the human brain processes incoming information. We present NeuroGen, a machine learning framework that couples an encoding model of human vision and a deep generative network to synthesize images predicted to achieve a target pattern of macro-scale brain activation. We demonstrate that the reduction of noise that the encoding model provides, coupled with the generative network's ability to produce images of high fidelity, results in a robust discovery architecture for visual neuroscience. We begin by validating NeuroGen with known image-brain response relationships, i.e. face/body/word/place areas in the visual system. Then we demonstrate NeuroGen's use in a discovery context by showing that only a few synthetic images can capture novel individual- and region-level variations in the level of activity response to dog compared to human faces. We further demonstrate that NeuroGen can create synthetic images predicted to achieve regional response patterns not achievable by the best-matching natural images. Finally, we present some prospective functional MRI results where we recorded brain activity in individuals viewing NeuroGen's synthetic images. We propose that NeuroGen can enable vision neuroscience discoveries and may allow modulation of the activity of regions within the human visual system (and beyond!) in a controlled way.

E0759: **Harmonizing correlated data structures in multi-center neuroimaging studies**
Presenter: **Russell Shinohara**, University of Pennsylvania, United States

While magnetic resonance imaging (MRI) studies are critical for the diagnosis, monitoring, and study of a wide variety of diseases, their use in quantitative analysis can be complex. An increasingly recognized issue involves the differences between MRI scanners that are used in large multi-center studies. To address this, the current state of the art is to "regress out" or "adjust for" scanner differences. Our group has found these methods to be insufficient, and has advocated for the adaptation of methods pioneered in genomics to help mitigate inter-scanner differences, which can vary across the brain and result in both mean and variance shifts. We further study the implications of differences in correlation structures across and between images, and how this affects downstream inference.

CO364 Room BH (S) 1.01 Lecture Theatre 1 RECENT ADVANCES IN PROBABILISTIC FORECASTING Chair: Gael Martin

C0425: **Evaluation of copula forecasts**
Presenter: **Tobias Fissler**, Vienna University of Economics and Business, Austria

Forecasts for uncertain events Y provide guidance in decision-making. The past decades have witnessed a paradigm shift from point forecasts to distributional forecasts, capturing the inherent uncertainty of Y. The accuracy of distributional forecasts is evaluated in terms of scoring rules, which assign to each predictive distribution F and observation from Y a penalty $S(F,Y)$. To incentivise truthful forecasts, scoring rules should be (strictly) proper, meaning that the expected score is (strictly) minimised by the correctly specified distribution of Y. If Y is multivariate, the predictive distribution can be decomposed into the marginals and the copula, capturing the dependence structure. It has been an open problem if there exist strictly proper copula scores S_{Copula} in the sense that the arguments are a predictive copula C and an observation Y, and the expectation over Y is strictly minimised by providing the correct copula of Y. It will be shown that such strictly proper copula scores cannot exist. As a remedy, the usage of bivariate scores equipped with the lexicographic order is suggested and discussed. They decouple the influence of predictive marginal distributions and the predictive copula. As such, they are a tool to control if a predictive distribution outperforms another one based on its marginals or its copula.

C0447: **Evaluating probabilistic classifiers: The triptych**
Presenter: **Timo Dimitriadis**, Heidelberg University, Germany
Co-authors: Tilmann Gneiting, Alexander Jordan, Peter Vogel

Predicting the occurrence probability of binary events is presumably the most common forecasting task throughout the sciences. Hence, a unified methodology for evaluating and comparing these forecasts is of great importance. We propose a new "triptych" of evaluating displays consisting of the receiver operating characteristic (ROC) curve, the CORP reliability diagram, and the Murphy diagram. Individually, these three displays focus on different and complementary aspects of the forecast's performance. The ROC curve assesses discrimination, the reliability diagram evaluates calibration, and the Murphy diagram combines both properties and visualizes overall predictive ability. In combination, these displays visualize the full generality of a forecast's predictive ability. This intuition is supported by showing the first theoretical result connecting these plots in full generality: For autocalibrated forecasts, the ROC curve and the Murphy diagram display congruent information. We illustrate our proposal through four case studies ranging from astrophysics, meteorology, economics, and social science.

C0921: **Estimating federal funds rates: A regime-mixture approach**
Presenter: **Herman van Dijk**, Erasmus University Rotterdam, Netherlands

Research on the Taylor rule using post-WWII data tends to exogenously partition the data based on either pre-Volcker and Volcker, Greenspan periods. A finer partition is based on the terms of the chair of the Board of Governors of the Federal Reserve System. We contribute to modelling Taylor Rule fundamentals by proposing a novel dynamic mixture model that endogenises structural changes in Taylor Rule fundamentals in real time. Bayes's estimates of regime indicators suggest that the Fed's response to variations in inflation and output gap switches in different phases of the business cycle. In particular, Fed behaviour is usually consistent with the Taylor principles during normal times, but then switching to output

targeting during early phases of recessions. Our results also suggest that splitting post-WWII macroeconomic data into sub-samples based on the Fed chair may generate misleading results.

C0506: **Solving the forecast combination puzzle**
Presenter: **Gael Martin**, Monash University, Australia

The purpose is to demonstrate that the so-called forecasting combination puzzle is a consequence of the methodologies commonly used to produce forecast combinations. By the combination puzzle, we refer to the empirical finding that predictions formed by combining multiple forecasts in ways that seek to optimize forecast performance often do not outperform more naive, e.g. equally-weighted approaches. In particular, we demonstrate that, due to the manner in which such forecasts are typically produced, tests that aim to discriminate between the predictive accuracy of such competing combinations can have low power, and can lack size control, leading to an outcome that favors the simpler approach. In short, we show that this counter-intuitive result can be completely avoided by the adoption of more efficient estimation strategies in the production of the combinations. We illustrate these findings both in the context of forecasting a functional of interest and in terms of predictive densities.

CO618 Room BH (SE) 1.01 REGIME SWITCHING MODELS Chair: Malvina Marchese

C0240: **Signal processing approach to forecasting seasonal load in electricity markets**
Presenter: **Ritvana Rrukaj**, NTNU: Norwegian University of Science and Technology, Norway

Medium-term load forecasting is a critical tool utilized by power market system operators for system planning and load-serving entities for procuring power supply contracts. Techniques from the field of signal processing are employed to model seasonal load in the New York wholesale electricity market (NYISO). We begin by using mathematical filtration techniques to smooth raw NYISO load and price data that spans 2006 to 2018. The resulting filtered data series express load and prices as percentage deviations from their annual moving averages. Next, we develop a nonlinear load model in price and time by using the 90-day moving averages of the filtered price and load data and time-dependent periodic terms to capture the cyclicality of the load. In the final step, we model the noisy residuals using an autoregressive (AR) process. We conclude by performing out-of-sample forecasts of our model using test data from 2019 and 2020. The forecasting results reveal that our model can predict seasonal load with a high degree of accuracy and potentially assist market participants with their medium-term planning objectives.

C1558: **Structural change in asset correlations and macroeconomic fundamentals**
Presenter: **Malvina Marchese**, City University of London, United Kingdom
Co-authors: Ioannis Kyriakou, michael tamvakis, Francesca Di Iorio

The relation between parameter instability in asset returns correlations and macro-economic fundamentals is revisited using a new correlation component model, the Regime Switching DCC-MIDAS (RSDCC-MIDAS), that distinguishes regime switches in the short and long-run correlations. Breaks in the secular component are associated with low-frequency macro-economic fundamentals via a Smooth Transition MIDAS regression, while the short-run correlations are characterized by abrupt breaks due to market constraints. After a discussion of estimation and inference, and simulation-based evaluations, the model is applied to the prediction of energy future returns. Results from an extensive forecasting exercise reveal the benefits of the specification in terms of forecasting performance at medium and long horizons. In addition, the inclusion of breaks in the short-run component increases the model out-of sample performance over periods of pronounced market instability, such as the Covid-19 crisis period.

C1857: **Foreign exchange hedging using a regime-switching model**
Presenter: **Taehyun Lee**, City University of London, United Kingdom
Co-authors: Ioannis Moutzouris, Nikolaos Papapostolou, Mahmoud Fatouh

The research investigates a four-state regime-switching model for optimal foreign exchange hedging using forward contracts with one-, three- and six-month terms for the United States dollar, euro, Japanese yen, Turkish lira and Indian rupee against the pound sterling. The hedging result of the proposed regime-switching model is illustrated, but also compares the figures are compared with the results of other hedging approaches, including two static hedging strategies, nave and ordinary least squares, and two other dynamic methods, the generalised orthogonal generalised autoregressive conditional heteroskedasticity and the Markov regime-switching model. As a result, the proposed model shows the highest level of risk reduction for the United States dollar, euro, Japanese yen and Turkish lira and the second-best performance for the Indian rupee.

C1883: **Forecasting benchmarks of long-term stock returns via machine learning**
Presenter: **Parastoo Mousavi**, Cass business school, United Kingdom
Co-authors: Jens Perch Nielsen, Ioannis Kyriakou, Michael Scholz

Recent advances in pension product development seem to favour alternatives to the risk-free asset often used in financial theory as a performance standard for measuring the value generated by an investment or a reference point for determining the value of a financial instrument. To this end, we apply the simplest machine learning technique, namely, a fully nonparametric smoother with the covariates and the smoothing parameter chosen by cross-validation to forecast stock returns in excess of different benchmarks, including the short-term interest rate, long-term interest rate, earnings-by-price ratio, and the inflation. We show that net of inflation, the combined earnings-by-price ratio and long-short rate spread form one of our best-performing two-dimensional sets of predictors to forecast both one and five-year horizon stock returns. This is a crucial conclusion for actuarial applications aiming to provide pensioners with real-income forecasts.

CO142 Room BH (SE) 1.02 BUSINESS CYCLES AND MACROECONOMIC POLICY Chair: Laura Jackson Young

C0858: **Constructing high-frequency measures of income and consumption inequality**
Presenter: **Laura Jackson Young**, Bentley University, United States
Co-authors: Michael Owyang, Ashley Stewart, Hoang Le

Measures of income and consumption inequality are many and varied. While data from the Current Population Survey are of higher quality (at least at the lower and middle of the income distribution), they are annual, making them ill-suited for empirical analysis, such as in VARs. The Survey of Consumer Expenditure data are higher frequency but considered lower quality. Recently, a high-frequency income inequality dataset using CPS and IRS data, temporally disaggregated on labor income, has been produced. As an alternative, we produce income and consumption inequality based on CPS data, temporally disaggregated on CEX data using a Bayesian version of Chow-Lin. We compare our series to BSZ in a monetary policy application, where we examine the effects of monetary contractions on income inequality.

C0866: **Asymmetric effects of monetary policy on firms**
Presenter: **Ezgi Kurt**, Bentley University, United States

Firm-level evidence on the asymmetric effects of monetary policy in the US is documented. Focusing on firm-level data from 1980q3 to 2016q2, we find that monetary contractions triple the effects of monetary expansions on firms' employment, investment rate, and sales. Furthermore, we examine the role of alternative financial characteristics in propagating these asymmetric effects. My findings show a higher level of asymmetry for firms with a small size, low leverage, high liquidity, or a no dividend-paying status. These results provide evidence that financial characteristics play a role in propagating the asymmetric effects of monetary policy.

C0982: **Macro forecasting with adaptive learning**
Presenter: **Irina Panovska**, University of Texas at Dallas, United States

Co-authors: Prajyna Barua Soni, Azharul Islam, Srikanth Ramamurthy

The aim is to study how incorporating adaptive learning-based inflation expectations can improve the forecasting performance of Unobserved Components (UC) models when it comes to predicting output, inflation, and unemployment. Our model directly integrates the expectations dynamics of the Hybrid New Keynesian Philips Curve while also retaining the appealing statistical features of the UC framework, allowing us to extract information about the output gap. Three interesting sets of results stand out. First, while the perceived persistence of inflation fell during the early stages of the pandemic, it increased sharply and substantially during the period 2021Q2-2022Q2. Second, the estimated output gap started decreasing in mid-2019, decreased sharply during the early stages of the pandemic, and bounced back rapidly. Finally, and most importantly, including information about the output gap and about the inflation expectations process helps improve both inflation forecasts and output growth forecasts relative to benchmark reduced-form models, with the largest improvements in predictive power during recessions and recovery stages.

C1629: **Age and gender differentials in unemployment and hysteresis**

Presenter: **Amy Guisinger**, Lafayette College, United States

Co-authors: Laura Jackson Young, Michael Owyang

A time-varying panel unobserved components model is used to estimate unemployment gaps disaggregated by age and gender. Recessions before COVID affected men's labor market outcomes more than women's; however, the reverse was true for the COVID recession, with effects amplified for younger workers. The aggregate Phillips curve flattens over time and hysteresis is countercyclical for all groups. We find heterogeneity in both the Phillips curve and hysteresis coefficients, with wages responding more to workers with an outside option (high school- and retirement-age) and larger effects of hysteresis for younger workers.

CO470 Room BH (SE) 1.06 PERSISTENT TIME SERIES Chair: Vivien Less

C0229: **Fractional unobserved components models**

Presenter: **Tobias Hartl**, University of Regensburg, Germany

The decomposition of time series into trend and cycle is addressed for the general state space model $y_t = x_t + c_t$, where c_t represents a stationary cyclical component, the d-th difference of the trend x_t is assumed to be a stationary martingale difference sequence, and both x_t and c_t are unobserved. The model allows for d in the set of positive real numbers, thus generalizes unobserved components models to fractionally integrated trends, and does not require any prior knowledge about d. A closed-form solution for the estimation of trend and cycle is provided and is identical to the Kalman filter and smoother but computationally superior. In addition, a conditional sum-of-squares estimator allowing for the joint estimation of d together with all other model parameters is introduced and is shown to be consistent. Monte Carlo studies reveal good estimation properties of the proposed estimators for parameters and unobserved components in finite samples, both in comparison to nonparametric estimators and integer-integrated unobserved components models. The practical benefits of the new methods are demonstrated in several applications, among others, to global temperature curves.

C0234: **Efficient realized variance estimation in time-changeddiffusion processes**

Presenter: **Roxana Halbleib**, University of Freiburg, Germany

Co-authors: Timo Dimitriadis, Jeannine Polivka, Sina Streicher

The aim is to analyze the statistical properties of realized variance estimators under the assumption that financial logarithmic prices follow a time-changed diffusion process. The time change takes the form of a counting process, implying that the logarithmic price is a pure jump process with stochastic and time-varying tick volatility. This framework is more appropriate to capture the dynamics of observed logarithmic price processes than the standard fusion model. It is also more general than the compound Poisson process with constant tick volatility. We show that our approach is particularly suited to model the logarithmic transaction prices of stocks, as they exhibit time-varying tick volatility. Our analysis deals with three types of sampling schemes, namely clock-time sampling, business-time sampling and transaction-time sampling. We theoretically show that, under no market microstructure noise, realized variance is an unbiased estimator of integrated variance and that business time sampling is optimal in terms of mean squared error. To deal with market microstructure noise, we theoretically and empirically consider various bias-corrected realized variance estimators. Our simulation results show that transaction time sampling outperforms business time sampling for high sampling frequencies and large levels of market microstructure noise.

C0372: **Modelling daily power generation time series: Germany's transition towards renewable energies**

Presenter: **Teresa Flock**, Leibniz University Hannover, Germany

Co-authors: Philipp Sibbertsen

As renewable energies provide a major instrument for slowing climate change, we expect the German energy mix to transition towards lower-emission power sources. To explore this expectation, we consider the recent daily power generation time series of Germany's twelve main energy sources up to 2021. In a univariate setting, we identify their seasonal patterns and long-term trends, and consider external shocks like the COVID-19 pandemic by accounting for structural breaks. We further perform multivariate analyses to model the interdependencies between the different energy sources, identifying common drivers and substitution effects. We place special emphasis on the German Renewable Energy Sources Act (EEG), comparing its effects on Germany's electricity mix with its stated intentions.

C1274: **Constrained QML estimation for multivariate asymmetric MEM with spillovers: The practicality of matrix inequalities**

Presenter: **Menelaos Karanasos**, Brunel University, United Kingdom

The aim is to review and generalize results on the derivation of tractable non-negativity (necessary and sufficient) conditions for N-dimensional asymmetric MEM and GARCH/HEAVY models with spillovers. We show that the non-negativity constraints are translated into simple matrix inequalities, which are easily handled. In practice, these conditions may not be fulfilled. To deal with these cases, we propose a constrained QML estimation. We also obtain new theoretical results about the second-moment structure and the optimal forecasts of such multivariate processes. Four empirical examples are included to show the effectiveness of the proposed method.

CO206 Room BH (S) 2.02 DYNAMIC ANALYSIS OF CRYPTOCURRENCY Chair: Joann Jasiak

C0433: **Forecasting cryptocurrency prices with machine learning: How important is market volatility**

Presenter: **Perry Sadorsky**, York University, Canada

Co-authors: Irene Henriques

Cryptocurrencies are decentralized digital currencies that use blockchain technology to settle transactions. Cryptocurrencies are an emerging asset class that has the potential to increase efficiency in the financial services sector greatly. Forecasting cryptocurrency prices is critical for making well-informed investment decisions concerning this important new asset class. We use machine learning techniques to forecast daily Bitcoin, Ethereum, Cardano, and Ripple price direction. The analysis reveals that random forests, extremely randomized trees, and support vector machines have higher prediction accuracy than Lasso or Nave Bayes. We find that the 10- to 20-day forecasts using random forests, extremely randomized trees, and support vector machines achieve prediction accuracies greater than 85% with some prediction accuracy reaching 90%. For a 20-day forecast, Shapley values show that MA50, MA200, and WAD technical indicators are important features for each of the four cryptocurrencies studied. US three-month T bills, ten-year bond yields and inflation expectations tend to be more important features than market volatility. The importance of market volatility varies by cryptocurrency. Ripple is unique in that emerging market stock market volatility is the most important

feature. Our results reveal the high prediction accuracy of using machine learning methods in forecasting cryptocurrency price direction and provide valuable information on variable importance.

C0500: **Nonlinear forecasts and impulse responses for causal-noncausal (S)VAR models**
Presenter: **Christian Gourieroux**, University of Toronto and CREST, Canada

The closed-form formulas of nonlinear forecasts and nonlinear impulse response functions (IRF) are introduced for the mixed causal-noncausal (Structural) Vector Autoregressive (S)VAR models. We also discuss the identification of nonlinear causal innovations of the model to which the shocks are applied. Our approach is illustrated by a simulation study and an application to a bivariate process of Bitcoin/USD and Ethereum/USD exchange rates.

C0695: **Estimation of multivariate mixed causal and noncausal models: A review**
Presenter: **Francesco Giancaterini**, Maastricht University, Netherlands
Co-authors: Alain Hecq, Gianluca Cubadda

Several strategies to estimate multivariate mixed causal and noncausal models have been proposed in recent years. The performance of the two most common estimators of these models are investigated, both when population parameters are known and unknown. The first estimator aims to maximize the approximate log-likelihood function, requiring a parametric specification of the error distribution. The second is a semi-parametric estimator that minimizes a specific objective function. The two existing estimation methods are compared using a bivariate process of Bitcoin-USD and Ethereum-USD exchange rates.

C1217: **Double autoregressive time-varying coefficient model for stablecoin prices: An application to Tether**
Presenter: **Emre Inan**, York University, Canada
Co-authors: Antoine Djogbenou, Joann Jasiak

The dynamics of the largest seven stablecoins are examined in terms of market capitalization as of July 2021, including Tether, USD Coin, Binance USD, Dai, Terra USD, True USD, and Pax Dollar. We show that the distributional and dynamic properties of Tether and other stablecoins have been evolving from 2017 to 2021. We use local analysis methods to detect and describe local explosive patterns, such as short-lived trends and bubbles, time-varying volatility, and persistence. We introduce a time-varying parameter Double Autoregressive DAR(1) model accommodating the local explosive patterns. We estimate the model non-parametrically and test hypotheses on the functional parameters. The application to Tether, the stablecoin with the largest market capitalization, provides a good fit and reliable inference while being robust to persistent price dynamics and time-varying volatility.

CO639 Room BH (S) 2.03 HIGH DIMENSIONAL METHODS IN TRACKING INFLATION AND BUSINESS CYCLES Chair: Andrew Butters

C0391: **Monetary policy, inflation outlook, and recession probabilities**
Presenter: **Luca Benzoni**, Federal Reserve Bank of Chicago, United States
Co-authors: Andrea Ajello, Makena Schwinn, Yannick Timmer, Francisco Vazquez-Grande

Why does the short-term slope of the yield curve predict recessions? We explore the economic forces underlying Treasury yields' fluctuations and highlight the roles of a tight monetary policy stance and expectations of lower inflation in predicting downturns. While the monetary policy stance is still accommodative, indicating a low recession probability, the negative inflation slope points to higher odds of a recession within a year. Aggressive removal of policy accommodation increases the recession probability to 60%.

C1929: **Gaussian process vector autoregressions and macroeconomic uncertainty**
Presenter: **Niko Hauzenberger**, University of Salzburg, Austria
Co-authors: Florian Huber, M. Marcellino, Nico Petz

A non-parametric multivariate time series model is developed that remains agnostic on the precise relationship between a (possibly) large set of macroeconomic time series and their lagged values. The main building block of our model is a Gaussian Process prior on the functional relationship that determines the conditional mean of the model, hence the name of Gaussian Process vector autoregression (GP-VAR). We control for changes in the error variances by introducing a stochastic volatility specification. To facilitate computation in high dimensions and to introduce convenient statistical properties tailored to match stylized facts commonly observed in macro time series, we assume that the covariance of the Gaussian Process is scaled by the latent volatility factors. We illustrate the use of the GP-VAR by comparing it with other nonlinear and time-varying models in a forecasting exercise. Moreover, we use the GP-VAR to analyze the effects of macroeconomic uncertainty, with a particular emphasis on time variation and asymmetries in the transmission mechanisms.

C1975: **Multi-sector business cycle accounting in a data-rich environment**
Presenter: **Andrew Butters**, Indiana University, United States
Co-authors: Scott Brave

Motivated by a multi-sector general equilibrium model with input-output linkages, we use a structural dynamic factor model to decompose U.S. macroeconomic fluctuations into the contributions of shocks to the four "wedges" commonly used in business cycle accounting: (i) an efficiency, (ii) a labor, (iii) an investment, and (iv) a government wedge. We then evaluate the extent to which shocks to these wedges account for the degree of cross-sectional co-movement in a panel of nearly 150 macroeconomic indicators at business cycle frequencies. We find evidence that the investment and labor wedges are the most likely source of this qualitative feature of business cycles for the U.S., and that the investment wedge played a dominant role in contributing to the depth of the Great Recession and the prolonged weakness of the recovery.

C2010: **Government debt management and inflation with real and nominal bonds**
Presenter: **Alessandro Villa**, Federal Reserve Bank of Chicago, United States

In the wake of rising inflation in the aftermath of unprecedented debt-financed stimulus packages, the question is: Can governments use real bonds (TIPS) as part of their debt portfolio to commit to stable inflation rates? We propose a novel framework of optimal debt management in the presence of sticky prices with a government that can issue nominal and real non-state-contingent bonds. Nominal debt can be inflated away, giving ex-ante flexibility, whereas real bonds are cheaper but constitute a real commitment ex-post. Under Full Commitment, the government chooses a leveraged portfolio of nominal liabilities and real assets to use inflation effectively to smooth fiscal policy. When the government cannot commit to future policies, it reduces borrowing costs ex-ante using real debt strategically to mitigate incentives for the future government to monetize debt ex-post. Without commitment, the policies are quantitatively consistent with US data, suggesting that such a framework realistically captures the relevant constraints governments face.

CO100 Room BH (SE) 2.05 BAYESIAN METHODS FOR EMPIRICAL MACROECONOMICS Chair: Gary Koop

C0244: **Fast two-stage variational Bayesian approach to estimating panel SAR models with unrestricted spatial weights matrices**
Presenter: **Deborah Gefang**, University of Leicester, United Kingdom
Co-authors: Stephen Hall, George Tavlas

A fast two-stage variational Bayesian algorithm is proposed to estimate panel spatial autoregressive models with unknown spatial weights matrices. Using Dirichlet-Laplace global-local shrinkage priors, we are able to uncover the spatial impacts between cross-sectional units without imposing any a priori restrictions. Monte Carlo experiments show that our approach works well for both long and short panels. We are also the first in the

literature to develop VB methods to estimate large covariance matrices with unrestricted sparsity patterns. The method is important because of its relevance to other popular large data models, such as Bayesian vector autoregressions. Matlab code is provided.

C0270: **A new Bayesian MIDAS approach for flexible and interpretable nowcasting**
Presenter: **Galina Potjagailo**, Bank of England, United Kingdom
Co-authors: David Kohns

The T-SV-t-BMIDAS model is proposed for nowcasting quarterly GDP growth. The model incorporates a long-run time-varying trend (T) and t-distributed stochastic volatility accounting for outliers (SV-t) into a Bayesian multivariate MIDAS. To address the high dimensionality of the model, to account for group correlation in mixed frequency data, and to make the model interpretable to the policymaker, we propose a new group-shrinkage prior combined with a sparsification algorithm for variable selection. The prior flexibly accommodates between-group and within-group sparsity and allows communicating the importance of predictors over the data release cycle. We evaluate the model for UK GDP growth nowcasts over the period 1999 to 2021. Our model is competitive before the pandemic relative to various benchmark models, while yielding substantial nowcast improvements during the pandemic. First, accounting for a long-run trend and t-distributed stochastic volatility substantially improves forecast performance relative to a simple BMIDAS. Second, the shrinkage prior enhances nowcast performance by inducing group-wise sparsity while enabling the model to shift flexibly between signals. During the Covid-19 pandemic, the model reads stronger signals from indicators for services, which reflected spending shifts related to lockdowns, and less from production surveys. This helps to nowcast the recovery from the shock precisely, and to update the nowcast for the pandemic-related trough sooner.

C0283: **Combining large numbers of density predictions with Bayesian predictive synthesis**
Presenter: **Tony Chernis**, Bank of Canada, Canada

Bayesian Predictive Synthesis is a flexible method of combining density predictions. The flexibility comes from the ability to choose an arbitrary synthesis function to combine predictions. Being able to choose an arbitrary synthesis function is useful, but what is the correct choice? we study this issue when combining large numbers of predictions - a common occurrence in macroeconomics. Specifically, we examine a Canadian GDP nowcasting exercise with close to 100 models and a GDP forecasting exercise using the European Survey of Professional Forecasters with around 50 experts. Estimating combination weights with so many predictions is difficult, so we consider shrinkage priors and factor modelling techniques as ways to address this problem. These techniques provide an interesting contrast between the sparse weights implied by shrinkage priors and dense weights of factor modelling techniques. We find that the sparse weights of shrinkage priors perform well across exercises.

C0287: **Monthly GDP growth estimates for the U.S. states**
Presenter: **James Mitchell**, Federal Reserve Bank of Cleveland, United States
Co-authors: Gary Koop, Stuart McIntyre, Aristeidis Raftapostolos

A mixed frequency vector autoregressive model (MF-VAR) is developed that exploits available state- and US-level data at the monthly, quarterly, and annual frequencies to produce timely nowcasts and historical monthly estimates of GDP growth for the 50 states of the US (plus Washington, DC) from 1964 through the present day. The model imposes temporal and cross-sectional constraints to ensure that the monthly estimates both 'add up' to published quarterly/annual data and that the GDP estimates for the 50 states (plus DC) sum to published GDP data for the US as a whole. We develop a computationally-fast approximate Bayesian Markov Chain Monte Carlo (MCMC) algorithm for estimating and nowcasting with this large-scale MF-VAR. The model is used to produce historical estimates of monthly GDP for the 50 (plus DC) US states back to the 1960s, and the utility of these estimates is illustrated by using them to understand business cycle dynamics and cross-state dependencies better. A nowcasting application, using real-time data, then shows how the model can be used to produce density estimates of state-level GDP two months ahead of the BEA's quarterly state-level estimates. We show the importance for nowcast accuracy of conditioning these state-level nowcasts on the latest estimates of US GDP from the BEA.

CO332 Room BH (SE) 2.09 ECONOMETRIC FORECASTING Chair: Onno Kleen

C1254: **A forest full of risk forecasts for managing volatility**
Presenter: **Anastasija Tetereva**, Erasmus University Rotterdam, Netherlands
Co-authors: Onno Kleen

A heterogeneous autoregressive (HAR) model is proposed with time-varying parameters in the form of a local linear random forest. In contrast to conventional random forests that approximate the volatility nonparametrically using local averaging, the building blocks of our forest are HAR panel models. The local HAR panel models cover the established linear relationship in realized variances, while the trees model nonlinearities and interaction effects. The approach allows the model coefficients to depend on idiosyncratic stock information and overall changing market conditions. We observe superior risk forecasting performance of the HAR forest across multiple forecast horizons and across 186 S&P 500 constituents. This leads to a significantly higher utility for volatility-managed portfolios. Superior forecast performance is especially pronounced for firms with high leverage.

C1299: **Testing for differences in survey-based density expectations: a compositional data approach**
Presenter: **Alexander Glas**, FAU Erlangen-Nuernberg, Germany
Co-authors: Jonas Dovern, Geoff Kenny

Survey-based density expectations are proposed to be treated as compositional data when testing either for changes in density forecasts over time or for heterogeneity across different groups of forecasters. Monte Carlo simulations show that our test has more power relative to both a bootstrap approach based on the KLIC and an approach which involves multiple testing for differences of individual parts of the density. In addition, the test is much faster than the KLIC-based one since it does not rely on simulations and allows for comparisons across multiple groups. Using density expectations from the ECB Survey of Professional Forecasters and the U.S. Survey of Consumer Expectations, we show the usefulness of the test in detecting possible changes in density forecasts over time and across different types of forecasters.

C1359: **Real-time nowcasting growth at risk**
Presenter: **Manuel Schick**, Heidelberg University, Germany
Co-authors: Christian Conrad

Macroeconomic and financial indicators are exploited to nowcast Growth at Risk (GaR) in a Mixed-Data Sampling (MIDAS) framework. Indicators are used at high-frequency with exact timing, i.e., as they become available to forecasters in real-time according to exact publication dates. A small number of factors for both mean and conditional variance are incorporated in the MIDAS model, resulting in a parsimoniously parameterized prediction model that provides daily updates of GaR of the current quarter. In particular, the conditional variance features a long-term volatility component of the S&P 500 index that is extracted by means of the MF2-GARCH model. Therefore, financial conditions are monitored on a daily basis which provides a timely signal for downside risk. The results show that the new model provides good out-of-sample GaR nowcasts.

C1824: **Forecasting performance of dynamic approaches for spillovers in networks**
Presenter: **Rebekka Buse**, Karlsruhe Institute of Technology, Germany

The purpose is to study the out-of-sample forecasting performance of smooth versus rolling window approaches for evaluating spillovers in dynamic networks. Each approach has particular advantages: The frequentist approach with rolling window vector autoregression (VAR), on the one hand, is very precise with respect to the dynamics within each window, while the Bayesian approach with time-varying parameter VAR, on the other

hand, yields dynamics that are smooth and independent of window size. We compare the out-of-sample forecast performance of both approaches in order to distinguish the most precise and up-to-date measure for real-time interconnected systems. In a comprehensive simulation study, we investigate both level data as well as resulting network measures for different types of dynamics, including not only times of steady evolution, but also inflection points of different shapes and succession. We put a particular emphasis on evaluating the performance for these different types of settings and provide guidance on the choice of the method depending on the setting type taking tuning parameter options into account. Our empirical application to U.S. market sectors reveals that not only financial sectors, but also the real economy and, in particular, the commodity industries play an equally important role in understanding interaction effects.

CO136 Room BH (SE) 2.10 MACHINE LEARNING IN ASSET PRICING Chair: Markus Pelger

C0872: **Asset-pricing factors with economic targets**
Presenter: **Sicong Allen Li**, London Business School, United Kingdom
Co-authors: Victor DeMiguel, Svetlana Bryzgalova, Markus Pelger

A method is proposed for estimating latent asset-pricing factors that incorporate economic information. Our estimator generalizes Principal Component Analysis (PCA) by penalizing deviations of the factors from economically motivated cross-sectional and time-series targets. Using a high-dimensional dataset containing decile-portfolio returns for 37 characteristics, we show that a cross-sectional target that aligns factor composition with firm characteristics and a time-series target that nudges factors to explain decile-portfolio alphas help to span pricing kernels with substantially higher Sharpe ratio and lower pricing error than conventional PCA factors.

C1486: **Factor models for conditional asset pricing**
Presenter: **Paolo Zaffaroni**, Imperial College London, United Kingdom

A methodology is developed for inference on conditional asset pricing models robust to omitted risk factors and misspecified conditional dynamics. All the features of the asset pricing model, such as risk premia, factors exposures, factors variances and covariances, idiosyncratic risk, and number of risk factors, are potentially time-varying. The limiting results hold when the number of assets diverges, but the time-series dimension is fixed, possibly very small, and applicable to a variety of data frequencies. An extensive empirical application based on individual asset returns data demonstrates the powerfulness of the methodology, allowing us to tease out the empirical content of the time-variation elicited by asset pricing theory.

C1436: **Term structure of characteristic-sorted portfolios and multi-horizon investment**
Presenter: **Svetlana Bryzgalova**, London Business School, United Kingdom
Co-authors: Serhiy Kozak, Markus Pelger

The term structure of expected returns is explored across multiple horizons of portfolios sorted on contemporaneous characteristics. At the core of our approach is the tensor factor model — a generalization of APT which explains returns on contemporaneously-sorted characteristics portfolios across many investment horizons. Equivalently, the model provides a parsimonious representation of contemporaneous returns sorted on multiple lags of characteristics. The model fits the data well and generates robust multi-horizon predictions and MVE portfolio weights for long-horizon investment.

C0848: **Conditional latent factor models via econometrics-based neural networks**
Presenter: **Hao Ma**, Queen Mary University of London, United Kingdom

A hybrid methodology is developed that incorporates an econometric identification strategy into artificial neural networks when studying conditional latent factor models. The time-varying betas are assumed to be unknown functions of numerous firm characteristics, and the statistical factors are population cross-sectional OLS estimators for given beta values. Hence, identifying betas and factors boils down to identifying only the function of betas, which is equivalent to solving a constrained optimization problem. For estimation, we construct neural networks customized to solve the constrained optimization problem, which gives a feasible non-parametric estimator for the function of betas. Empirically, we conduct my analysis on a large unbalanced panel of monthly data on US individual stocks with around 30, 000 firms, 516 months, and 94 characteristics. We find that 1) the hybrid method outperforms the benchmark econometric method and the neural network's method in terms of explaining out-of-sample return variation, 2) betas are highly non-linear in firm characteristics, 3) two conditional factors explain over 95% variation of the factor space, and 4) hybrid methods with literature-based characteristics (e.g., book-to-market ratio) outperform ones with COMPUSTAT raw features (e.g., book value and market value), emphasizing the value of academic knowledge from an angle of Man vs Machine.

CO306 Room BH (SE) 2.12 FINANCIAL ENGINEERING Chair: Genevieve Gauthier

C0187: **Long memory in option pricing: A fractional discrete-time approach**
Presenter: **Jean Francois Begin**, Simon Fraser University, Canada
Co-authors: Maciej Augustyniak, Alex Badescu, Sarath Kumar Jayaraman

The impact of long memory on asset return modelling and option pricing is studied. We propose a general discrete-time pricing framework based on affine multi-component volatility models that admit ARCH(∞) representations. It not only nests a plethora of option pricing models from the literature but also allows for the introduction of novel fractionally integrated processes for option valuation purposes. Using an infinite sum characterization of the unconditional cumulant generating function of the log-asset price, we derive semi-explicit expressions for European option prices under a volatility-dependent stochastic discount factor. We carry out an extensive empirical analysis which includes returns-only as well as return and option joint estimations of a variety of short- and long-memory models for the S&P 500 index. Our results indicate that the inclusion of long memory into return modelling substantially improves the option pricing performance. Using a set of out-of-sample option pricing errors, we show that long-memory models outperform richer-parametrized one- and two-component models with short-memory dynamics.

C0195: **Option pricing under stochastic volatility models with latent volatility**
Presenter: **Frederic Godin**, Concordia University, Canada
Co-authors: Jean-Francois Begin

An important challenge regarding the pricing of derivatives is related to the latent nature of volatility. Most studies disregard the uncertain nature of volatility when pricing options; the few authors who account for it typically consider the risk-neutral posterior distribution of the latent volatility. As the latter distribution differs from its physical measure counterpart, this leads to at least two issues: (1) it generates some unwanted path dependence, and (2) it oftentimes requires simultaneously tracking the physical and risk-neutral distributions of the latent volatility. Pricing approaches purging such a path-dependence issue are presented. This is achieved by modifying conventional pricing approaches (e.g., the Girsanov transform) to formally recognize the uncertainty about the latent volatility during the pricing procedure. The two proposed risk-neutral measures circumventing the aforementioned undesired path-dependence feature are based on either the extended Girsanov principle or the Esscher transform. We also show that such pricing approaches are feasible, and we provide numerical implementation schemes.

C0264: **Foreseeing the worst: Forecasting electricity DART spikes**
Presenter: **Remi Galarneau-Vincent**, HEC Montreal, Canada
Co-authors: Genevieve Gauthier, Frederic Godin

Statistical learning models are proposed for the prediction of the probability of a spike in the electricity DART (day-ahead minus real-time price)

spread. Assessing the likelihood of DART spikes is of paramount importance for virtual bidders, among others. The model's performance is evaluated on historical data for the Long Island zone of the New York Independent System Operator (NYISO). A tailored feature set encompassing novel engineered features is designed. Such a set of features makes it possible to achieve excellent predictive performance and discriminatory power. Results are shown to be robust to the choice of the predictive algorithm. Lastly, the benefits of forecasting the spikes are illustrated through a trading exercise, confirming that trading strategies employing the model-predicted probabilities as a signal generate consistent profits.

C0837: **How does price discovery take place in the option market: Evidence from S&P 500 index options**

Presenter: **Diego Amaya**, Wilfrid Laurier University, Canada

Co-authors: Fabricio Perez

A novel methodology is proposed to quantify the price discovery contribution of options to the underlying asset value. Using options on the S&P 50 index, we show fundamental changes in this market during the sample period between 2004 and 2018. We find an important shift in price discovery from call to put options, and from long-term contracts to short-term contracts. We show that volatility and jump risks have different effects on the information share of options, and that these effects vary across years, contract types, and maturities. Our results suggest that information flows in the option market change across time and depend on option characteristics, implying that these contracts are actively used for risk-sharing opportunities in this market.

CC803 Room BH (SE) 1.05 REALIZED VOLATILITY Chair: Edward Knotek

C0476: **Monitoring European realized volatility during recent structural change**

Presenter: **Robinson Kruse-Becher**, FernUniversitat in Hagen, Germany

Realized volatility measures are likely to experience structural shifts during crises. These permanent shifts are important in a long-run perspective and informative about changes in market risk. Timely detection of mean shifts in volatility can be beneficial for portfolio choice, risk controlling and forecasting. Related to the literature on GARCH models with shifting intercepts, we focus on realized volatility measures and long memory features instead. We contribute to long memory and level shifts by studying the monitoring properties of popular CUSUM detectors. To this end, we compare the CUSUM detector (i) applied to fractionally differenced data and (ii) using memory autocorrelation consistent long-run variance estimation to a naive version ignoring fractional integration. Another issue under investigation is the role of the Local Whittle estimator for the fractional integration parameter, which is typically biased due to low-frequency contaminations like random level shifts. In an empirical application, we demonstrate the merits and limitations of real-time CUSUM monitoring for realized volatility of European stock market indices under recent crises.

C0731: **Long memory realised GAS model**

Presenter: **Ekaterina Ugulava**, University of Amsterdam, Netherlands

A new score-driven model is introduced, which explicitly incorporates long memory dynamics in the conditional variance of daily returns. First, we model the conditional variance as a fractionally integrated (FI) process and, importantly, allow the long memory parameter to cross the stationarity boundary $d = 0.5$. Second, we adopt a heterogeneous autoregressive (HAR) model that parsimoniously approximates long memory dynamics through mixed-frequency mean components. The new model accommodates heavy-tailed densities for both the daily returns and realised measures. This choice of observational densities ensures an automatic correction for the influential observations through the score. In an empirical study conducted for fifteen individual components of the DJI index over the period 2001-2010, we find that likelihood favours the new FI model with $d > 0.5$. Our out-of-sample analysis identifies that accounting for long memory through either the FI or the HAR form is particularly useful for volatility level evaluation and return risk assessment during non-crisis periods.

C1413: **Realized volatility forecasting using extreme gradient boosting**

Presenter: **Andreas Teller**, Friedrich Schiller University Jena, Germany

Co-authors: Uta Pigorsch, Christian Pigorsch

Extreme Gradient Boosting (XGBoost) is adopted to forecast realized volatility. This is motivated by XGBoost's strong forecasting performance in other forecast applications and its ability to capture non-linearities, a feature that is also oftentimes reported in the context of realized volatility. We examine the forecasting precision of linear and non-linear XGBoost models for different forecast horizons and compare it to that of Long Short-Term Memory (LSTM) networks as well as heterogeneous autoregressive (HAR) models. We find that XGBoost exhibits a better forecast performance. In particular, XGBoost models significantly outperform both HAR and LSTM models for one-step-ahead predictions. For longer forecast horizons, linear models such as XGBoost with linear base learners perform better than non-linear specifications, suggesting that accounting for non-linearities is only important if short-term forecasts are of interest.

C1664: **Revisiting the evidence for rough volatility in daily realized variances**

Presenter: **Hiroyuki Kawakatsu**, Dublin City University, Ireland

The evidence for rough volatility in the daily log realized variance series is reexamined. For the class of ARMA(1,1) models with fractional Gaussian noise, there are two nearly observationally equivalent models, one with short memory and one with long memory. The practical difference between the two models is explored through comparisons of persistence in autocorrelations and (pseudo)out-of-sample performance.

Saturday 17.12.2022 18:05 - 19:20 Parallel Session F – CFE-CMStatistics

EO669 Room S-2.23 ADVANCES IN MHEALTH METHODS Chair: Walter Dempsey

E1379: **Testing non-stationarity and quantifying associations in the presence of missing data in time series of mHealth studies**
Presenter: **Linda Valeri**, Columbia University, United States
Co-authors: Charlotte Fowler, Xiaoxuan Cai

The use of digital devices to collect data in mobile health (mHealth) studies introduces a novel application of time series methods, with the constraint of potential data missing at random (MAR) or missing not at random (MNAR). In time series analysis, testing for stationarity is an important preliminary step to inform appropriate later analyses. Further, appropriately accounting for missing data is crucial to estimate exposure effects in multivariate time series settings validly. The augmented Dickey-Fuller (ADF) test was developed to test the null hypothesis of unit root non-stationarity, under no missing data. We propose maximum likelihood estimation and multiple imputation using a state space model approach to adapt the ADF test and evaluate associations among multivariate time-series in a context with missing data. We further develop sensitivity analysis techniques to examine the impact of MNAR data. We evaluate the performance of existing and proposed methods across different missing mechanisms in extensive simulations and their application to a multi-year smartphone study of bipolar patients.

E1951: **Sleep classification with artificial synthetic imaging data using convolutional neural networks**
Presenter: **Peter Song**, University of Michigan, United States

A new analytic framework is proposed, Artificial Synthetic Imaging Data (ASID) Workflow, for sleep classification from a wearable device comprising: 1) the creation of ASID from data collected by a non-invasive wearable device that permits real-time multi-modal physiological monitoring of heart rate (HR), 3-axis accelerometer, electrodermal activity, and skin temperature and 2) the use of an image classification supervised learning algorithm, convolutional neural network (CNN), to classify periods of sleep. We compare our ASID Workflow with competing machine/deep learning classification algorithms, including logistic regression, support vector machine, random forest, k-nearest neighbors, and Long Short-Term Memory. The ASID Workflow achieves excellent performance with high mean weighted accuracy and is superior to the Competing Workflow. We explore specifically the influence of data resolution and HR modality on the Workflow's performance in order to achieve the desirable cost-and-effectiveness of data collection. Applying CNN to ASID allows us to capture both temporal and spatial dependency among physiological variables and modalities by using 2D images' topological structure that competing algorithms fail to utilize.

E1346: **Reinforcement learning in possibly nonstationary environments**
Presenter: **Zhenke Wu**, University of Michigan at Ann Arbor, United States
Co-authors: Mengbing Li, Chengchun Shi, Piotr Fryzlewicz

Reinforcement learning (RL) methods are considered in offline nonstationary environments. Many existing RL algorithms in the literature rely on the stationarity assumption that requires the system transition and the reward function to be constant over time. However, the stationarity assumption is restrictive in practice and is likely to be violated in a number of applications, including traffic signal control, robotics and mobile health. We introduce a consistent procedure to test the nonstationarity of the optimal policy based on pre-collected historical data, without additional online data collection. Based on the proposed test, we further develop a sequential change point detection method that can be naturally coupled with existing state-of-the-art RL methods for policy optimisation in nonstationary environments. The usefulness of our method is illustrated by theoretical results, simulation studies, and a real data example from the 2018 Intern Health Study. A Python implementation of the proposed procedure is available at GitHub.

EO362 Room S-2.25 STATISTICAL AND MACHINE LEARNING METHODS FOR ANALYSIS OF EHR/RWD Chair: Emily Getzen

E0546: **Automated harmonization of multi-institutional electronic health records data**
Presenter: **Xu Shi**, University of Michigan, United States

Current practice for electronic health records (EHR) data harmonization involves standardizing data elements via a common data model, a critical step that unifies the medical coding "vocabulary" across study sites. However, despite a common vocabulary, the coding "dialect" (i.e., the use and interpretation of codes for a particular clinical procedure or diagnosis) may differ across data partners due to heterogeneity in care practice and financial drivers. With increasingly diverse health systems and coding systems, there is more and more potential variation in the way a clinical concept can be coded. Existing manually curated ontology and mapping to reduce heterogeneity and harmonize data are not scalable and error-prone. Data-sharing constraints bring additional challenges to statistical analysis across institutions. We will present data-driven and privacy-preserving statistical methods for detecting and reducing coding differences between healthcare systems. We will share our findings from a case study of EHR data harmonization between two healthcare institutions.

E0637: **Double sampling and semiparametric methods for informatively missing data**
Presenter: **Sebastien Haneuse**, Harvard TH Chan School of Public Health, United States

Large observational databases, such as those derived from electronic health records (EHR), are increasingly being used for clinical and public health research. Despite the many benefits, these data are often subject to complex and poorly understood patterns of missing data, such that the typical missing-at-random assumption may be untenable. In contrast to traditional methods of sensitivity analysis and estimation of parameter bounds, we explore double sampling in which complete data can be obtained on a subsample via intensive follow-up. We discuss assumptions and designs under which the joint density of interest is identified, and present a general approach for constructing estimators in the augmented sample. From this analysis, we show when the initial missingness process itself is identified, and how the associated missing-at-random assumption can be tested. Further, we apply the framework to derive semiparametric efficient and multiply robust estimators of causal average treatment effects from double-sampled observational data when outcome data are initially missing, not at random.

E0880: **Mining for equitable health: Assessing the impact of missing data in electronic health records**
Presenter: **Emily Getzen**, University of Pennsylvania, United States
Co-authors: Lyle Ungar, Danielle Mowery, Xiaoqian Jiang, Qi Long

Electronic health records (EHRs) contain multiple years of health information to be leveraged for disease detection and treatment evaluation. However, they do not have standardized formatting, and can present significant analytical challenges– they contain multi-scale data from heterogeneous domains and include both structured and unstructured data. Data for individual patients are collected at irregular time intervals and with varying frequencies. In addition to the analytical challenges, EHRs can reflect inequity– patients belonging to different groups will have differing amounts of data. The consequence is that the data for marginalized groups may be less informative due to more fragmented care, which can be viewed as a missing data problem. For EHR data in this complex form, there is currently no framework for introducing missing values. There has also been little to no work in assessing the impact of missing data in EHRs. We simulate realistic missing data scenarios in EHRs to adequately assess their impact on predictive modeling. We incorporate the use of a medical knowledge graph to capture dependencies between medical events to create a more realistic missing data framework. In an intensive care unit setting, we found that missing data have a greater negative impact on the performance of disease prediction models in groups that tend to have less access to healthcare.

EO438 Room S-1.01 FINITE POPULATION INFERENCE USING ML AND LATENT VARIABLE MODELS Chair: Maria Giovanna Ranalli

E1042: **Latent variable models and machine learning for prediction of employment status in Italy**
Presenter: **Roberta Varriale**, University of Rome La Sapienza, Italy
Co-authors: Marco Alfo

The increasing availability of a large amount of multi-source information in national statistical institutes makes it necessary to investigate new methodological approaches, based on combining primary and secondary data, for the production of estimates. Primary data are collected by NSIs for statistical purposes, usually using a statistical sample survey. Secondary data, such as administrative registers and big data, are not collected by NSIs, and are not collected for statistical purposes. Still, they may be used by NSIs for producing statistics. In the context of qualitative/categorical data, there are different methodological approaches to produce estimates by exploiting all available information. Latent variable models may help take explicitly into account deficiencies in the measurement process of both survey and administrative sources. Machine learning techniques are frequently used to classify large amounts of data. The use of Hidden Markov Model and Machine Learning methods is described in the labour statistics context to predict the individual employment status. The relevant data may be drawn from the labour force survey conducted by Istat and from several administrative sources that Istat regularly acquires from external bodies.

E0700: **Design-based ensemble learning for individual prediction in finite populations**
Presenter: **Li-Chun Zhang**, University of Southampton, United Kingdom

One can distinguish between ensemble methods which use a single base learning algorithm to produce homogeneous base predictors (or learners), such as bagging, and those using heterogeneous (component) predictors of different types. The two basic combination methods for heterogeneous ensembles are voting and averaging, whereas gating and stacking are examples of methods which combine by learning a meta-learner. We present a design-based approach to ensemble learning by voting or averaging based on the expected cross-validation results, given the sampling design and the sample-splitting design for cross-validation. Valid inference of the uncertainty of ensemble prediction for finite populations is defined and obtained with respect to the known sampling design, regardless if the assumed model that facilitates prediction is correct or not.

E1844: **Multivariate small area estimation of educational poverty with latent variable models**
Presenter: **Gaia Bertarelli**, Sant'Anna school of Advanced Studies, Italy
Co-authors: Maria Giovanna Ranalli, Monica Pratesi

Educational Poverty (EP) for young adults can be read as a deprivation of opportunities and rights related to culture, participation, environment, and social relations. It means being excluded from acquiring the skills needed to live in a world characterized by a knowledge-based economy and innovation. It is a latent trait, only indirectly measurable through a collection of observable variables and indicators purposively selected as micro-aspects, contributing to the latent macro-dimension. EP is measured in Italy by the Educational Poverty Index. A problem with this index is that it is based on direct estimates, which are reliable only at a regional level, while to intervene in the phenomenon, it is important to obtain information at a finer geographical level. This problem has been overcome by considering estimates based on a Fay and Herriot model in the aggregation process. However, none of the proposed indicators considers the true latent nature of the phenomenon. We aim to go beyond this limit and develop a new multidimensional indicator at a small area level, which is based on a unit-level latent variable model in order to capture the underlying hidden dimensions of EP from a set of binary manifest indicators. The proposed model is applied to data from the aspects of everyday life survey in Italy focusing on Provinces and on suburbs in Italian Regions.

EO426 Room S-1.06 NEW ALTERNATIVES TO SIGNIFICANCE TESTING Chair: Simon Vandekar

E1155: **A single framework for the analysis of effect sizes in cross-sectional and longitudinal studies**
Presenter: **Kaidi Kang**, Vanderbilt University, United States
Co-authors: Kristan Armstrong, Suzanne Avery, Maureen McHugo, Stephan Heckers, Simon Vandekar

Effect size indices are useful tools for communicating study findings. Reporting effect size estimates with their confidence intervals (CIs) can be an excellent way to simultaneously communicate the strength of the observed evidence and the precision of the evidence. However, the existing effect size indices are all highly restricted by model type. They are limited to the cross-sectional study setting, and the indices for longitudinal analysis are poorly defined. These restrictions unavoidably lead to difficulties in effect size communication not only between cross-sectional studies using different models but also between cross-sectional and longitudinal studies and longitudinal studies with different numbers of measurements. We previously proposed a robust effect size index (RESI) which is advantageous over common indices, especially because it is widely applicable across different models such that cross-sectional studies using different models can easily report effect sizes with CIs in an analysis of variance (ANOVA) table format. In the current research, we extended the framework to longitudinal analysis and defined a new RESI that is not affected by the number of measurements. Thus, researchers studying the same scientific questions but using different data types or study designs (cross-sectional or longitudinal) can easily communicate their observed effect sizes and CIs without having to translate between them.

E1522: **Accurate estimation of effect sizes: A sequential approach for scientific advancement**
Presenter: **Ken Kelley**, University of Notre Dame, United States

Sequential estimation (SE) is a well-recognized approach to inference in statistical theory. In SE the sample size to use is not specified at the start of the study, and instead, the data itself and the goals of the researcher guide when a predefined stopping rule is met. Thus, rather than a fixed sample size approach to study design, which is usually based on supposed values, the final sample size in SE is unknown. This is positive because sampling stops once the goal is met, but it is negative because the necessary sample size might be larger than a researcher is able or willing to obtain. SE for accurate estimation is discussed. Then, a general effect size is discussed. Then, the two are combined into a method for obtaining an accurate estimate of this general effect size. Accurate estimation is operationalized as a sufficiently narrow confidence interval, where the goals of the research determine the desired narrowness of the confidence interval since narrower intervals illustrate less uncertainly in the estimated effect, holding the confidence level constant. Termed sequential accurately in parameter estimation, which does not require the pre-specification of supposed population parameters, as is generally necessary for power analysis in a null hypothesis significance testing framework. The premise is that if an effect size is of interest for a research study, the study should be such that an accurate effect size is obtained.

E1700: **An introduction to second-generation p-values**
Presenter: **Jeffrey Blume**, University of Virginia, United States

Despite decades of controversy, p-values remain a popular tool for assessing when the data are incompatible with the null hypothesis. While it is widely recognized that p-values are imperfect, the consequences of ignoring their flaws remain elusive and p-values continue to flourish in the scientific literature. We will introduce the second-generation p-value, a novel and intuitive extension that better serves the intended purpose. We will illustrate their use in several examples, including a high-dimensional genomics setting, and show how their implicit emphasis on scientific relevance leads to improved statistical performance (as measured by error rates or false discovery rates).

EO196 Room S-1.22 SCREENING AND VARIABLE SELECTION IN HIGH-DIMENSIONAL SURVIVAL DATA Chair: Marialuisa Restaino

E0186: **Effects selection via likelihood-based boosting in Cox frailty models**
Presenter: **Andreas Groll**, Technical University Dortmund, Germany

In regression tasks, dealing with high-dimensional data has become more and more important with lots of potentially influential covariates. A

possible solution is to apply estimation methods that aim at the detection of the relevant effect structure by using regularization methods such as, e.g. boosting or penalization. The effect structure in the Cox frailty model, which is the most widely used model that accounts for heterogeneity in survival data, is investigated. Since in survival models, one has to account for a possible variation of the effect strength over time, the selection of the relevant features has to distinguish between several cases, covariates can have time-varying effects, can have time-constant effects or be irrelevant. A likelihood-based boosting approach is presented, which is able to distinguish between these types of effects to obtain a sparse representation that includes the relevant effects in a proper form. This idea is applied to a real-world data set, illustrating that the complexity of the influence structure can be strongly reduced by using such a regularization approach.

E1036: **Feature selection for competing risks models: A comparison**
Presenter: **Marialuisa Restaino**, University of Salerno, Italy

In the analysis of time-to-event data, competing risks data are encountered when individuals may fail from multiple causes, and the occurrence of one failure event precludes the others from happening. To analyze the effects of covariates on the hazard function, both the cause-specific hazard (CSH) model and the subdistribution hazard (SDH) model. The main difference is in the definition of the risk set. In CSH, subjects who experience the competing events are treated as censored, while in SDH they are included in the risk set. In both settings, and in presence of a large number of variables, it becomes crucial to identify those variables that may affect the hazard. While in the CSH model, screening and variable selection methods developed for Cox model can be easily extended, for the SDH approach, naive applications of these procedures may be problematic and not suitable. It is due to the definition of the risk set. In the present work, the aim is to compare the performance of some existing methods for screening and selecting the most significant variables, for both CSH and SDH models, for highlighting their main advantages and disadvantages and proposing a new procedure able to identify the relevant covariates in the framework of high and ultra-high dimensions and also in presence of highly correlated variables.

E1256: **Model-free variable screening for ultrahigh dimensional survival data with FDR control**
Presenter: **Chenlu Ke**, Virginia Commonwealth University, United States

A novel framework is proposed for variable screening for ultrahigh dimensional survival data. The contribution of each individual predictor to the survival outcome is quantified in the presence of the other candidates by kernel-based R-squared statistics. Compared with existing marginal screening methods, our proposal does not require an intermediate estimation of the survival function and relaxes the commonly imposed assumption of independent censoring. Moreover, our method can capture hidden important predictors that are marginally independent but jointly dependent on the survival outcome. We establish the sure screening property and the rank consistency property of the proposed approach in the notion of sufficiency. A knockoff procedure is also developed for controlling false discoveries. The advantages of the proposed method are demonstrated by simulation studies and an application to high-throughput gene expression data.

EO420 Room S-1.27 ADVANCES IN STATISTICAL NEUROIMAGING AND SPATIO-TEMPORAL MODELING Chair: Rajarshi Guhaniyogi

E0845: **Bayesian image-on-scalar regression with a spatial global-local spike-and-slab prior**
Presenter: **Meng Li**, Rice University, United States
Co-authors: Zijian Zeng, Meng Li, Marina Vannucci

Image-on-Scalar regression has wide-ranging applications in discovering the relationship between the image data and covariates measured on the same subjects. This remains challenging partly because of the highly complex spatial dependency in image data as well as the demand for the selection and interpretation of influential covariates at more than one level. We develop a method for simultaneous image smoothing, parameter estimation, and variable selection at both the image and pixel levels. We consider a Bayesian hierarchical Gaussian process model for image smoothing, that uses a flexible Inverse-Wishart process prior to handle within-image dependency, and propose a general global-local spatial selection prior that extends a rich class of well-studied selection priors. Unlike existing constructions, we achieve simultaneous global (i.e, at covariate-level) and local (i.e., at pixel/voxel-level) selection by introducing "participation rate" parameters that measure the probability for the individual covariates to affect the observed images. This, along with a hard-thresholding strategy, leads to dependency between selections at the two levels, introduces extra sparsity at the local level, and allows the global selection to be informed by the local selection, all in a model-based manner. We design an efficient Gibbs sampler that allows inference for large image data. The proposed method is demonstrated by using data from the Autism Brain Imaging Data Exchange (ABIDE) study.

E0853: **Quantum mechanics uncertainty, data science inference, and AI in complex time (kime)**
Presenter: **Ivo Dinov**, University of Michigan, United States
Co-authors: Milen Velev, Yueyang Shen

By using complex time (kime) to lift the classical 4D space-time into the 5D space-kime manifold, translating quantum mechanics principles to address data science and predictive analytics challenges will be demonstrated. We extend the physical laws of velocity, momentum, Lorentz transformations, and 4D solutions of Einstein's equations to their corresponding counterparts in 5D spacekime. Direct AI applications include transforming classical random sampling in spacetime to spacekime phase-uncertainty and a Bayesian formulation of spacekime analytics. Using neuroimaging and macroeconomics data, we will show examples mapping longitudinal data (e.g., time-series) to 2D manifolds (e.g., kime-surfaces) and discuss the subsequent modeling, inference, and AI based on space-kime representations.

E1437: **Latest developments in Bayesian image analysis in Fourier space (BIFS) models**
Presenter: **John Kornak**, University of California, San Francisco, United States
Co-authors: Karl Young, Eric Friedman, Ross Boylan

For more than 30 years now, Bayesian image analysis has been a leading approach to image reconstruction and enhancement. The idea of the approach is to balance a priori expectations of image characteristics (the prior) with a model for the image degradation process (the likelihood). The conventional Bayesian modeling approach, as defined in image space, implements priors that describe inter-dependence between spatial locations on the image lattice (commonly through Markov random field, MRF, models) and can therefore be difficult to model and compute. Bayesian image analysis in Fourier space (BIFS) provides for an alternate approach that can generate a wide range of models, including ones with similar properties to conventional models, but with reduced computational burden; the originally complex high-dimensional estimation problem in image space can be similarly modeled as a series of (trivially parallelizable) independent one-dimensional problems in Fourier space. A range of prior models that can be formulated in Fourier space will be illustrated, including MRF-matched models and frequency-selective models, and these will be compared to conventional models. In addition, extensions will be briefly discussed based on a) a data-driven prior approach and b) transforming to the wavelet domain.

EO388 Room K0.16 STATISTICAL METHODS FOR CONSTRUCTING AND ANALYZING NETWORKS (VIRTUAL) Chair: Simon Preston

E0357: **Network regression with graph Laplacians**
Presenter: **Yidong Zhou**, University of California, Davis, United States
Co-authors: Hans-Georg Mueller

Network data are increasingly available in various research fields, motivating statistical analysis for populations of networks where a network as a whole is viewed as a data point. Due to the non-Euclidean nature of networks, basic statistical tools available for scalar and vector data are no longer applicable when one aims to relate networks as outcomes to Euclidean covariates, while the study of how a network changes in dependence

on covariates is often of paramount interest. This motivates us to extend the notion of regression to the case of responses that are network data. We propose to adopt conditional Fréchet means implemented with both global least squares regression and local weighted least squares smoothing, extending the Fréchet regression concept to networks that are quantified by their graph Laplacians. The challenge is to characterize the space of graph Laplacians so as to justify the application of Fréchet regression. This characterization then leads to asymptotic rates of convergence for the corresponding M-estimators by applying empirical process methods. We demonstrate the usefulness and good practical performance of the proposed framework with simulations and with network data arising from resting-state fMRI in neuroimaging, as well as New York taxi records.

E0505: **Causal inference under network interference with noise**
Presenter: **Wenrui Li**, University of Pennsylvania, United States
Co-authors: Daniel Sussman, Eric Kolaczyk

Increasingly, there is a marked interest in estimating causal effects under network interference due to the fact that interference manifests naturally in networked experiments. However, network information generally is available only up to some level of error. We study the propagation of such errors to estimators of average causal effects under network interference. Specifically, assuming a four-level exposure model and Bernoulli random assignment of treatment, we characterize the impact of network noise on the bias and variance of standard estimators in homogeneous and inhomogeneous networks. In addition, we propose method-of-moments estimators for bias reduction where a minimal number of network replicates are available. We show our estimators are asymptotically normal and provide confidence intervals for quantifying the uncertainty in these estimates. We illustrate the practical performance of our estimators through simulation studies in British secondary school contact networks.

E1433: **Statistical approaches for networks and trees arising from natural language data**
Presenter: **Simon Preston**, University of Nottingham, United Kingdom
Co-authors: Katie Severn, Ian Dryden

Networks and trees arise as natural representations of text data, for example, in characterising word-pair co-occurrence, or the syntactic structure of individual sentences. We will discuss networks and trees as "object data", i.e. in which they are treated as the statistical unit of observation, but with non-Euclidean sample space, and outline some statistical approaches we have developed for regression, two-sample testing and classification.

EO658 Room K0.18 ADVANCES IN BAYESIAN METHODS TO RECORD LINKAGE AND SURVEY METHODOLOGY Chair: Jairo Fuquene

E1510: **Fast Bayesian record linkage for streaming data contexts**
Presenter: **Andee Kaplan**, Colorado State University, United States
Co-authors: Ian Taylor, Andee Kaplan, Brenda Betancourt

Record linkage is the task of combining records from multiple files which refer to overlapping sets of entities when there is no unique identifying field in the records. In streaming record linkage, files arrive sequentially in time and estimates of links are updated after the arrival of each file. This problem arises in settings such as longitudinal surveys, electronic health records, and online events databases, among others. The challenge in streaming record linkage is to efficiently update parameter estimates as a new data file arrives. We approach the problem from a Bayesian perspective with estimates in the form of posterior samples of parameters and present methods for updating link estimates after the arrival of a new file that is faster than fitting a joint model with each new data file. We generalize a two-file Bayesian Fellegi-Sunter model to the multi-file case and propose two methods to perform streaming updates. We examine the effect of prior distribution on the resulting linkage accuracy as well as the computational trade-offs between the methods when compared to a Gibbs sampler through simulated and real-world survey panel data. We achieve near-equivalent posterior inference at a small fraction of the compute time.

E1618: **Microclustering for record linkage applications**
Presenter: **Brenda Betancourt**, NORC at the University of Chicago, United States

In database management, record linkage aims to identify multiple records that correspond to the same individual. Record linkage can be treated as a clustering problem in which one or more noisy database records are associated with a unique latent entity. In contrast to traditional clustering applications, a large number of clusters with a few observations per cluster is expected in this context. Hence, two new classes of prior distributions based on exchangeable sequences of clusters and allelic partitions are proposed for the small cluster setting of record linkage. The proposed priors facilitate the introduction of information about the cluster size distribution at different scales, and naturally enforces sublinear growth of the maximum cluster size, known as the microclustering property. In addition, a set of novel microclustering conditions are introduced in order to impose further constraints on the cluster sizes a priori. The performance of the proposed classes of priors is evaluated using simulated data and official statistics data sets.

E1735: **A Bayesian approach to modeling the variances of estimates in small area estimation**
Presenter: **Sirapat Watakajaturaphon**, University of California, Davis, United States

The estimates of variances in small area estimation play an important role. We discuss common problems of assuming a frequentist framework for modeling the variances of small area estimates and propose a new Bayesian framework to deal with these problems in practice. We propose suitable Markov chain Monte Carlo algorithms and study the theoretical properties of our proposed model. Finally, we implement our model in a real data set.

EO572 Room K0.19 ADVANCES IN BAYESIAN COMPUTATIONS Chair: Riccardo Corradin

E0901: **Improving MCMC convergence diagnostic with a local version of R-hat**
Presenter: **Theo Moins**, Inria, France
Co-authors: Julyan Arbel, Anne Dutfoy, Stephane Girard

Diagnosing the convergence of Markov chain Monte Carlo is crucial and remains an essentially unsolved problem. Among the most popular methods, the potential scale reduction factor, commonly named R-hat, is an indicator that monitors the convergence of output chains to a target distribution, based on a comparison of the between- and within-variances. Several improvements have been suggested since its introduction in the 90s. Here, we aim at better understanding the R-hat behavior by proposing a localized version that focuses on the quantiles of the target distribution. This new version relies on key theoretical properties of the associated population value. It naturally leads to proposing a new indicator R-hat-infinity, which is shown to allow both for localizing the Markov chain Monte Carlo convergence in different quantiles of the target distribution, and at the same time, for handling some convergence issues not detected by other R-hat versions.

E1384: **Neutral to the right posterior computations**
Presenter: **Alan Riva-Palacio**, Universidad Nacional Autonoma de Maxico, Mexico

Computations in neutral to the right models for Bayesian nonparametric survival analysis are discussed. Applications for regression modelling and population comparison based on efficient algorithms for the simulation of posterior survival curves and density estimation for posterior cumulants are presented. In particular, we present a package in the Julia programming language for survival analysis with neutral to the right priors.

E1294: **Robust inference in mixture models maximum mean discrepancy relaxations**
Presenter: **Yordan Raykov**, University of Nottingham, United Kingdom

Bayesian inference delivers principled rules for learning from data and integrating out uncertainty. As data grows through, posterior estimates concentrate around the likelihood, and the robustness properties embedded in the prior diminish. This problem is often approached by different

forms of model-agnostic robust inference frameworks, such as generalised Bayesian inference. We propose a model-specific approach for robust inference of mixture-type densities to any potential likelihood of misspecification; we call this the neighbourhood mixture model. We propose to use the maximum mean discrepancy (MMD) to relax the assumptions of the component parameters(s) from being points to neighbourhoods. The proposed framework is shown to lead to superior maximum-a-posterior point estimates across many practical tasks, such as clustering in the presence of model misspecification, learning the number of mixture components, and clustering of single-cell data in the presence of sequencing depth.

EO579 Room K0.20 RECENT ADVANCES IN CAUSAL MEDIATION ANALYSIS Chair: Linbo Wang

E1955: **Proximal mediation analysis**
Presenter: **Oliver Dukes**, Ghent University, Belgium
Co-authors: AmirEmad Ghassami, Ilya Shpitser, Eric Tchetgen Tchetgen

A common concern when trying to draw causal inferences from observational data is that the measured covariates are insufficiently rich to account for all sources of confounding. In practice, many of the covariates may only be proxies of the latent confounding mechanism. Recent work has shown that in certain settings where the standard 'no unmeasured confounding' assumption fails, proxy variables can be leveraged to identify causal effects. Results currently exist for the total causal effect of an intervention, but little consideration has been given to learning about the direct or indirect pathways of the effect through a mediator variable. We will describe three separate proximal identification results for natural direct and indirect effects in the presence of unmeasured confounding. We will then develop a semiparametric framework for inference on natural (in)direct effects, which leads us to locally efficient, multiply robust estimators.

E1952: **Defining and estimating principal stratum-specific natural mediation effects with semi-competing risks data**
Presenter: **Fei Gao**, Fred Hutchinson Cancer Center, United States
Co-authors: Fan Xia, Kwun Chuen Gary Chan

In many medical studies, an ultimate failure event such as death is likely to be affected by the occurrence and timing of other intermediate clinical events. Both event times are subject to censoring by loss-to-follow-up, but the nonterminal event may further be censored by the occurrence of the primary outcome, but not vice versa. To study the effect of an intervention on both events, the intermediate event may be viewed as a mediator, but the conventional definition of direct and indirect effects is not applicable due to the semi-competing risks data structure. We define three principal strata based on whether the potential intermediate event occurs before the potential failure event, which allows a proper definition of direct and indirect effects in one stratum, whereas total effects are defined for all strata. We discuss the identification conditions for stratum-specific effects, and propose a semiparametric estimator based on a multivariate logistic stratum membership model and within-stratum proportional hazards models for the event times. By treating the unobserved stratum membership as a latent variable, we propose an EM algorithm for computation. We study the asymptotic properties of the estimators by the modern empirical process theory and examine the performance of the estimators in numerical studies.

E1950: **Feature identification on high-dimensional mediators using causal mediation tree model**
Presenter: **Yao Li**, University of Tortonto, Canada

High-dimensional mediation analysis plays an important role in recent biomedical research as a large number of mediators, such as microbiomes, could modulate the effect of exposure to the outcome of interest. Most of the current studies focus on modeling independent mediators. However, these methods do not consider the correlation between the mediators and their non-linear interactive effect. On the other hand, identifying the mediators with significant effects from the high-dimensional mediator space is challenging. We propose an innovative non-parametric approach to build a causal mediation tree to select important mediators and assess their nonlinear effects. The data are recursively partitioned into subpopulations constructed by the mediators with the largest mediation effect. We aim to incorporate this nonlinear relationship into the mediation framework and evaluate the total effect. Simulation studies were conducted to assess the performance of our algorithm under different scenarios of the interactive mediation effects. We applied the method to analyze vaginal microbiome data from the reproductive-age women study. We investigated the causal relationship between ethnic groups and the virginal PH levels mediated by the virginal microbiomes. We identified two important microbiome taxa with strong mediation effects and estimated the total effect of the mediation tree model.

EO632 Room K0.50 DESIGN OF EXPERIMENTS Chair: Peter Goos

E0278: **Hierarchical identifiable saturated models**
Presenter: **Janet Godolphin**, University of Surrey, United Kingdom

The class of Hierarchical Identifiable Saturated models corresponding to a given design is termed the Statistical Fan of the design. The relationship between a design matrix and its statistical fan is investigated with regards to three aspects: (1) The advantages and disadvantages of a Grobner basis approach to finding a Hierarchical Identifiable Saturated model for a given design matrix are discussed; (2) An alternative, practitioner-led, approach to finding a Hierarchical Identifiable Saturated model is outlined; (3) Conditions on design matrices are developed to ensure the inclusion of specific terms in every leaf of the fan. The results are illustrated by examples.

E0929: **Experimental designs to test for heteroscedasticity in a regression model**
Presenter: **Chiara Tommasi**, University of Milan, Italy
Co-authors: Alessandro Lanteri, Jesus Lopez-Fidalgo, Samantha Leorato

The goal is to design an experiment to detect a specific kind of heteroscedasticity in a non-linear Gaussian regression model. To test the homoscedastic case (under the null hypothesis) against local alternatives, a likelihood-based test is usually applied. Suitable design criteria for this task are Ds- and KL-criteria because they are related to the noncentrality parameter of the asymptotic chi-squared distribution of a likelihood-based test. Thus, they maximize the asymptotic power of the test. Specifically, when the variance function depends just on one parameter, these criteria coincide asymptotically, and in particular, the D1-criterion is proportional to the noncentrality parameter. Differently, when the variance function depends on a vector of parameters, the two criteria are not asymptotically equivalent anymore; the KL-optimum design outperforms the Ds-optimal design because it converges to the design that maximizes the noncentrality parameter. A simulation study, concerning the computation of asymptotic and exact powers of the log-likelihood ratio statistic, confirms these theoretical results.

E0948: **Response surface models: To reduce or not to reduce**
Presenter: **Maria Weese**, Miami University, United States
Co-authors: David Edwards, Byran Smucker

In classical response surface methodology, the optimization step uses a small number of important factors. However, in practice, experimenters sometimes fit a second-order model without previous experimentation. In this case, the true model is uncertain, and the full model may overfit. We use an extensive simulation to evaluate several analysis strategies in terms of their optimum locating ability, and use both simulation and published experiments to evaluate their general prediction facility. We consider traditional (reducing via p-values; forward selection), regularization (LASSO; Gauss-LASSO), and Bayesian analysis methods.

EO573 Room S0.03 SPATIAL AND SPATIO-TEMPORAL STATISTICS IN URBAN AND NATURAL CONTEXTS Chair: Paolo Maranzano

E0339: **Local characteristics of functional marked point processes with applications to seismic data**
Presenter: **Nicoletta D Angelo**, Universita degli Studi di Palermo, Italy
Co-authors: Giada Adelfio, Jorge Mateu, Ottmar Ottmar Cronie

A family of local inhomogeneous mark-weighted summary statistics is presented for general marked point processes. These capture various types of local dependence structures depending on the specified involved weight function. We use them to propose a local random labeling test. This procedure enables us to identify points and, thus, regions where the random labeling assumption does not hold, for example, when the (functional) marks are spatially dependent. We further present an application to a seismic point pattern with functional marks provided by seismic waveforms. Indeed, despite the relatively long history of point process theory, few approaches to analyzing spatial point patterns where the features of interest are functions (i.e. curves) rather than qualitative or quantitative variables have been developed. Forest patterns with associated functional data, curves representing the incidence of an epidemic over time, and the evolution of distinct economic parameters such as unemployment and price rates, all for distinct spatial locations, are examples of point patterns with associated functional data.

E0806: **Gaussian process mapping with uncertainty measures for the urban building models**
Presenter: **Qianqian Zou**, Leibniz University Hannover, Germany
Co-authors: Monika Sester

Mapping with probabilistic uncertainty for urban scenes is required in many research domains, such as localization and sensor fusion. Although there are many uncertainty explorations in the pose estimation of an ego-robot with map information, the quality of the reference maps is often neglected. To avoid the potential problems caused by the errors of maps and a lack of uncertainty quantification, an adequate uncertainty measure for the maps is required. The modelling for urban buildings with the implicit surface using Gaussian Process (GP) is proposed to measure the mapping uncertainty in a probabilistic fashion. To reduce the redundant computation for simple planar objects, explicit facets from a Gaussian Mixture Model (GMM) are combined with the GP map while sparse GP techniques are used as well. The proposed method is evaluated on LiDAR point clouds of city buildings collected by a mobile mapping system. Compared to the performances of methods using standard GP, sparse GP and GMM, our method has shown lower RMSE and higher log-likelihood with less computational complexity.

E1268: **Spatial statistical calibration for linear network data: The analysis of traffic volumes**
Presenter: **Andrea Gilardi**, University of Milano - Bicocca, Italy
Co-authors: Riccardo Borgoni , Jorge Mateu

Analyzing traffic volumes at the street network level represents a crucial step to improving transport planning protocols and developing effective road safety interventions. A common practice to estimate traffic figures involve manual counts with ad-hoc cameras or automatic counts with road-fixed sensors (e.g., inductive loops and spirals). Unfortunately, these traditional techniques have several limitations due to their limited spatial coverage and high economic costs of installation and maintenance. For these reasons, in the last years, several authors developed statistical methods to derive counts from traffic information collected using geo-referenced mobile sensors (e.g., smartphones and sat-navs). Mobile sensors have several advantages over traditional instruments, but they underestimate real flows. For these reasons, we developed a spatial statistical calibration technique to combine accurate counts from fixed cameras and extensive GPS mobile data to estimate traffic flows, re-adapting the methodology to the linear network context. We also propose an algorithm that can be used to optimise the size and the spatial allocation of fixed sensors in an urban environment by evaluating their importance for spatial calibration. The suggested methodology is exemplified by considering data collected in the City of Leeds (UK).

EO282 Room S0.11 RECENT ADVANCEMENTS IN POINT PROCESS MODELS Chair: Ganggang Xu

E0631: **Statistical analysis of point patterns on linear networks**
Presenter: **Jesper Moeller**, Aalborg University, Denmark

A point process is a mathematical model for randomly distributed point patterns in a given space. While the mathematical and statistical theory for point processes on one, two, or higher dimensional Euclidean space is fairly well-developed with accompanying user-friendly software for statistical analysis, notably the R package spatstat, the research on point processes defined on more general spaces such as spheres and linear networks is in its infancy. A state-of-the-art review of statistical models, simulation procedures, and methods for estimation and model checking when analysing point patterns observed on linear networks, e.g. crime cases observed on a street network or spine positions on a dendrite network.

E1046: **A $K-$function for inhomogeneous fiber patterns**
Presenter: **Rasmus Waagepetersen**, Aalborg University, Denmark

A $K-$function is presented for assessing second-order properties of inhomogeneous fiber patterns generated by marked point processes. The $K-$function takes into account the geometric features of the fibers, such as tangent directions. The $K-$function requires an estimate of the inhomogeneous density function of the fiber pattern. We introduce parametric estimates for the density function based on parametric models that represent large-scale features of the inhomogeneous fiber pattern. The proposed methodology is applied to simulated fiber patterns as well as a three-dimensional data set of steel fibers in concrete.

E1091: **Group network Hawkes process**
Presenter: **Ganggang Xu**, University of Miami, United States
Co-authors: Guanhua Fang, Haochen Xu, Xuening Zhu, Yongtao Guan

The event occurrences of individuals interacting in a network are studied. To characterize the dynamic interactions among the individuals, we propose a group network Hawkes process (GNHP) model whose network structure is observed and fixed. In particular, we introduce a latent group structure among individuals to account for the heterogeneous user-specific characteristics. A maximum likelihood approach is proposed to cluster individuals in the network and estimate model parameters simultaneously. A fast EM algorithm is subsequently developed by utilizing the branching representation of the proposed GNHP model. Theoretical properties of the resulting estimators of group memberships and model parameters are investigated under both settings when the number of latent groups G is over-specified or correctly specified. A data-driven criterion that can consistently identify the true G under mild conditions is derived. Extensive simulation studies and an application to a data set collected from Sina Weibo are used to illustrate the effectiveness of the proposed methodology.

EO272 Room S0.12 NEW BAYESIAN APPROACHES FOR VARIABLE SELECTION Chair: Daniel Kowal

E0622: **A Bayesian zero-inflated Dirichlet-multinomial regression model for multivariate compositional count data**
Presenter: **Matthew Koslovsky**, Colorado State University, United States

The Dirichlet-multinomial (DM) distribution plays a fundamental role in modern statistical methodology development and application. Recently, the DM distribution and its variants have been used extensively to model multivariate count data generated by high-throughput sequencing technology in omics research due to its ability to accommodate the compositional structure of the data as well as overdispersion. A major limitation of the DM distribution is that it is unable to handle excess zeros typically found in practice which may bias inference. To fill this gap, we propose a novel Bayesian zero-inflated DM model for multivariate compositional count data with excess zeros. We then extend our approach to regression

settings and embed sparsity-inducing priors to perform variable selection for high-dimensional covariate spaces. Throughout, modeling decisions are made to boost scalability without sacrificing interpretability, imposing limiting assumptions, or relying on approximation techniques. We apply the model to a benchmark human microbiome data set and compare the performance of the proposed method to existing approaches.

E0636: **Going beyond spike-and-slab: Sparse modeling with the L1-ball priors**
Presenter: **Leo Duan**, University of Florida, United States

The l1-regularization is very popular in high dimensional statistics – it changes a combinatorial problem of choosing which subset of the parameters are zero, into a simple continuous optimization. Using a continuous prior concentrated near zero, the Bayesian counterparts are successful in quantifying the uncertainty in the variable selection problems; nevertheless, the lack of exact zeros makes it difficult for broader problems such as change-point detection and rank selection. Inspired by the duality of the l1-regularization as a constraint onto an l1-ball, we propose a new prior by projecting a continuous distribution onto the l1-ball. This creates a positive probability on the ball boundary, which contains both continuous elements and exact zeros. Unlike the spike-and-slab prior, this l1-ball projection is continuous and differentiable almost surely, making the posterior estimation amenable to the Hamiltonian Monte Carlo algorithm. We examine the properties, such as the volume change due to the projection, the connection to the combinatorial prior, and the minimax concentration rate in the linear problem. We demonstrate the usefulness of exact zeros that simplify combinatorial problems, such as the change-point detection in time series, the dimension selection of the mixture model and the low-rank-plus-sparse change detection in the medical images.

E0662: **Bayesian subset selection and variable importance for interpretable prediction**
Presenter: **Daniel Kowal**, Rice University, United States

Subset selection is a valuable tool for interpretable learning, scientific discovery, and data compression. However, classical subset selection is often avoided due to selection instability, lack of regularization, and difficulties with post-selection inference. We address these challenges from a Bayesian perspective. Given any Bayesian model M, we extract a family of near-optimal subsets of variables for linear prediction. This strategy deemphasizes the role of a single "best" subset and instead advances the broader perspective that often many subsets are highly competitive. The acceptable family of subsets offers a new pathway for model interpretation and is neatly summarized by key members and new (co-) variable importance metrics. More broadly, we apply Bayesian decision analysis to derive the optimal linear coefficients for any subset of variables. These coefficients inherit both regularization and uncertainty quantification via M. For both simulated and real data, the proposed approach exhibits better prediction, interval estimation, and variable selection than competing Bayesian and frequentist selection methods. These tools are applied to a large education dataset with highly correlated covariates. Our analysis provides unique insights into the combination of environmental, socioeconomic, and demographic factors that predict educational outcomes, and identifies over 200 distinct subsets of variables that offer near-optimal out-of-sample predictive accuracy.

EO568 Room Virtual R01 RECENT ADVANCES IN STOCHASTIC MODELS Chair: Anna Panorska

E1995: **What can we learn from the skewed log-logistic model of the epidemic curve?**
Presenter: **Anna Panorska**, University of Nevada, United States
Co-authors: Tomasz Kozubowski, Fares Qeadan, Alina Giri

A skewed log-logistic model is introduced for the epidemic curve and its connection with the corresponding differential equation. The model allows some comparisons between the behavior of an epidemic in different countries or regions. The skewness parameter, which controls the speed of growth and later leveling of the epidemic curve, is shown to correlate closely with the socioeconomic measures for different countries. In addition, we show the possibility of using the skewed model for some comparisons of the effectiveness of the nonpharmaceutical interventions. We present data analyses from several countries around the world.

E1997: **Kaniadakis functions beyond statistical mechanics: Power-law tails, and modified lognormal distribution**
Presenter: **Anastassia Baxevani**, University of Cyprus, Cyprus
Co-authors: Dinissios Hristopoulos

Probabilistic models with flexible tail behavior have important applications in engineering and earth science. We introduce a nonlinear normalizing transformation and its inverse based on the deformed lognormal and exponential functions proposed by Kaniadakis. The deformed exponential transform can be used to generate skewed data from normal variates. We apply this transform to a censored autoregressive model for the generation of precipitation time series. We also highlight the connection between the heavy-tailed k-Weibull distribution and weakest-link scaling theory, which makes the k-Weibull suitable for modeling the mechanical strength distribution of materials. Finally, we introduce the k-lognormal probability distribution and calculate the generalized (power) mean of k-lognormal variables. The k-lognormal distribution is a suitable candidate for the permeability of random porous media. In summary, the k-deformations allow modifying the tails of classical distribution models (e.g., Weibull, lognormal), thus enabling new directions of research in the analysis of spatiotemporal data with skewed distributions.

E1839: **Sample cross-covariance function for testing of bi-dimensional Gaussian processes**
Presenter: **Katarzyna Maraj-Zygmat**, Wroclaw University of Science and Technology, Poland

A novel framework is introduced that allows efficient bi-dimensional Gaussian process discrimination. The underlying test statistic is based on the sample cross-covariance function. There is a lack of methods for these processes. We present the analysis of probabilistic properties of the sample cross-covariance function for selected multivariate Gaussian processes (bi-dimensional Brownian motion, bi-dimensional fractional Brownian motion, and bi-dimensional Ornstein-Uhlenbeck process). Also, the test based on the sample cross-covariance function for multivariate Gaussian processes is presented. The results and properties are checked by Monte Carlo simulation. For completeness, it is also shown how to embed this methodology into bi-dimensional real data analysis.

EO613 Room Virtual R02 COMBINING CLINICAL TRIALS AND OBSERVATIONAL STUDY DATA Chair: Jonathan Schildcrout

E0887: **Methods to improve the efficiency of RCTs using observational studies**
Presenter: **Amir Asiaee**, Vanderbilt University Medical Center, United States

The aim is to develop a framework for precisely estimating the Conditional Average Treatment Effect (CATE), which characterizes treatment effect heterogeneity. Due to cost constraints, RCTs are often small in size and scope and are often significantly underpowered to detect treatment effect heterogeneity; on the other hand, the estimated effect from one population may not be transportable to another. The transportability of RCT findings has been a subject of methodological and practical interest, and much progress has been made. In contrast, the integration of observational studies (OS) in the analysis of RCT for variance reduction is under-explored. The target population of heterogeneous data integration is the RCT, and the goal is to utilize large observational studies, usually from a non-comparable population, to compensate small sample size of RCTs for CATE estimation. We address a few critical theoretical challenges presented in real-world scenarios. First, we study the settings where the set of measured covariates in the RCT and OS are not entirely identical. Further, almost all of the current work in transportability assume that the CATE is invariant across populations. The CATE transportability assumption or its more stringent forms (trial participation ignorability) are cornerstones of current theoretical results. Still, they are rarely satisfied, e.g., where there are systematic, unrecorded differences in those who participate in RCTs versus the OS.

E0844: **Incorporating external data into the analysis of clinical trials via Bayesian additive regression trees**
Presenter: **Tianjian Zhou**, Colorado State University, United States
Co-authors: Yuan Ji

Most clinical trials involve the comparison of a new treatment to a control arm (eg, the standard of care) and the estimation of a treatment effect. External data, including historical clinical trial data and real-world observational data, are commonly available for the control arm. With proper statistical adjustments, borrowing information from external data can potentially reduce the mean squared errors of treatment effect estimates and increase the power of detecting a meaningful treatment effect. We propose to use Bayesian additive regression trees (BART) for incorporating external data into the analysis of clinical trials, with a specific goal of estimating the conditional or population average treatment effect. BART naturally adjusts for patient-level covariates and captures potentially heterogeneous treatment effects across different data sources, achieving flexible borrowing. Simulation studies demonstrate that BART maintains desirable and robust performance across a variety of scenarios and compares favorably to alternatives. We illustrate the proposed method with an acupuncture trial and a colorectal cancer trial.

E0836: **Survival analysis of randomized controlled trials with observational studies**
Presenter: **Xiaofei Wang**, Duke University, United States

In the presence of heterogeneity between the randomized controlled trial (RCT) participants and the target population, evaluating the treatment effect solely based on the RCT often leads to biased quantification of the real-world treatment effect. To address the problem of lack of generalizability for the treatment effect estimated by the RCT sample, we leverage observational studies with large samples that are representative of the target population. The focus is on evaluating treatment effects on survival outcomes for a target population. A broad class of estimands is considered that are functionals of treatment-specific survival functions, including differences in survival probability and restricted mean survival times. We propose a semiparametric estimator through the guidance of the efficient influence function. The proposed estimator is doubly robust in the sense that it is consistent for the target population estimands if either the survival model or the weighting model is correctly specified, and is locally efficient when both are correct. Simulation studies confirm the theoretical properties of the proposed estimator and show it outperforms competitors. We apply the proposed method to estimate the effect of adjuvant chemotherapy on survival in patients with early-stage resected non-small lung cancer.

EO496 Room Virtual R04 STATISTICAL INFERENCE AND EXPLAINABLE MACHINE LEARNING Chair: Ali Shojaie

E0620: **Inference for model-agnostic longitudinal variable importance**
Presenter: **Brian Williamson**, Kaiser Permanente Washington Health Research Institute, United States
Co-authors: Susan Shortreed, Peter Gilbert, Noah Simon, Marco Carone

In many applications, it is of interest to assess the relative contribution of features (or subsets of features) toward the goal of predicting a response; in other words, to gauge the variable importance of features. We will discuss a general framework for nonparametric inference on interpretable algorithm-agnostic variable importance, where variable importance is defined as a population-level contrast in oracle prediction potential between two nested groups of features. In this framework, valid confidence intervals and tests may be constructed, even when machine learning techniques are used. We will further discuss several approaches to summarizing the longitudinal importance of variables and to making inferences from these summaries.

E0991: **Nonparametric methodology for causal inference with continuous exposures**
Presenter: **Aaron Hudson**, Fred Hutchinson Cancer Center, United States
Co-authors: Mark van der Laan

In many scientific studies, it is of interest to assess whether there exists a causal relationship between an exposure variable and an outcome. When the exposure is continuous, a target statistical parameter that is commonly of interest is the dose-response function. Most of the available methodology for statistical inference on the dose-response function requires strong parametric assumptions on the probability distribution. Such parametric assumptions are typically untenable in practice and lead to an invalid inference. It is often preferable to instead use nonparametric methods for inference, which only make mild assumptions about the data-generating mechanism. We propose a nonparametric test of the null hypothesis that the dose-response function is equal to a constant function. We argue that when the null hypothesis holds, the dose-response function has zero variance. Thus, one can test the null hypothesis by assessing whether there is sufficient evidence to claim that the variance is positive. We construct a novel estimator for the variance of the dose-response function, for which we can fully characterize the null limiting distribution and thus perform well-calibrated tests of the null hypothesis. We also present an approach for constructing simultaneous confidence bands for the dose-response function by inverting our proposed hypothesis test.

E1675: **Black box tests for algorithmic stability**
Presenter: **Byol Kim**, University of Washington, United States
Co-authors: Rina Foygel Barber

Algorithmic stability is a concept from learning theory that expresses the degree to which changes to the input data (e.g., removal of a single data point) may affect the outputs of a regression algorithm. Knowing an algorithm's stability properties is often useful for many downstream applications; for example, stability is known to lead to desirable generalization properties and predictive inference guarantees. However, many modern algorithms currently used in practice are too complex for a theoretical analysis of their stability properties, and thus we can only attempt to establish these properties through an empirical exploration of the algorithm's behavior on various data sets. We lay out a formal statistical framework for this kind of black box testing without any assumptions on the algorithm or the data distribution, and establish fundamental bounds on the ability of any black box test to identify algorithmic stability.

EO166 Room Virtual R05 STATISTICS OF EXTREME VALUES Chair: Gilles Stupfler

E0593: **A refined extreme quantiles estimator of Weibull tail-distributions**
Presenter: **Jonathan El Methni**, Universite Paris Cite, France
Co-authors: Stephane Girard

In the case of Weibull tail distributions, the most commonly used methodology for estimating extreme quantiles is based on two estimators: an order statistic to estimate an intermediate quantile and an estimator of the Weibull tail coefficient. The common practice is to select the same intermediate sequence for both estimators. We show how an adapted choice of two different intermediate sequences leads to a reduction of the asymptotic bias associated with the resulting refined estimator. The asymptotic normality of the latter estimator is established, and a data-driven method is introduced for the practical selection of the intermediate sequences. Our approach is compared to various bias-reduced estimators in a simulation study. An illustration of an actuarial real data set is also provided.

E0727: **Partial tail correlation for extremes**
Presenter: **Jeongjin Lee**, University of Namur, Belgium

A method is developed for investigating conditional extremal relationships between pairs of variables. We consider an inner product space constructed from transformed-linear combinations of independent regularly varying random variables. By developing the projection theorem for the inner product space, we derive the concept of partial tail correlation via the projection theorem. We show that the partial tail correlation can be

understood as the inner product of the prediction errors associated with the transformed linear prediction. Similar to Gaussian cases, we connect the partial tail correlation to the inverse of the inner product matrix and show that a zero in this inverse implies a partial tail correlation of zero. We develop a hypothesis test for the partial tail correlation of zero and demonstrate the performance in a simulation study as well as in two applications: extreme river discharges in the upper Danube basin and high nitrogen dioxide levels in Washington DC.

E0938: **Multivariate sparse clustering for extremes**
Presenter: **Nicolas Meyer**, Universite de Montpellier, France
Co-authors: Olivier Wintenberger
Identifying directions where extreme events occur is a major challenge in multivariate extreme value analysis. We use the concept of sparse regular variation introduced in previous work to infer the tail dependence of a random vector X. This approach relies on the Euclidean projection onto the simplex, which better exhibits the sparsity structure of the tail of X than the standard methods. Our procedure, based on a rigorous methodology, aims at capturing clusters of extremal coordinates of X. It also includes the identification of the threshold above which the values taken by X are considered as extreme. We provide an efficient and scalable algorithm called MUSCLE and apply it to numerical examples to highlight the relevance of our findings. Finally, we illustrate our approach with financial return data.

EO616 Room Virtual R06 RECENT DEVELOPMENTS IN NONPARAMETRIC STATISTICS Chair: Shubhadeep Chakraborty

E1693: **A high dimensional dissimilarity measure**
Presenter: **Reza Modarres**, George Washington University, United States
A new dissimilarity measure for high-dimensional, low-sample size settings is proposed to compare high-dimensional probability distributions. The asymptotic behavior of the new dissimilarity index is studied theoretically. Numerical experiments from high dimensional distributions exhibit the usefulness of the method. The eigenvalues of the matrix of dissimilarities for comparing two high-dimensional samples are determined and shown to be related to the asymptotic value of the dissimilarity index. A dissimilarity visualization plot that is useful for the detection of outliers and change points is proposed and utilized to find the change points in S&P500 stock return data.

E1756: **A multivariate permutation test for the analysis of C paired samples in the presence of multiple data types**
Presenter: **Riccardo Ceccato**, University of Padova, Italy
Co-authors: Rosa Arboretti, Elena Barzizza, Luigi Salmaso
The Nonparametric Combination (NPC) is a flexible permutation-based methodology that can be adopted to deal with a wide range of complex problems, including the comparison of two or more populations when a multivariate outcome is observed. We propose on a new NPC-based testing procedure to address a specific multivariate problem in which C 2 paired samples and multiple data types are available. A simulation study is proposed to evaluate the performances of our proposal under several challenging scenarios. A real-data application is also considered. Data were gathered through a questionnaire that was submitted to multiple respondents, asking them to evaluate a product in terms of a certain set of KPIs after multiple time frames. A number of experiments were also conducted and several continuous KPIs were measured after the same time frames. The NPC-based test was therefore adopted to compare the performances of the product across time.

E2038: **Change-point testing of high-dimensional spectral density matrices**
Presenter: **Ansgar Steland**, University Aachen, Germany
The analysis of multivariate time series in the frequency domain may be based on the spectral density matrix in terms of auto- and cross-spectra and has applications in various areas such as finance, brain research or sensor monitoring. Identifying inhomogeneities in the form of significant change-points (breaks), e.g. in coherences, is a relevant statistical issue, often complicated by the fact that additional structure, such as a factor effect, needs to be taken into account. A flexible framework is proposed to analyze high-dimensional nonlinear time series, which is formally based on self-standardized CUSUM statistics based on localized linear combinations of bilinear forms of spectral average statistics calculated from local lag-window spectral estimators, and, more generally, nonlinear functions of the spectral density matrix. In this way, one can easily analyze contrasts between cross-spectra or coherencies, test for a change in a treatment effect in a frequency band or test the equality of spectral density matrices. All asymptotic results are shown under a general nonlinear time series model. A wild bootstrap procedure is proposed to determine critical values. Simulations indicate that the approach performs well in terms of type I and type II error rates. The method is illustrated by analyzing SP500 returns.

EO566 Room Virtual R07 DATA INTEGRATION IN SURVEY SAMPLING Chair: Saumen Mandal

E2024: **Combining surveys in small area estimation using area level model**
Presenter: **Carolina Franco**, NORC at the University of Chicago, United States
For many surveys, researchers, policymakers, and other stakeholders are interested in obtaining estimates for various domains, such as for geographic subdivisions, for demographic groups, or a cross-classification of both. Often, the demand for estimates at a disaggregated level exceeds what the sample size can support when estimation is done by traditional design-based estimation methods. Small area estimation involves exploiting relationships among domains and borrowing strength from multiple sources of information to improve inference relative to direct survey methods. This typically involves the use of models whose success depends heavily on the quality and predictive ability of the sources of information used. One rich source of information is that of other surveys, especially in countries like the United States, where multiple surveys exist that cover related topics. We will provide a review of the topic of combining information from multiple surveys in small area estimation, focusing on area-level models. We will provide practical advice and a technical introduction, and illustrate with applications.

E2023: **Statistical data integration using multilevel models to predict employee compensation**
Presenter: **Andreea Erciulescu**, Westat, United States
Co-authors: Jean Opsomer, Benjamin Schneider
The focus is on the case where two surveys collect data on a common variable, with one survey being much smaller than the other. The smaller survey collects data on an additional variable of interest, related to the common variable collected in the two surveys, and out-of-scope with respect to the larger survey. Estimation of the two related variables is of interest at domains defined at a granular level. We propose a multilevel model for integrating data from the two surveys, by reconciling survey estimates available for the common variable, accounting for the relationship between the two variables, and expanding estimation for the other variable, for all the domains of interest. The model is specified as a hierarchical Bayes model for domain-level survey data, and posterior distributions are constructed for the two variables of interest. A synthetic estimation approach is considered as an alternative to the hierarchical modeling approach. The methodology is applied to wage and benefits estimation using data from the National Compensation Survey and the Occupational Employment Statistics Survey, available from the Bureau of Labor Statistics, Department of Labor, United States.

EO637 Room K2.31 (Nash Lec. Theatre) INFERENCE UNDER HETEROGENEITY Chair: Gourab Mukherjee

E1978: **A burden shared is a burden halved: A fairness-adjusted approach to classification**
Presenter: **Bradley Rava**, University of Sydney, Australia
Fairness in classification is studied, when one wishes to make automated decisions for people from different protected groups. When individuals are

classified, the decision errors can be unfairly concentrated in certain protected groups. We develop a fairness-adjusted selective inference (FASI) framework and data-driven algorithms that achieve statistical parity in the sense that the false selection rate (FSR) is controlled and equalized among protected groups. The FASI algorithm operates by converting the outputs from black-box classifiers to R-values, which are intuitively appealing and easy to compute. Selection rules based on R-values are provably valid for FSR control, and avoid disparate impacts on protected groups. The effectiveness of FASI is demonstrated through both simulated and real data.

E1989: **A new central limit theorem for the augmented IPW estimator: variance inflation, cross-fit covariance and beyond**
Presenter: **Rajarshi Mukherjee**, Harvard T.H. Chan School of Public Health, United States
Co-authors: Kuanhao Jiang, Subhabrata Sen, Pragya Sur

In recent times, inference for the ATE in the presence of high-dimensional covariates has been extensively studied. Among the diverse approaches that have been proposed, augmented inverse propensity weighting (AIPW) with cross-fitting has emerged as a popular choice in practice. We study this cross-fit AIPW estimator under well-specified outcome regression and propensity score models in a high-dimensional regime where the number of features and samples are both large and comparable. Under assumptions on the covariate distribution, we establish a new CLT for the suitably scaled cross-fit AIPW that applies without any sparsity assumptions on the underlying high-dimensional parameters. Our CLT uncovers two crucial phenomena among others: (i) the AIPW exhibits substantial variance inflation that can be precisely quantified in terms of the signal-to-noise ratio and other problem parameters, (ii) the asymptotic covariance between the pre-cross-fit estimates is non-negligible even on the root-n scale. Finally, we complement our theoretical results with simulations that demonstrate both the finite sample efficacy of our CLT and its robustness to our assumptions.

E1991: **A scalable dynamic bayesian mixture model for fine-grained marketing mix analysis of digital coupons**
Presenter: **Gourab Mukherjee**, University of Southern California, United States

A novel dynamic mixture model is developed for analyzing the effects of varied marketing components in a digital promotion campaign that uses online coupons. A key feature of the proposed model is that it segments customers based on their purchase history and provides fine-grained estimates of the heterogeneous effects that marketing mix variables have on the different consumer segments. The proposed model captures not only long-term heterogeneous segments in the customer pools, but also tracks short-term changes in customer engagement through dynamic indices that tabulate stocks of unresponded recent coupons. We conduct Bayesian estimation of the model parameters by using a novel Gibbs algorithm, which is highly scalable due to the usage of Polya-Gamma distributions based data-augmentation strategy in handling Binomial likelihoods of customer responses to promotional coupons. Finally, through a path-algorithm we provide an integrated framework for providing fine-grained analysis of the marketing component effects at various levels of heterogeneity. We establish large-sample properties on the operational characteristics of the developed algorithm. We apply the proposed model to recent consumer response data from the apparel industry and obtain encouraging results.

EO587 Room K2.40 CAUSAL INFERENCE: RECENT CHALLENGES AND DEBATES Chair: Enrico Ripamonti

E1501: **Propensity score analysis: lessons for observational studies from the design and analysis of RCTs**
Presenter: **Peter Austin**, ICES, Canada

I will discuss how the use of propensity score methods should be guided by the design and analysis of a similar randomized controlled trial (RCT).

E1746: **Using design strategies to improve non-experimental study designs**
Presenter: **Elizabeth Stuart**, Johns Hopkins Bloomberg School of Public Health, United States

Many important research questions can only be answered using non-experimental studies. Propensity scores and related methods are a key tool in the design of non-experimental studies – allowing the design of the study to proceed without the use of the outcome data, and with clear diagnostics. An overview of these methods is provided, common misperceptions regarding their use, and insights into recent advances in the field, including strategies for using propensity scores with complex survey data, with covariates, or in longitudinal policy evaluation contexts. Examples of their use will come from public health, including suicide prevention. The discussion will also include areas for potential future work in propensity score methods, especially for use in the behavioral and social sciences.

C0574: **Causal impact of policy measures and behavior on the COVID pandemic in Germany**
Presenter: **Jenny Bethaeuser**, Justus Liebig University Giessen, Germany

Critics protest loudly against restrictions imposed by politicians during the COVID pandemic: Mandatory masks, lockdowns, school and business closures. The aim is to examine (1) the extent to which these policies have indirectly contributed to limiting the number of COVID cases and deaths by forcing people to practice social distancing, and (2) the extent to which people have adjusted their social distancing behavior on their own based on information about national case and fatality numbers and therefore directly limit the number of COVID cases and deaths. The panel analysis at the federal-state level in Germany between 03/2020 and 12/2021 finds that substantial declines in COVID case and death growth rates are attributable to private behavioral response, but policies played an important role as well. A change in policies explains a large fraction of changes in social distancing behavior, why both policies and national information are important determinants of federal COVID cases and deaths. Due to the lack of cross-sectional variation, there is uncertainty about the effect of mask mandate.

EO238 Room K2.41 HIGH-DIMENSIONAL DATA ANALYSIS AND SPECTRAL METHODS (VIRTUAL) Chair: Christopher McKennan

E0241: **Dataset matching and its applications in single-cell multi-omics**
Presenter: **Shuxiao Chen**, University of Pennsylvania, United States
Co-authors: Zongming Ma

One-way matching of a pair of datasets with low-rank signals is studied. Under a stylized model, we first derive information-theoretic limits of matching. We then show that linear assignment with projected data achieves fast rates of convergence and sometimes even minimax rate optimality for this task. We further design a new matching algorithm that accommodates the case where the covariates are only partially aligned. The practical use of the proposed algorithms is illustrated in several single-cell multi-omics data examples, including batch effect removal in sequencing data, integration of proteomics data, and spatial transfer in multiplex imaging data.

E0870: **Learning low-dimensional nonlinear structures from high-dimensional noisy data: An integral operator approach**
Presenter: **Rong Ma**, Stanford University, United States
Co-authors: Xiucai Ding

A kernel-spectral embedding algorithm is proposed for learning low-dimensional nonlinear structures from noisy and high-dimensional observations, where the datasets are assumed to be sampled from a nonlinear manifold model and corrupted by high-dimensional noise. The algorithm employs an adaptive bandwidth selection procedure which does not rely on prior knowledge of the underlying manifold. The obtained low-dimensional embeddings can be further utilized for downstream purposes such as data visualization, clustering and prediction. Our method is theoretically justified and practically interpretable. Specifically, for a general class of kernel functions, we establish the convergence of the final embeddings to their noiseless counterparts when the dimension and the sample size are comparably large, and characterize the effect of the signal-to-noise ratio on the rate of convergence and phase transition. We also prove the convergence of the embeddings to the eigenfunctions of an integral operator defined by the kernel map of some reproducing kernel Hilbert space capturing the underlying nonlinear structures. Our results hold even when the dimension of the manifold grows with the sample size. Numerical simulations and analysis of three real datasets show the

superior empirical performance of the proposed method, compared to many existing methods, on learning various nonlinear manifolds in diverse applications.

E1160: **Inference for heteroskedastic PCA with missing data**
Presenter: **Yuling Yan**, Princeton University, United States
Co-authors: Yuxin Chen, Jianqing Fan

The purpose is to study how to construct confidence regions for principal component analysis (PCA) in high dimension, a vastly under-explored problem. While computing measures of uncertainty for nonlinear/nonconvex estimators is in general difficult in high dimensions, the challenge is further compounded by the prevalent presence of missing data and heteroskedastic noise. We propose a suite of solutions to perform valid inference on the principal subspace based on two estimators: a vanilla SVD-based approach, and a more refined iterative scheme called HeteroPCA. We develop non-asymptotic distributional guarantees for both estimators, and demonstrate how these can be invoked to compute both confidence regions for the principal subspace and entrywise confidence intervals for the spiked covariance matrix. Particularly worth highlighting is the inference procedure built on top of HeteroPCA, which is not only valid but also statistically efficient for broader scenarios (e.g., it covers a wider range of missing rates and signal-to-noise ratios). Our solutions are fully data-driven and adaptive to heteroskedastic random noise, without requiring prior knowledge about the noise levels and noise distributions.

EC808 Room S-1.04 COPULAS Chair: Stefanie Biedermann

E0692: **On copulas constructed with Bernoulli and Coxian-2 distributions**
Presenter: **Christopher Blier-Wong**, Universita Laval, Canada

A new generalized Farlie-Gumbel-Morgenstern copula is constructed that naturally scales to high dimensions. The copula can model moderate positive and negative dependence, can cover different types of asymmetries and admits exact expressions for many quantities of interest, such as measures of association or risk measures in actuarial science. We construct this copula through a stochastic representation based on multivariate Bernoulli random vectors and Coxian-2 distributions. The construction of the copula and the study of its measures of multivariate association and stochastic ordering is addressed. We explain how to sample random vectors in high dimensions. Then, we study the bivariate copula in detail. Finally, we consider subfamilies of the new family of copulas that exhibit specific shapes of dependence.

E0711: **Exchangeable FGM copulas**
Presenter: **Etienne Marceau**, Laval University, Canada
Co-authors: Helene Cossette, Christopher Blier-Wong

Copulas are a powerful tool to model dependence between the components of a random vector. One well-known class of copulas when working in two dimensions is the Farlie-Gumbel-Morgenstern (FGM) copula since their simple analytic shape enables closed-form solutions to many problems in applied probability. However, the classical definition of high-dimensional FGM copula does not enable a straightforward understanding of the effect of the copula parameters on the dependence, nor a geometric understanding of their admissible range. We circumvent this issue by studying the FGM copula from a probabilistic approach based on multivariate Bernoulli distributions. High-dimensional exchangeable FGM copulas are studied, a subclass of FGM copulas. We show that dependence parameters of exchangeable FGM can be expressed as convex hulls of a finite number of extreme points and establish partial orders for different exchangeable FGM copulas (including maximal and minimal dependence). We also leverage the probabilistic interpretation to develop efficient sampling and estimating procedures and provide a simulation study. Throughout, we discover geometric interpretations of the copula parameters that assist one in decoding the dependence of high-dimensional exchangeable FGM copulas.

E1999: **A new family of smooth copulas with arbitrarily irregular densities**
Presenter: **Robert Zimmerman**, University of Toronto, Canada
Co-authors: Michael Lalancette

Copulas are known to satisfy a number of regularity properties, and one might therefore believe that their densities, when they exist, admit a certain degree of regularity themselves. We show that this is not true in general by constructing a broad family of copulas which admit densities that can hardly be considered regular. The copula densities are constructed from arbitrary univariate densities supported on the unit interval, and we show by example that the copula densities can inherit pathological behaviour from the underlying univariate densities. In particular, we construct a nontrivial univariate density which is unbounded in every open set of the unit interval, and show that it induces a copula density which is finite everywhere but unbounded in every neighbourhood of the unit hypercube. Nevertheless, all of our copulas are shown to enjoy attractive smoothness properties.

EC774 Room S0.13 METHODOLOGICAL STATISTICS Chair: Richard Guo

E1748: **Estimating the logarithm of characteristic function and stability parameter for symmetric stable laws**
Presenter: **Annika Krutto**, University of Oslo, Norway

Consider an i.i.d. sample from a symmetric continuous stable distribution with stability parameter and scale parameter. Based on the empirical characteristic function, we prove a uniform large deviation inequality for given preciseness and probability. As an application, we show how it can be used in estimating the unknown stability parameter.

E1902: **Inference for multiple data splitting and exchangeable p-values**
Presenter: **Richard Guo**, University of Cambridge, United Kingdom
Co-authors: Rajen D Shah

Many modern procedures for hypothesis testing employ data splitting. Typically, the dataset is randomly split into two parts: certain complex nuisance functions are estimated from the first part, while the final statistic is computed by evaluating the estimated functions in the second part. Constructed as such, the errors from the two parts are independent, which is often essential to prevent "double dipping", controlling bias and ensuring asymptotic normality of the final statistic. However, such a practice has obvious drawbacks. First, the test is randomized and can yield inconsistent results on two analyses of the same data. Second, using only part of the sample hurts power. One remedy is to combine the statistics or p-values resulting from multiple data splits. We introduce a general method for large-sample inference of the combined statistic under minimal assumptions. We apply our method to a variety of problems: (1) testing conditional mean independence, (2) testing cluster structure in high dimensions and (3) testing no direct effect (Verma constraint) in a sequentially randomized trial. For these problems, our proposal is able to derandomize and improve power. Moreover, in contrast to existing p-value aggregation approaches that can be highly conservative, our method enjoys type-I error control that asymptotically approaches the nominal level.

E0228: **Tight bounds for augmented KL divergence in terms of augmented total variation distance**
Presenter: **Michele Caprio**, University of Pennsylvania, United States

Optimal variational upper and lower bounds for the Kullback-Leibler divergence are provided in terms of the total variation distance between two probability measures defined on two Euclidean spaces having different dimensions.

CO038 Room Virtual R03 REGIME CHANGE MODELING I Chair: Willi Semmler

C1472: **The COVID-19 pandemic and ethical stock markets**
Presenter: **Fredj Jawadi**, University of Lille, France
Co-authors: Nabila Jawadi, Abdoulkarim Idi cheffou

The evolution of stock returns is assessed for two classes of ethical investments: Islamic and sustainable funds, over the period 2001-2021 that covers technological (the dot-com bubble crash in 2000), financial (the 20082009 global financial crisis), and healthcare (the COVID-19 pandemic) shocks. Second, we analyze the dynamics of the financial returns of conventional and ethical markets in the COVID-19 context and model the impact of COVID-19 news on these investments. Hence, we perform different time-series analyses by applying distinct time-varying tests and estimated an ARX-GARCH model to apprehend market reactions to the COVID-19 pandemic. Accordingly, we present two interesting results. First, the COVID-19 pandemic has had a time-varying impact on the stock market, evolving with the progression of the pandemic. Basically, a close to zero effect was observed at the early stage of the pandemic, followed by a negative and significant effect during the first wave between March 2020 and June 2020. However, this has since been attenuated owing to social restriction measures, teleworking, government stimulus policies, and massive vaccine rollouts. Second, among all markets, the Islamic stock market is the most resilient and least impacted by the pandemic, suggesting evidence of a new moral shock.

C1176: **Regime-dependent health care employment dynamics during recessionary periods**
Presenter: **Luiggi Donayre**, University of Minnesota - Duluth, United States
Co-authors: Lacey Loomer

The assumption that recessions are all alike is relaxed in studying whether U.S. healthcare employment is recession-proof. Because healthcare services are inelastic and largely driven by costs, we argue that economic conditions only impact healthcare employment to the extent that they affect the minimum number of necessary healthcare workers or healthcare costs. Specifically, using U.S. monthly data for the January 1991-June 2022 period, we estimate a threshold VAR that explicitly allows for regime-dependent negative demand and negative supply shocks in examining how healthcare employment responds to recessionary periods. When healthcare workers fall below a necessary minimum (or when labor costs are too high) as determined by an endogenously-estimated threshold, we find a large and significant reduction in healthcare employment during demand-induced recessions and a much smaller response during supply-induced recessions. Meanwhile, the response of healthcare employment to negative demand or supply shocks is negligible in the high-worker (high-cost) regime. By identifying the source of recessions and their effects on the minimum number of necessary workers and labor costs, our findings reveal that healthcare employment is not necessarily recession-proof.

C1448: **Econometrics of the distributional effects and effectiveness of carbon taxes**
Presenter: **Andreas Lichtenberger**, The New School, United States

The carbon tax is one of the first pricing tools that was used to prevent climate change by bringing down carbon emissions and driving investment into cleaner options. We study the econometric effects of carbon taxation in two different ways: We jointly research (a) the income distributional effect based on the quasi-experimental introduction of a carbon tax with revenue recycling in British Columbia (BC) vs other Canadian territories with a difference-in-difference approach, and (b) estimate the macroeconomic impacts of the carbon taxes of five European countries (DK, FI, NO, PL, SW) on the economy measured by GDP per capita and the success of the carbon tax on reducing emissions using various specifications of the Vector Autoregression (VAR) and regime switching model (RSM) model. The findings suggest that (a) the BC revenue recycling approaches were not successful in mitigating expenditure increases for the lowest income groups and (b) that the European carbon tax initiatives are effective in reducing carbon emission in the high carbon tax regime in all five countries.

CO454 Room Virtual R08 INFLATION DYNAMICS Chair: Edward Knotek

C1000: **Greater than the sum of its parts: Aggregate vs. aggregated inflation expectations**
Presenter: **Edward Knotek**, Federal Reserve Bank of Cleveland, United States

Using novel survey evidence on consumer inflation expectations for personal consumption expenditure (PCE) categories, we document the paradox that consumers' aggregate inflation expectations differ systematically from an expenditure-weighted sum of their category beliefs, with the "whole" (aggregate inflation expectations) usually greater than the sum of its parts. The inconsistency between aggregate and aggregated inflation expectations rises with subjective uncertainty and is related to socioeconomic characteristics. We show that behavioral weighting schemes that focus on food and gasoline prices or a small number of categories tend to align more closely with aggregate inflation expectations than an expenditure-weighted sum of category beliefs. The choice between aggregate and aggregated inflation expectations has consequential policy implications for monitoring consumer inflation expectations and consequential economic implications: aggregated inflation expectations explain a greater share of planned consumer spending than aggregate inflation expectations, suggesting that the former are more important for consumer decision-making.

C1097: **Central bank communication and house price expectations**
Presenter: **Pei Kuang**, University of Birmingham, United Kingdom

A large online survey is used to study how US consumers' house price expectations respond to communication about interest rate changes. Average house price growth expectations respond little to interest rate hikes, while large heterogeneity exists among households with different mortgage statuses or education levels. Communication about rate hikes combined with a simple explanation of the mortgage rate channel causes large downward revisions to housing price expectations, especially for the less well-educated and non-mortgage payers. Personal experiences are closely associated with the mechanisms recalled by households, which crucially determine the direction and size of their house price predictions.

C1508: **Credibility gains from communicating with the public: Evidence from the ECBs new monetary policy strategy**
Presenter: **Dimitris Georgarakos**, European Central Bank, Germany
Co-authors: Michael Ehrmann, Geoff Kenny

With the rapid increase in euro area inflation, it is ever more important that the ECB maintains its credibility, also among the wider public. Although it is hard to reach out to this group, we show that explaining and communicating key elements of the ECBs new monetary policy strategy can enhance the perceived credibility that price stability will be maintained. In particular, randomised information treatments in the new Consumer Expectations Survey reveal that effective communication about the symmetric inflation target can raise credibility among survey respondents, especially if the stabilising role of monetary policy is also explained. However, the communication of a decision to take better account of climate considerations and a promise to better capture housing costs in inflation measures yield neither marginal credibility gains nor losses.

CO088 Room BH (S) 1.01 Lecture Theatre 1 SUSTAINABLE FINANCE Chair: Kris Boudt

C1621: **Corporate delisting and investors' attention to climate risk**
Presenter: **Stefano Colonnello**, Ca' Foscari University of Venice, Italy
Co-authors: Monica Billio, Ivan Gufler

The focus is on the universe of public companies from the largest European economies to explore the impact of shifting investors' attention towards climate risk in recent years on firms' decision to delist and go private, possibly to dodge increasingly stringent disclosure requirements and scrutiny. In a prima facie aggregate analysis, we show that, despite growing attention to the issue, the presence of listed (relative to non-listed) firms has not declined in several of the sectors most exposed to climate risk. Even at the micro-level, we do not find evidence that firms operating in

climate-sensitive sectors delist at higher rates. We confirm this result by means of a quasi-natural experiment centered on the enhanced disclosure requirements introduced by the EU Non-Financial Reporting Directive of 2014. Whereas firms affected by the directive become more likely to get an ESG rating, they are not more likely to go private.

C1742: **Surfing the green wave**

Presenter: **Carmelo Latino**, Leibniz Institute for Financial Research SAFE; Ca Foscari University of Venice, Germany

A quasi-natural experiment is exploited to study the impact of greenwashing on stock prices. More specifically, we investigate a new channel whereby adopting a new green name (i.e. a name that evokes green and sustainability feelings) could influence investor behavior even in the case of companies that are not related to green activities. Results show that greenwashing has a short-term significant positive impact on companies' value. Cumulative abnormal returns for non-green companies are three times higher than cumulative abnormal returns of green firms over a 10-days window around the announcement date of the name change. However, this effect is transitory and vanishes as soon as investors realize the inconsistency of the name change. In the long term, greenwashing harms companies' value. Results suggest the presence of a green mania in the market with investors eager to buy anything green and some companies keen to surf the green wave.

C1895: **Does credit risk reflect climate transition risk: Evidence from the CDS market for the utility sector**

Presenter: **Michele Costola**, Ca' Foscari University of Venice, Italy

Co-authors: Stefano Battiston

The low-carbon transition can only be achieved if firms reallocate CAPEX to low-carbon technology. Hence transition risk is the risk arising from the mismanagement of the technological shift. Finance acknowledges climate risk, yet financial investments into high-carbon assets have not decreased, and no substantial differential in risk indicators is reported. We aim to test if the technological profile of firms with respect to the energy transition is reflected in their Credit Default Swap (CDS) spreads. We consider the utility electricity sector, where technologies are readily observable. Specifically, we proxy the technology profile of the percentage of capacity in electricity generation from fossil versus renewable sources and control for usual credit risk drivers. Preliminary findings show that the technology profile has no impact on the CDS spreads globally. However, we find evidence that after the Paris Agreement, high fossil (renewable) utility firms in EU27 are associated with higher (lower) CDS spreads.

CO034 Room BH (SE) 1.01 ADVANCES IN TIME SERIES ECONOMETRICS Chair: Martin Wagner

C0519: **A fixed-b perspective on the Phillips-Ouliaris non-cointegration Z-tests**

Presenter: **Sebastian Veldhuis**, University of Klagenfurt, Austria

Co-authors: Martin Wagner

Fixed-b asymptotic theory is extended to the nonparametric Phillips-Ouliaris (PO) non-cointegration Z-tests. We show that the fixed-b limits depend on nuisance parameters in a complicated way. These non-pivotal limits provide an alternative theoretical explanation for the well-known finite-sample problems of the PO Z-tests. Based on these results, we introduce modified PO Z-tests that allow for pivotal fixed-b inference. We consider the asymptotic behavior under the spurious regression null, under local (near-integrated) and fixed alternatives. The performance of the original and modified PO Z-tests is compared by means of local asymptotic power and a finite-sample simulation study.

C1418: **Sources and channels of nonlinearities and instabilities of the Phillips curve in the Euro area and its member states**

Presenter: **Martin Wagner**, University of Klagenfurt, Bank of Slovenia and Institute for Advanced Studies, Vienna, Austria

Co-authors: Karsten Reichold, Milan Damjanovic, Marija Drenkovska

Evidence is presented for sources and channels of nonlinearities and instabilities of the new Keynesian Phillips curve (NKPC) for the euro area and all but four member states over the last two decades prior to the COVID-19 crisis. The approach rests upon misspecification testing using auxiliary regressions based on the standard open-economy hybrid NKPC. Using a large number of specifications, this approach allows to systematically, i. e., based on a literature review, disentangle the evidence for nonlinearities and instabilities of the NKPC according to sources and channels. There is substantial evidence for nonlinearities and instabilities for the euro area and most considered member states. The relatively most important channels of nonlinearities and instabilities are similar across countries, whereas the relatively most important sources differ across countries. The results strongly indicate the need for considering nonlinear NKPC relationships in empirical analyses and also point towards potentially useful nonlinear specifications.

C1654: **Mixed-frequency dynamic factor models**

Presenter: **Leopold Soegner**, Institute for Advanced Studies, Austria

Co-authors: Manfred Deistler, Philipp Gersing, Christoph Rust

Representation and estimation theory is provided for Generalised Dynamic Factor Models (GDFM) with mixed frequency data. We suppose a GDFM for the underlying high-frequency processes where the spectrum of the common component is rational. We look at the structure of the blocked process running on the slow frequency sampling rate containing all observable outputs. With this approach, we aim to build "information efficient" methods for denoising, parameter estimation and factor extraction with observations under mixed sampling frequency. We prove that the blocked process is again a GDFM with a rational spectrum in the common component and define a canonical state space representation for the blocked high-frequency common component.

CO737 Room BH (SE) 1.05 ECONOMETRIC ANALYSIS OF FINANCIAL INSTITUTIONS Chair: Ekaterina Kazak

C0218: **Evaluating hedge fund performance when models are misspecified**

Presenter: **Olivier Scaillet**, University of Geneva and Swiss Finance Institute, Switzerland

Evaluating the performance of hedge funds is challenging because any benchmark model is unlikely to capture their numerous strategies. To assess the level of misspecification among models, we develop a formal comparison approach. This comparison sharpens performance evaluation by improving the separation between pure alphas and factor premium exposures. We find that using the standard benchmark models is problematic because they deliver the same performance as the simplest possible benchmark, the CAPM. In contrast, a parsimonious model based on economically-motivated factors (including carry, time-series momentum, and variance) tracks some alternative hedge fund strategies and achieves a sizable performance reduction relative to the CAPM.

C1227: **Activist hedge fund performance**

Presenter: **Juha Joenvaara**, Aalto University, Finland

Co-authors: Christian Lundblad, Philip Howard, Greg Brown

Using a novel data set containing almost all activist hedge funds, the following questions are posed: Do activist hedge funds create value and/or risk for investors? What are the risk-reward characteristics of activist hedge funds? Who reaps the rewards from activist campaigns: shareholders, fund managers, and/or fund investors? We contribute to the existing literature by examining the effects of activism from investors' perspectives, not the target companies or their shareholders and creditors. Although the literature has documented the positive short-term and long-term effects related to firm performance and the real economy, it is not clear whether investors providing funding to activist hedge funds earn excess returns or whether fund managers extract all economic rents.

C1728: **Reverse stress testing and multivariate extremes**
Presenter: **Natalia Nolde**, The University of British Columbia, Canada

Reverse stress testing of a financial portfolio aims to identify scenarios for risk factors that lead to a specified adverse portfolio outcome, typically a portfolio loss of a given magnitude. The stress scenarios of interest naturally need to be probable yet extreme. In order to capture movements of risk factors that result in large portfolio losses, we propose a method to estimate stress scenarios using extrapolation based on techniques from multivariate extreme value theory. Such a method effectively addresses data scarcity in the joint tail regions while allowing for more flexible model assumptions focused on extremes. We study the asymptotic behaviour of the proposed estimator, investigate its finite-sample performance in simulation studies and apply it to real-life financial portfolios in a case study.

CO144 Room BH (SE) 1.06 HIGH-DIMENSIONAL TIME SERIES ANALYSIS AND APPLICATIONS Chair: Tommaso Proietti

C1052: **Band-pass filtering with high-dimensional time series**
Presenter: **Alessandro Giovannelli**, University of L'Aquila, Italy
Co-authors: Marco Lippi, Tommaso Proietti

The focus is on the construction of a synthetic indicator of economic growth obtained by projecting a measure of aggregate economic activity, such as gross domestic product (GDP), onto high-frequency smooth principal components representative of the medium-to-long-run component of growth in a high-dimensional time series. The smooth principal components result from applying a suitable cross-sectional filter. The result is a monthly nowcast of the medium-to-long-run component of GDP growth. After discussing the theoretical properties of the indicator, we deal with the assessment of its reliability and predictive validity with reference to US data.

C1358: **An analysis of CO2 emissions in Spain using many macroeconomic predictors**
Presenter: **Pilar Poncela**, Universidad Autonoma de Madrid, Spain
Co-authors: Esther Ruiz, Aranzazu de Juan

The aim is to analyze how the (macro) economic activity in Spain affects CO2 emissions. To do so, we build an extensive database of macroeconomic predictors and proceed in two ways: on the one hand, we use variable selection techniques to find a small set of economic predictors with a higher correlation with CO2 emissions and analyze the resulting relations. On the other hand, we build a dynamic factor model with a large dataset of macroeconomic predictors. The extracted common factors capture macroeconomic conditions in Spain. In the second step, we check the relation of the common factors with CO2 emissions. We also estimate joint models that incorporate both the common factors and the individual predictors as explanatory variables of CO2 emissions. Our results indicate that private consumption, industrial production and maritime transport are the most significant variables in order to explain C02 emissions. Once these individual predictors are considered, the information contained in the macroeconomic data set only has negligible explanatory power for emissions. Finally, we generate 1-step ahead forecasts of C02 emissions and evaluate the out-of-sample performance of different models.

C1484: **The dual U.S. labor market uncovered**
Presenter: **Hie Joo Ahn**, Federal Reserve Board, United States

Aggregate U.S. labor market dynamics are well approximated by a dual labor market supplemented with a third home-production segment. We estimate a Hidden Markov Model, a machine-learning method, to uncover this structure. The different market segments are identified through (in-)equality constraints on labor market transition probabilities. This method yields time series of stocks and flows for the three segments for 1980-2021. Primary sector workers, who make up around 55 percent of the population, are almost always employed and rarely experience unemployment. The secondary sector, which constitutes 14 percent of the population absorbs most of the short-run fluctuations in the labor market, both at seasonal and business cycle frequencies. Workers in this segment experience 6 times higher turnover rates than those in the primary tier and are 10 times more likely to be unemployed than their primary counterparts. The tertiary segment consists of workers who infrequently participate in the labor market but nevertheless experience unemployment when they try to enter the labor force. While we find that young workers, racial minorities, and workers with lower educational attainment are more likely to belong to the secondary sector, the bulk of labor market segment variation across individuals cannot be explained by observables. Our findings imply that aggregate stabilization policies, such as monetary policy, predominantly work through the small but turbulent secondary market.

CO731 Room BH (SE) 2.09 MACHINE LEARNING MEETS ECONOMETRICS Chair: Edvard Bakhitov

C1278: **Causal inference with proxy controls in high-dimensional linear models**
Presenter: **Ben Deaner**, UCL, United Kingdom

Recent literature considers causal inference using two vectors of noisy proxies for unobserved confounding factors. We consider linear models in which the vectors of proxies are potentially high-dimensional, and there may be many unobserved confounders. A key insight is that if each group of proxies is strictly larger than the number of confounding factors, this implies rank restrictions on matrices of nuisance parameters. We can exploit the rank-restriction to reduce the number of free parameters to be estimated. The number of unobserved confounders is not known a prior, but we show that it is identified, and we apply penalization methods to adapt to this quantity. We develop doubly-robust estimation and inference methods. We examine the asymptotic properties of these techniques and provide simulation evidence that they are effective.

C1395: **Robust inference for the variance of the CATE in randomized experiments using machine learning**
Presenter: **Alejandro Sanchez Becerra**, Emory University, United States

Machine learning tools are increasingly used to estimate the conditional average treatment effect function (CATE). We propose an efficient estimator of its unconditional variance, the VCATE, which measures the overall effect heterogeneity explained by covariates. First, we introduce a novel way of interpreting VCATE as a bound for the gains of policy learning. We show that the incremental welfare of policies targeted using CATE over policies without targeting has a sharp upper bound equal to root-VCATE/2. Second, we propose novel adaptive confidence intervals (CIs). We start by showing that off-the-shelf CIs with normal critical values and standard errors derived from the efficient influence function have issues covering VCATE. We explain the boundary problem that arises when the treatment effects are homogeneous; this manifests even for simple two-step estimators of VCATE assuming a linear CATE. For the regression case, we prove that the second-step limiting distribution is a generalized chi-square and construct CIs with exact coverage. We then show how to (i) extend these intervals to a class of efficient, debiased, machine learning estimators with a regression step, and (ii) construct variational extensions with conservative coverage that account for sample splitting uncertainty. We document excellent performance in simulations using LASSO.

C1547: **Asymmetric autoencoders for factor-based covariance matrix estimation**
Presenter: **Kevin Huynh**, University of Basel, Switzerland
Co-authors: Gregor Lenhard

Estimating high dimensional covariance matrices for portfolio optimization is challenging because the number of parameters to be estimated grows quadratically in the number of assets. When the matrix dimension exceeds the sample size, the sample covariance matrix becomes singular. A possible solution is to impose a (latent) factor structure for the cross-section of asset returns as in the popular capital asset pricing model. Recent research suggests dimension reduction techniques to estimate the factors in a data-driven fashion. An asymmetric autoencoder neural network-based estimator is presented that incorporates the factor structure in its architecture and jointly estimates the factors and their loadings. The method is tested against well-established alternatives from the literature in an empirical experiment using stock returns of the past five decades. Results

show that the proposed estimator is very competitive across different scenarios. The estimated loadings further reveal that the constructed factors are related to the stocks' sector classification.

CO140 Room BH (SE) 2.10 DEVELOPMENTS IN RISKY ASSET RETURNS DECOMPOSITION METHODS Chair: Mohammad Jahan-Parvar

C1408: **Asset valuations and distributions of returns: Cross-country perspective**
Presenter: **Yuriy Kitsul**, Federal Reserve Board, United States
Co-authors: Stephanie Curcuru

The aim is to investigate the relationship between asset valuations and the probability distributions of future asset price moves in domestic and foreign markets. We find that (1) elevated (subdued) valuations help explain probabilities (and in some cases magnitudes) of large equity price declines (increases) over medium term, (2) U.S. price-earnings (P-E) ratio carries incremental predictive information for explaining probabilities of large foreign equity price changes even after conditioning on foreign countries' own P-E ratios, and (3) U.S. P-E ratio appears to be at least as informative as the first principal component of P-Es of all the considered countries. The last two results echo the findings in the financial cycle literature on the global importance of financial conditions in the United States.

C0172: **A stock return decomposition using observables**
Presenter: **Benjamin Knox**, Federal Reserve Board, United States

A method is proposed to decompose realized stock returns period by period. First, we argue that one can directly estimate expected stock returns from securities available in modern financial markets (using the real yield curve and a given equity risk premium). Second, we derive a return decomposition which is based on stock price elasticities with respect to expected returns and expected dividends, where elasticities can be calculated from dividend futures. An application to the COVID crisis in 2020 reveals that risk premium changes drove much of the crash and rebound in the S&P500 while a fall in long-term real yields drove a strong positive return for 2020 as a whole.

C0258: **Policy spillovers and asset prices: Evidence from the United States and euro area**
Presenter: **Mohammad Jahan-Parvar**, Federal Reserve Baord of Governors, United States

Following and expanding previous methods, a scheme is implemented to identify economic shocks from exchange rates, stock returns, sovereign yield changes, and inflation compensation changes at a daily frequency. We consider growth, monetary policy, hedging premium, common premium, and inflation shocks. This method combines the finance perspective following previous work that studies cash-flow and discount-rate news as drivers of asset prices with the macro view that focuses on structural disturbances.

CO458 Room BH (SE) 2.12 ECONOMETRICS OF ART MARKETS Chair: Douglas Hodgson

C0185: **Creative innovation in golf course architecture, retrospective judgments of quality, and magazine golf course rankings**
Presenter: **Douglas Hodgson**, UQAM, Canada

In creative production, there exists a competitive struggle to persuade the field to value the new work. This is a challenge as cultural innovators do not produce in response to existing demand, but must create new demand for what is supplied. A common tactic is to demarcate the new style by appealing to the virtues of an earlier style of classic status and lacking the corruption of the prevailing style. The proposed innovation is purported, by the creators themselves or the critics who champion them, to renew the classic principles of the historical style, and in their polemic, the new breed of creators attempt to persuade the field to evaluate the historical style, and thus by affinity the new style, to the detriment of the prevailing style, which must be devalued. In golf course architecture, the stylistic revolution unleashed in the early 1990s was accompanied by the new architects and their critics extolling the pre-war Golden Age of architecture to the detriment of what they called the Dark Ages of post-war architecture. We measure the effect of this polemic on the field's overall judgment through an empirical analysis of 30 years of widely read and discussed biannual rankings of the 100 Greatest Courses in the United States as assembled and published by the major magazines Golf Digest and Golf Magazine, and find significant evidence that the rankings evolved during this period in favour of pre-war courses as opposed to post-war courses built before 1990.

C0212: **Public support for performing arts: Efficiency and productivity gains in eleven European countries**
Presenter: **Concetta Castiglione**, University of Calabria, Italy
Co-authors: Davide Infante, Marta Zieba

The importance of public cultural expenditure for the efficiency and productivity of performing arts (PA) firms is investigated. To this aim, we estimate a translog production function using the stochastic frontier approach, and we obtain the estimates of both technical efficiency and its determinants for the PA firms in eleven EU countries from 2009 to 2017. The large panel data set enables the application of robust true random-effects SFA techniques, which control for noise, unobserved firms heterogeneity and endogeneity of the inputs. The empirical results demonstrate that PA firms are technically inefficient, implying that the investigated firms could increase their artistic output between 32 and 42 percent and that decreasing returns to scale are prevalent, due to the presence of too many micro and large-scale firms in the European PA sectors. The results also show that technological progress is not present for the PA firms in the EU eleven countries. However, in contrast to previous research, we demonstrate that the total factor productivity (TFP) increased in the EU PA firms over the examined time period, which is the result of the positive technical efficiency change, as opposed to scale efficiency or technological change. We find that, contrary to the common wisdom on its negative effects on firm efficiency, public spending on culture increases the efficiency of PA firms.

C0267: **The Economic valuation of the cultural services delivered by Belsay Hall: a choice modelling study**
Presenter: **Brenda Dorpalen**, Environment Agency, United Kingdom
Co-authors: Thomas Colwill

The approach to conservation practices of cultural heritage is ever-changing. Academics and policymakers have shifted from viewing the intrinsic value of cultural heritage to considering its social and environmental spill-over effects and the contribution it makes to inclusive and sustainable development. However, to what extent do these professional discourses on the value of heritage reflect public preferences? We put forward a choice experiment using data collected at Belsay Hall, a heritage site in North East England. We contribute to the knowledge of the economic valuation of cultural services by evaluating the performance of conservation, educational and networking services, and facilities, including those targeted towards adults and children. Several econometric specifications are tested. We find a strong public preference for conservation and the intrinsic value of heritage.

CC799 Room BH (S) 2.03 ELECTRICITY MARKETS Chair: Gabriele Torri

C0435: **Risk measures forecasting for diversification of trading in electricity markets**
Presenter: **Andrzej Puc**, Wroclaw University of Science and Technology, Poland
Co-authors: Joanna Janczura

The price risk related to trading in electricity markets has increased significantly in recent years, due to the ongoing market liberalization and the growing renewable energy sources production. The limited storability makes the prices much more volatile than in any other market, and they are characterized by unique features like spikes and the possibility of negative values. We derive and evaluate probabilistic forecasts of electricity prices using the ARX-GARCH model combined with the variance-stabilizing transformation using the inverse hyperbolic sine function. Based

on these forecasts, we calculate predictions of different risk and profit measures taking into account a possible split of the traded energy among markets. Next, we propose a short-term risk management strategy for an electricity supplier/consumer, that utilizes diversification of the markets for electricity trade. Strategies aiming at risk minimization, profit maximization or finding optimal trade-off between risk and return are applied to the German and Polish electricity markets. The obtained results show that diversifying the markets at which electricity is traded leads to improvement in the objective strategy.

C1054: **The coverage probability of forecast intervals in the presence of unpredictable and predictable spikes**
Presenter: **Samaneh Sheybanivaziri**, Norwegian School of Economics, Norway
Co-authors: Jonas Andersson

Methods to compute forecast intervals for electricity price forecasts are systematically compared. In particular, we investigate the complication caused by price spikes. The electricity prices are modeled with a mixture model with two regimes, one for regular prices and one for spikes; a specification that we argue captures the most essential features of distributional and temporal properties of electricity prices. Our first findings are that, not surprisingly, fitting a model accounting for the possibility of spikes helps in getting a correct coverage probability. Also, using simple models, e.g., ARMA models, in combination with a non-parametric bootstrap approach, often give coverage probabilities close to the nominal levels. Ignoring the spikes is often not that consequential for the coverage probabilities for nominal levels 95% and 99%. 80% prediction intervals are conservative, i.e., have a coverage probability well above 80%. While giving an approximately correct coverage probability, simple ARMA models yield substantially wider prediction intervals than the correctly specified mixture model.

C0336: **Understanding volatility spillovers between European electricity spot markets**
Presenter: **Ainhoa Zarraga**, University of the Basque Country, Spain
Co-authors: Evelyn Lizeth Chanatasig, Aitor Ciarreta, Cristina Pizarro-Irizar

The process towards integration of European electricity markets continues despite many obstacles in the path. The benefits of further market integration are not being accomplished due to uncompleted market coupling of national markets, insufficient interconnections and a lack of harmonization of national regulatory policies. A significant feature of electricity prices is that they are significantly more volatile than other comparable financial or commodity markets. Our study examines volatility spillover effects across markets in a model that includes covariances with the aim of achieving a better understanding of the transmission of risks. We conduct both a static and dynamic analysis of aggregated spillover effects as well as their directional decomposition between markets. We find that there are significant spillovers across markets. We also relate the dynamic spillover patterns to specific market events, such as the creation of the single intraday continuous market. Our findings provide useful information for market stakeholders to understand the expected outcomes of further integration initiatives.

CC764 Room BH (SE) 2.05 MACROECONOMETRICS I Chair: Mikkel Plagborg-Moller

C0499: **Backward error and condition number analysis of linear DSGE solutions**
Presenter: **Alexander Meyer-Gohde**, Goethe-University Frankfurt, Germany

The aim is to develop and implement a backward and forward error analysis of the numerical reliability of the solutions of linear dynamic stochastic general equilibrium (DSGE) models. Comparing seven different solution methods from the literature, we demonstrate an economically significant loss of accuracy in the standard, generalized Schur (or QZ) decomposition-based solutions methods resulting from large backward errors in solving the associated matrix quadratic problem. This is illustrated in two production-based asset pricing models, a simple model of external habits with a readily available symbolic solution and a model that lacks such a symbolic solution - QZ-based numerical solutions miss the equity premium by up to several annualized percentage points for parameterizations that either match the chosen calibration targets or are nearby to the parameterization in the literature. While the numerical solution methods from the literature failed to give any indication of these potential errors, easily implementable backwards-error metrics and condition numbers are shown to warn of such potential inaccuracies successfully. The analysis is then performed for a database of roughly 100 DSGE models from the literature and a large set of draws from a given model. While economically relevant errors do not appear pervasive from these latter applications, accuracies that differ by several orders of magnitude persist.

C1815: **Testing instrument validity in proxy-SVARs**
Presenter: **Luca Fanelli**, University of Bologna, Italy
Co-authors: Giuseppe Cavaliere, Giovanni Angelini

A pre-test of instrument validity in proxy-SVARs (SVAR-IVs) is designed based on bootstrap resampling, free from the usual pre-testing issues. The null hypothesis is that (i) the proxies used for the instrumented structural shocks identify the proxy-SVAR, hence are 'strong' (relevance condition), and that (ii) the proxies are uncorrelated with the non-instrumented structural shocks (exogeneity condition). The tests for (i)-(ii) are run sequentially: a bootstrap test for relevance is computed and, conditional on not rejecting the null, a bootstrap test for the exogeneity condition is computed without relying on additional external information other than the proxies used for the shocks of interest. Notably, the outcome of the bootstrap pre-test of relevance does not affect asymptotically the bootstrap test of exogeneity and both do not affect post-test inference. A test for the exogeneity condition (ii) is also designed for cases where the test of relevance (i) rejects the null and the inference in the proxy-SVAR is carried out by weak-instrument robust methods.

C0317: **Simultaneous inference for generalized impulse responses in VAR Models**
Presenter: **Endong Wang**, McGill University, Canada
Co-authors: Jean-Marie Dufour

In macroeconomics, testing the non-causality hypothesis at multiple horizons is often crucial. It has been demonstrated that the test can be simply undertaken using linear regression through the coefficients in Vector Autoregression (VAR) at various horizon, named as generalized impulse responses. It is shown that lag-augmented Vector Autoregression (VAR) at multiple horizons can provide unit-roots robust inference and a simple formula for covariance matrix. The inference is derived by reparameterizing the model with real VAR parameters, which eliminates the concerns of unit roots and serially correlated residuals. The research results provide a theoretical foundation that conservative lag length selection is preferable in terms of preserving asymptotic efficiency and obtaining more accurate asymptotic variance estimates. In practice, the reparameterized model with VAR parameter estimates, analogous to using estimated residuals as regressors, can estimate the usual impulse responses jointly with a non-degenerated distribution. The simulation study illustrates that our estimates could save efficiency. The inference result for structural impulse responses with recursive identification is also presented. Lastly, an empirical application of generalized impulse responses is implemented to analyze the response of the macroeconomic variables to a positive shock on economic policy uncertainty, as measured by the Economic Policy Uncertainty (EPU) index.

Sunday 18.12.2022 08:15 - 09:55 Parallel Session G – CFE-CMStatistics

EO595 Room S-2.25 RECENT DEVELOPMENTS IN LEARNING THEORY Chair: Anirbit Mukherjee

E0752: **The difficulty of computing stable and accurate neural networks: On the barriers of deep learning & Smale 18th problem**
Presenter: **Matthew Colbrook**, University of Cambridge, United Kingdom
Co-authors: Vegard Antun, Anders Hansen

Deep learning (DL) has had unprecedented success and is now rapidly entering scientific computing (SC). However, DL suffers from a universal phenomenon: instability, despite universal approximation results that often guarantee the existence of stable and accurate neural networks (NNs). We show the following paradox. There are well-conditioned problems in SC where one can prove the existence of NNs with great approximation qualities; however, there does not exist any algorithm that can train such a NN. For any positive integers $n > 2$ and M, there are cases where simultaneously: (a) no algorithm can train a NN correct to n digits, (b) there exists an algorithm that trains a NN with $n-1$ correct digits, but any such algorithm needs arbitrarily many training data, (c) there exists an algorithm that trains a NN with $n-2$ correct digits using no more than M training samples. These results provide basic foundations for Smale's 18th problem and imply a classification theory describing conditions under which (stable) NNs with a given accuracy can be trained. We begin this theory by initiating a unified theory for compressed sensing and DL, leading to sufficient conditions for the existence of training algorithms for stable NNs in inverse problems. We introduce FIRENETs, which we prove and numerically verify are stable. FIRENETs only require $O(\log(1/\varepsilon))$ hidden layers for an ε-accurate solution to the inverse problem.

E0764: **Implicit Bias of the Step Size in over-parameterized models**
Presenter: **Daniel Soudry**, Technion, Israel

Focusing on diagonal linear networks and shallow neural networks as a model for understanding the implicit bias in underdetermined models, we show how the gradient descent step size can have a large qualitative effect on the implicit bias toward "smoother" predictors, and thus on generalization ability. In particular, in diagonal linear networks, we show how using large step size for non-centered data can change the implicit bias from a "kernel" type behavior to a "rich" (sparsity-inducing) regime — even when gradient flow, studied in previous works, would not escape the "kernel" regime. We do so by using dynamic stability, proving that convergence to dynamically stable global minima entails a bound on some weighted L1-norm of the linear predictor, i.e. a "rich" regime. We prove this leads to good generalization in a sparse regression setting.

E1261: **The challenges of training models with differential privacy**
Presenter: **Soham De**, DeepMind, United Kingdom

Differential Privacy (DP) provides a formal privacy guarantee preventing adversaries with access to a machine learning model from extracting information about individual training points. Differentially Private Stochastic Gradient Descent (DP-SGD), the most popular DP training method for deep learning, realizes this protection by injecting noise during training. However, previous works have found that DP-SGD often leads to a significant degradation in performance on standard benchmarks. We will first describe the challenges of achieving good performance under differential privacy guarantees. We will then discuss some recent work that shows that using simple techniques to improve signal propagation and convergence on deep networks can significantly improve the performance of DP-SGD on large models.

E1403: **From differential learning to diffusion models**
Presenter: **Samuel Kaski**, The University of Manchester, United Kingdom
Co-authors: Markus Heinonen

Deep learning models construct successive representations across layers, and many DL models have recently been shown to converge to various stochastic processes (such as SDEs). This new perspective has fueled dramatic developments in machine learning, where generative models such as DALLE or Imagen exhibit jaw-dropping performance. We will survey some of the foundational principles of such models, and demonstrate how such principles can be used to re-think supervised learning as well.

EO704 Room S-1.01 ADVANCES IN HILBERT STATISTICS AND APPLICATION TO DISTRIBUTIONAL DATA Chair: Enea Bongiorno

E0924: **Generalization of cell-wise outlier identification for probability density functions**
Presenter: **Ivana Pavlu**, Palacky University Olomouc, Czech Republic
Co-authors: Karel Hron

When performing any data analysis, it is important to acknowledge the possible existence of outliers in the dataset. Observations deviating from the model assumptions may severely affect the results and their interpretability, making outlier detection an important step of the analysis. With multivariate data, it can be convenient to observe the outliers at the cellwise level, this means looking for deviations in individual cells of a data matrix. One well-established possibility for a comprehensive outlier identification is the use of Deviating Data Cells (DDC) algorithm which enables the search for both row-wise and cellwise outliers. The idea of the DDC algorithm is applied to a spline representation of probability density functions (PDFs), hence extending multivariate outlier detection to the functional distributional case. Using the information contained in the spline coefficients, it is possible to highlight parts of PDFs where their behavior deviates from the common trend. Theoretical developments will be demonstrated with a dataset containing particle size distributions from a geological survey in the Czech Republic.

E1370: **Additive density-on-scalar regression in Bayes Hilbert spaces with an application to gender economics**
Presenter: **Eva-Maria Maier**, Humboldt University of Berlin, Germany
Co-authors: Sonja Greven, Almond Stoecker, Bernd Fitzenberger

Functional additive regression models are presented for probability density functions as responses with scalar covariates. To respect the special properties of densities, i.e., nonnegativity and integration to one, we formulate the model for densities in a Bayes Hilbert space with respect to an arbitrary finite measure. Besides continuous and discrete densities, this also allows for, e.g., mixed densities, having discrete point masses at some points of an interval. For estimation, we propose a gradient boosting algorithm, which allows for potentially numerous flexible covariate effects and model selection. We apply our approach to a motivating data set from the German Socio-Economic Panel Study (SOEP) on the distribution of the woman's share in a couple's total labor income - an example for mixed densities since the woman's income share is a continuous variable having discrete point masses at zero and one for single-earner couples. Our approach assumes the response densities are observed directly. If we have individual-level data, we can apply our approach to estimated densities. Alternatively, we currently work on an approach to model the conditional densities directly from individual observations, which we will sketch in an outlook.

E1646: **Functional factor model for density functions**
Presenter: **Israel Martinez-Hernandez**, Lancaster University, United Kingdom

Air pollutants are the most studied phenomena due to their impact on our daily life. In particular, the study and understanding of different sources of particulate matter (PM). Recent evidence has shown that a mixture of particles from different sources can have different toxicity and health effects. For this reason, particle number size distribution (PNSD) measurements have received much attention and are used to investigate PM sources. Due to the high correlation and a large amount of data, current models are highly computationally demanding or use a simple surrogate model. We propose to use a functional data analysis approach. PNSD can be naturally considered as a sequence of density functions over time. With this motivation, we propose a functional factor model that considers the time dependency and overcomes several challenges the current models face.

We will illustrate our methodology using apportion PNSD measured near London Gatwick Airport (UK). Our model can identify the most common PM sources and provides accurate information on how each source contributes to the total PM.

E1770: **Density regression with functional data analysis**
Presenter: **Stefano Antonio Gattone**, University G. d'Annunzio of Chieti-Pescara, Italy
Co-authors: Tonio Di Battista

Recent technological advances have eased the collection of large amounts of data in many research fields. In this scenario, a useful statistical technique is density estimation, which represents an important information source. One-dimensional density functions represent a special case of functional data subject to the constraints to be non-negative and with a constant integral equal to one. Because of these constraints, density functions do not form a vector space, and a naive application of functional data analysis (FDA) methods may lead to non-valid estimates. To address this issue, two main strategies can be found in the literature. In the first, the probability density functions (pdfs) are mapped into a linear functional space through a suitably chosen transformation. Established methods for Hilbert space-valued data can be applied to the transformed functions, and the results are moved back into the density space by means of the inverse transformation. In the second strategy, pdfs are treated as infinite-dimensional compositional data since they are part of some whole that only carries relative information. By means of a suitable transformation, densities are proposed to be embedded in the Hilbert space of square-integrable functions where standard FDA methodologies can be applied.

EO709 Room S-1.06 MODELING COMPLEX DATA AND INTERACTIONS Chair: Alex Cucco

E0305: **Flexible species distributions modelling for spatiotemporal opportunistic surveys data**
Presenter: **Jafet Belmont**, University of Glasgow, United Kingdom
Co-authors: Claire Miller, Marian Scott, Craig Wilkie

Biodiversity monitoring programs have become essential to describe, predict and map species distributions across large geographic and temporal scales. Unfortunately, collecting species occurrence data in such large-scale studies can be difficult. Thus, citizen science projects, involving volunteers who help to collect data and monitor sites, offer a cost-effective solution to investigate species distributions at large spatial and temporal scales. However, analysing data from these opportunistic recording schemes is challenging because of uneven sampling efforts and species imperfect detection. Over the last decade, the increasing awareness of accounting for species imperfect detection in ecological studies has led to the development of different species distribution models. Particularly, dynamic occupancy models have proven to be a powerful tool for estimating temporal changes in species occurrences while correcting for imperfect detection and varying sampling effort. Thus, we discuss some of the challenges involved with occupancy models and opportunistic recording schemes, and propose a multiple-species flexible dynamic model that enables non-linear effects to be estimated. We applied this model to investigate how dragonflies' population dynamics are affected by temperature levels in waterbodies across the UK and to account for seasonal patterns in species' life cycles.

E0679: **Hilbert regression model for complex responses**
Presenter: **Agnese Maria Di Brisco**, University of Piemonte Orientale, Italy

In a standard regression model, it is generally assumed that the response is normally distributed. In case the response is a percentage or a rate, i.e., it is defined on a bounded interval, a beta-type regression model is preferable. If the response exhibits further complexities, such as bimodality, heavy tails, and outlying observations, proper regression models have to be tailored. A further source of complexity concerns the nature of the covariates, should they be high-dimensional or functional. To deal with these issues, the proposed regression model is the Hilbert flexible beta regression model. The latter is designed to cope with complex bounded responses being based on a special mixture of betas. Moreover, it accounts for Hilbert covariates, either high-dimensional or functional, with a principal component analysis strategy, whereas the estimation issues are addressed with MCMC techniques. Finally, the selection of the most significant coefficients of the expansion of the Hilbert terms is performed through Bayesian variable selection methods that take advantage of shrinkage priors. The effectiveness of the proposal is illustrated with intensive simulation studies. Results concerning a real application aimed at regressing the percentage of milk fat onto spectrometric curves are also illustrated, showing a fit behavior of the proposed model that is more satisfactory in comparison to competing approaches.

E0786: **Nonparametric density estimation over multidimensional domains**
Presenter: **Eleonora Arnone**, University of Padua, Italy
Co-authors: Federico Ferraccioli, Livio Finos, Laura Sangalli

A nonparametric method for density estimation over complicated multidimensional spatial domains is presented. The method combines a likelihood approach with a regularization based on a differential operator, and the estimator is proven to be consistent. The discretization of the estimator is based on finite elements, ensuring high computational efficiency and enabling great flexibility. The proposed method efficiently deals with data scattered over two-dimensional regions having complicated shapes, two-dimensional Riemannian manifolds, or planar networks. Moreover, it captures very well complicated signals having multiple modes with different directions and intensities of anisotropy.

E1077: **Exploratory graph analysis for configural invariance assessment of a test**
Presenter: **Sara Fontanella**, Imperial College London, United Kingdom
Co-authors: Lara Fontanella, Alex Cucco, Nicola Pronello, Pasquale Valentini

In cross-country comparative analyses, self-report survey tools are widely used to examine variations among respondents from different groups, such as citizens of various nations. An important methodological issue, in this situation, relates to the configural invariance of the measurement tool, which holds if the latent structure exhibits the same pattern across various groups. To address this issue, we take an exploratory approach grounded in the paradigm of graph theory. We discuss the use of exploratory graph analysis to assess the configural invariance in the context of a multi-group comparative analysis with measurement instruments comprised of ordered categorical indicators. In this framework, networks are utilised to represent latent constructs, and the covariance between observable indicators is explained through a pattern of causal interactions between the items. Therefore, we postulate that group-specific correlation-based networks would have a comparable structure if the measuring instrument operates consistently across groups. Network embedding will be utilised to look into the similarity of the network structures estimated using a Bayesian approach with sparse-inducing priors and mixture models to identify subgroups of homogeneous graphs. We show through a simulation analysis and real-world applications that the suggested technique can distinguish differences in the latent structure.

EO708 Room S-1.22 RELIABILITY AND STOCHASTICS: THEORY AND APPLICATIONS Chair: Alexandros Karagrigoriou

E1796: **Bootstrapped and kernel-type estimators of reliability indicators in semi-Markov processes**
Presenter: **Eirini Votsi**, Le Mans University, France
Co-authors: Salim Bouzebda

Semi-Markov processes are stochastic processes that are widely used in reliability and related fields. They generalize both jump Markov processes and renewal processes. We consider semi-Markov processes in continuous-time and finite state space. The stochastic behavior of such processes is governed by the semi-Markov kernel. Empirical estimators of the semi-Markov kernel and its functionals have been proposed. The limiting distributions of such estimators have usually complicated expressions, and therefore explicit computation in practice is rather infeasible. To overcome this difficulty, we propose a general bootstrap of empirical semi-Markov kernels and of the conditional transition distributions. We consider a general bootstrap that allows for a unified treatment for resampling methods and provides a flexible framework to handle practical problems. In particular, we present three different types of estimators for the semi-Markov kernel and its functionals: the bootstrapped, the kernel-

type and the bootstrapped kernel-type estimators. We further establish their asymptotic properties and focus on reliability indicators, such as the functions of reliability, availability and maintainability, as well as different failure rates. The asymptotic properties of the latter indicators are obtained by means of martingale techniques. The advantages of the use of such estimators are discussed.

E1814: **Reliability inference for actuarial-financial mathematics**
Presenter: **Andreas Makrides**, University of the Aegean, Cyprus
Co-authors: Christos Meselidis, Alexandros Karagrigoriou
An innovative approach to loss ratio forecasting is developed using a special type of semi-Markov process. Three levels of loss ratio are considered as the states of a semi-Markov process, and semi-Markov process methodology is employed for estimating transition probabilities of loss ratio levels transit from a predefined level to another one.

E1963: **ROCOF of higher order for semi-Markov processes**
Presenter: **Guglielmo Damico**, University G. dÁnnunzio of Chieti-Pescara, Italy
Co-authors: Filippo Petroni
The rate of occurrence of failures (ROCOF) of higher order for continuous time semi-Markov processes (SMP) is studied. This indicator gives information on whether there are a lot of failures or only a few within a time interval. It also considers the relative positioning of tuples of failures in time. We consider SMP with a mixed probability distribution for the initial law of the system, taking into account the possible random starting from any state with any duration. Furthermore, under suitable assumptions on the transition rates, we determine an explicit formula for the ROCOF of higher order, and we recover as particular cases previous results obtained in the literature. A numerical example demonstrates the possibility of using this index in real applications.

E1965: **On the identification of the tail-type of distributions in reliability and actuarial science**
Presenter: **Alexandros Karagrigoriou**, University of The Aegean, Greece
Co-authors: Ilia Vonta, Georgia Papasotiriou, Ioannis Mavrogiannis
The aim is to fill the gap regarding the verification of the log-concavity property, which is a widely studied topic due to the fact that it provides desirable estimating properties. At the same time, log-concavity together with log-convexity are vital in reliability, engineering and stochastic modeling for distinguishing between exponential and light- or heavy-tailed distributions. For the above purpose, we propose a goodness-of-fit exponentiality test which is based on the conspiracy and catastrophe principles which provide a characterization for the exponential distribution. The proposed test is thoroughly discussed and its performance is investigated via simulation studies. A fire insurance dataset is used for demonstration.

EO054 Room S-1.27 ADVANCES ON MODELS FOR TIME SERIES AND LONGITUDINAL DATA Chair: Sabrina Giordano

E1432: **Multi-scale modelling of time series data using state-switching varying-coefficient stochastic differential equations**
Presenter: **Timo Adam**, University of Copenhagen, Denmark
Varying-coefficient stochastic differential equations (SDEs) are popular tools for uncovering mechanistic relationships underlying time series data. By modelling the parameters of the process of interest as smooth functions of covariates, they provide an extension of basic SDEs that allows us to capture more detailed, non-stationary features of the data-generating process. However, in practice, these parameters often vary at multiple time scales, which is illustrated using dive data of beaked whales: while changes in pitch and roll exhibited within some dives can be described by some varying-coefficient SDE, other dives can be better characterised by other varying-coefficient SDEs; a pattern that is not readily accommodated for by the existing approach. We propose state-switching varying-coefficient SDEs as a novel class of statistical models for time series that accounts for such state-switching patterns between dives while simultaneously allowing us to make inferences on the underlying behavioural processes that occur within dives.

E1549: **Maximum likelihood estimation of multivariate regime switching Student-t copula models**
Presenter: **Federico Cortese**, University of Milano-Bicocca, Italy
Co-authors: Fulvia Pennoni, Francesco Bartolucci
A multivariate regime-switching model is presented based on a Student-t copula function with parameters governed by a latent Markov process of the first order. We consider a two-step procedure carried out through the Expectation-Maximization algorithm to estimate model parameters. The main computational burdens involve estimating the correlation matrix R and the number of degrees of freedom ν of the Student t-copula. At this aim, we propose to perform the M-step of the algorithm by computing R given ν using a closed-form solution obtained from a constrained optimization of the log-likelihood using Lagrange multipliers. Then, we numerically maximize the log-likelihood with respect to ν given the estimate of R obtained at the previous iteration. We validate the proposal through a simulation study which shows the computational efficiency and the good finite sample properties of the estimates. We consider an application concerning daily log-returns of the five cryptocurrencies Bitcoin, Ethereum, Ripple, Litecoin and Bitcoin Cash over a four years period. Results show that a regime-switching Student-t copula model with three states can identify bull, neutral and bear market periods based on the intensity of the correlations among cryptocurrencies.

E1771: **A comparative hierarchical analysis of financial literacy using three waves of PISA survey**
Presenter: **Mariangela Zenga**, Universita degli Studi di Milano-Bicocca, Italy
Co-authors: Adele Marshall
Over the past few decades, financial literacy has emerged as a key objective among policymakers in several countries. The PISA survey measures and compares the level of financial literacy skills of young people (adolescent students) in different countries through three waves in 2012, 2015 and 2018. Using the comparisons of hierarchical models, we are able to explain differences in the level of financial literacy across countries considering the three waves of the PISA survey, given the characteristics of the students, their families and the structure of financial education and policies in the countries.

E1986: **Modelling longitudinal claims data using Markov-modulated marked Poisson processes**
Presenter: **Sina Mews**, Bielefeld University, Germany
Markov-modulated marked Poisson processes (MMMPPs) are explored as a natural framework for modelling patients' disease processes over time based on medical claims data. In claims data, patients' interactions with the healthcare system not only occur at random points in time but are also informative, i.e. driven by unobserved disease levels, as poor health conditions usually lead to more frequent interactions. Therefore, we model the observation process as a Markov-modulated Poisson process, where the rate of healthcare interactions is governed by a continuous-time Markov chain, whose states serve as proxies for the patients' latent disease levels. To provide further information on the latent states, we incorporate additional data collected at each observation time (so-called marks), corresponding to the amount of drug usage, for example, into the model. The distribution of these marks - like the event rates - is determined by the underlying (disease) states. Overall, MMMPPs thus jointly model observations and their informative time points by comprising two state-dependent processes: the observation process (corresponding to the event times) and the mark process (corresponding to event-specific information), which are both driven by an underlying continuous-time Markov chain. The approach is illustrated using claims data from patients diagnosed with chronic obstructive pulmonary disease (COPD), revealing inter-individual differences in the state-switching dynamics.

EO651 Room K0.16 GOODNESS-OF FIT AND MODEL SELECTION PROCEDURES Chair: Dimitrios Bagkavos

E0567: **Testing Poissonity of many populations**
Presenter: **Maria Dolores Jimenez-Gamero**, Universidad de Sevilla, Spain
Co-authors: Jacobo de Una-Alvarez

Univariate count data appear in many real-life situations, and the Poisson distribution is frequently used to model this kind of data. Testing the goodness-of-fit of given observations with a probabilistic model is a crucial aspect of data analysis. Because of these reasons, a number of authors have proposed tests for the Poisson law. Most papers on this issue deal with testing Poissonity for a single sample, and the properties of the proposed procedures are studied as the sample size increases. We consider the problem of simultaneously testing Poissonity for k samples, where k can increase with the sample sizes. Moreover, k will be allowed to be even larger than the sample sizes. This is important, for instance, in applications with high-dimensional data, such as those arising from DNA sequencing experiments. The cases of independent samples and weakly dependent samples are both of them studied.

E0563: **A new Gaussian mixture model clustering algorithm**
Presenter: **Polychronis Economou**, University of Patras, Greece

Gaussian mixture models (GMM) are widely used as a probabilistic model for density estimation for multivariate data and as an unsupervised clustering algorithm to provide a soft (fuzzy) clustering to the available data. The GMMs rely on the expectation-maximization algorithm for maximizing the likelihood. A new approach is proposed in the present work, which depends on approximate Bayesian computation and aims not only to estimate the population parameters but also to assign each observation to a specific subpopulation. The performance of the new approach is compared with the expectation-maximization algorithm for the GMM under several challenging simulation problems.

E0758: **The modified (Φ, α)-power divergence family: The zero count case**
Presenter: **Christos Meselidis**, University of the Aegean, Greece
Co-authors: Alexandros Karagrigoriou

Divergence measures play a crucial role in statistical inference. They can be used in the construction of tests of fit as well as for estimation purposes. In the presence of zero frequency cells, a modified version of the (Φ, α) power divergence family, which produces robust estimators that are more efficient than the classical ones, can be utilized. These estimators are then used in the proposed test statistic that involves four indices, through which the stability of the test is achieved in the case of contaminated data. All the aforementioned notions are presented through an extensive simulation study. Regarding the estimation procedure, conclusions are based on the mean squared error, while for the test statistic, both the size and the power of the test are taken into consideration.

E0565: **Goodness-of-fit tests for regression models with a doubly truncated response**
Presenter: **Jacobo de Una-Alvarez**, University of Vigo, Spain

In Survival Analysis, Epidemiology or Reliability, among other fields, doubly truncated data may appear. Double truncation means that the target variable is observed only when it falls within two random limits, which are also available in such a case. Unlike other phenomena of data incompleteness, nonparametric maximum-likelihood estimation with doubly truncated data does not have a closed form; this results in complicated asymptotics. An omnibus goodness-of-fit test for a regression model with a doubly truncated response will be introduced. The test statistic will be based on the distance between two empirical integrated regression functions: one purely nonparametric, and the other one driven by the model to be tested. The underlying process will be a marked empirical process based on weighted residuals, where the weights remove the observational bias induced by the double truncation. The asymptotic null distribution of the test statistic will be obtained for both a fully specified and a parametric regression model. A bootstrap algorithm will be proposed in order to approximate the null distribution of the test in practice. The method will be illustrated with both simulated and real data.

EO500 Room K0.50 DESIGN OF EXPERIMENTS AND APPLICATIONS Chair: Stella Stylianou

E0462: **Weighing matrices for Definitive screening designs**
Presenter: **Stella Stylianou**, RMIT University, Australia

A class of designs called definitive screening designs (DSDs) was proposed, and later it was shown how a few categorical factors could be introduced in those designs. We provide a new general method that can use weighing matrices to construct screening designs with some two-level qualitative factors. The methodology is illustrated through a few small examples. The generated designs are compared to the designs in the literature, and their advantages and disadvantages are discussed.

E0463: **Egde designs from skew-symmetric supplementary difference sets**
Presenter: **Stelios Georgiou**, RMIT University, Australia

The purpose of screening experiments is to identify the dominant variables from a set of many potentially active variables which may affect some characteristic y. Edge designs were introduced in the literature, are constructed by using conference matrices and were proved to be robust designs. We introduce a class of edge designs which are constructed from skew-symmetric supplementary difference sets. These designs are particularly useful since they can be applied for experiments with an even number of factors, and they may exist for orders where conference matrices do not exist. Using this methodology, examples of edge designs for 6, 14, 22, 26, 38, 42, 46, 58, and 62 factors are constructed. Of special interest are the edge designs for studying 22 and 58 factors because edge designs with these parameters have not been constructed in the literature, as conference matrices of the corresponding order do not exist. The suggested edge designs achieve the same model robustness as the traditional edge designs. We also suggest the use of a mirror edge method as a test for the linearity of the true underlying model. We give the details of the methodology and provide some illustrating examples for this new approach. We also show that the new designs have good D-efficiencies when applied to first-order models.

E0868: **Strategies to construct directly optimal and near-optimal symmetric paired choice experiments for main effects models**
Presenter: **Abdulrahman Sultan S Alamri**, RMIT University, Australia

Discrete choice experiments (DCEs) are increasingly used for identifying the underlying influences on an individual's choice behaviour in various fields, e.g., health resources, marketing, transport, economics, and the list goes on. Choosing the DCE design plays an essential role in defining which effects are observable. For paired choice experiments, we present the optimal form of the information matrix for attributes at two levels and main effects models. Moreover, we apply globally D-optimal designs to construct DCEs and address some identification issues by suitably modifying the constructions of D-optimal designs. For this situation, we cover the part where some practitioners somehow may need to use choice sets that are of size other than zero modulo 4, i.e. $N \not\equiv 0 mod 4$. Furthermore, as against the existing efficient designs, our designs have higher D-efficiencies for the same number of choice pairs. Also, our design techniques can be extended to be applied to include situations where attributes of DCEs have a higher number of levels with sufficiently small sample sizes.

E1783: **Supersaturated design-based statistical methods for variable selection**
Presenter: **Tharkeshi Dharmaratne**, RMIT University, Australia
Co-authors: Alysha De Livera, Stelios Georgiou, Stella Stylianou

In experimental studies, supersaturated screening design (SSD)-based statistical methods are commonly used to screen relevant factors when

the number of factors exceeds the run size. Based on simulation studies, several of these SSD methods have shown to be performing well in experimental settings. It motivated The exploration of the use of these SSD methods on observational data for variable selection. Variable selection is a widely-used approach for selecting variables of a statistical model in observational studies, which has often been criticised. Therefore, initially reviewed the latest recommendations and methods that are developed for variable selection in observational studies. The performance of the SSD-based statistical methods is then evaluated using both simulated and real-life datasets, followed by a comparison of their performance with the existing approaches.

EO672 Room S0.03 SPATIAL AND TEMPORAL MODELING IN THE CLIMATE AND ENVIRONMENTAL SCIENCES Chair: Peter Craigmile

E0755: **Recent developments in nonstationary time series modelling with application to the environmental sciences**
Presenter: **Matthew Nunes**, University of Bath, United Kingdom
Co-authors: Euan McGonigle, Rebecca Killick

In many application areas, including the environmental sciences, time series often show second-order characteristics which vary over time. Modelling and estimating this structure is vital for understanding and characterising the evolution of underlying processes. However, many processes also exhibit a first-order (trend) structure. Trend estimation in time series is often performed without consideration of the second-order nonstationary structure; on the other hand, it is common to remove trends prior to nonstationary time series analysis, risking inaccurate estimation of second-order properties. We introduce trend locally stationary wavelet (T-LSW) processes, a modelling framework which extends previous work to capture both first- and second-order nonstationarity, and describe an estimation scheme for the time series quantities, which ensures bias and consistency. We illustrate our model with climatic data examples and outline avenues of further work of interest to practitioners in the environmental sciences.

E0573: **Nonparametric regression on complex constrained domains**
Presenter: **Mu Niu**, University of Glasgow, United Kingdom

A class of intrinsic Gaussian processes (GPs) is proposed for interpolation, regression and classification on manifolds with a primary focus on complex constrained domains or irregularly shaped spaces arising as subsets or submanifolds of R, R2, R3 and beyond. For example, intrinsic GPs can accommodate spatial domains arising as complex subsets of Euclidean space. Intrinsic GPs respect the potentially complex boundary or interior conditions as well as the intrinsic geometry of the spaces. The key novelty of the approach proposed is to utilise the relationship between heat kernels and the transition density of Brownian motion on manifolds for constructing and approximating valid and computationally feasible covariance kernels. This enables intrinsic GPs to be practically applied in great generality, whereas existing approaches for smoothing on constrained domains are limited to simple special cases.

E0516: **Conditional modelling of spatio-temporal sea surface temperature extremes, with high-dimensional inference using INLA**
Presenter: **Emma Simpson**, University College London, United Kingdom
Co-authors: Jenny Wadsworth, Thomas Opitz

Recent extreme value theory literature has seen a significant emphasis on the modelling of spatial extremes, with comparatively little consideration of spatio-temporal extensions. This neglects an important feature of extreme events: their evolution over time. Many existing models for the spatial case are limited by the number of locations they can handle; this impedes extension to space-time settings, where models for higher dimensions are required. Moreover, the spatio-temporal models that do exist are restrictive in terms of the range of extremal dependence types they can capture. Recently, conditional approaches for studying multivariate and spatial extremes have been proposed, which enjoy benefits in terms of computational efficiency and an ability to capture both asymptotic dependence and asymptotic independence. We extend this class of models to a spatio-temporal setting, conditioning on the occurrence of an extreme value at a single space-time location. An application to modelling extreme Red Sea surface temperatures is considered, and Gaussian Markov random fields and the integrated nested Laplace approximation (INLA) are exploited to facilitate inference in higher dimensions.

E0904: **Spectral analysis of multivariate time series with applications to ocean waves**
Presenter: **Adam Sykulski**, Imperial College London, United Kingdom
Co-authors: Jake Grainger, Philip Jonathan, Kevin Ewans

Ocean waves are monitored using buoys which measure three-dimensional displacement at high temporal frequencies over one hertz. Such multivariate measurements allow not only the frequency of waves to be measured, but also their direction, and how these evolve over time. Oceanographers jointly characterise this information in what is called the frequency-direction spectrum. We will discuss what this is and why it is important. Oceanographers have developed numerous parametric models for the frequency-direction spectrum, which can then be fitted to observations by approximating the empirical frequency-direction spectrum; however, the reality is that existing methodologies do a very poor job of fitting these parameters. We will discuss a new approach we have developed to fix this problem. Specifically, we think about how the data is fundamentally collected - as a three-dimensional (orthogonal) multivariate time series. We, therefore, convert the frequency-direction parametric models into multivariate spectral models and then use quasi-likelihood approaches to fit the parameters directly to the Fourier transforms of the observed multivariate data. This yields vastly improved parameter estimates in terms of both reduced bias and variance, as we shall demonstrate with simulations and applications to real data.

EO236 Room S0.11 THEORY AND COMPUTATION IN INFERENCE FOR STOCHASTIC PROCESSES Chair: Hiroki Masuda

E0526: **Estimation for linear parabolic SPDEs in two space dimensions based on high-frequency data**
Presenter: **Masayuki Uchida**, Osaka University, Japan

Parametric estimation is addressed for linear parabolic second-order stochastic partial differential equations (SPDEs) in two space dimensions driven by Q-Wiener processes from high-frequency data in time and space. Minimum contrast estimators have been proposed for unknown parameters of the SPDE in one space dimension driven by the cylindrical Wiener process from high-frequency data and showed the asymptotic normality of the estimators. We first derive minimum contrast estimators for unknown parameters of the coordinate processes of the SPDEs in two space dimensions from the thinned data with respect to space. Next, we deduce approximate coordinate processes of the SPDEs in two space dimensions using the minimum contrast estimators. Finally, we obtain adaptive estimators of the coefficient parameters of the SPDEs in two space dimensions using the approximate coordinate processes from the thinned data with respect to time. It is proved that the adaptive estimators have asymptotic normality under some regularity conditions. Numerical simulations of the proposed estimators are also performed.

E0598: **A Hawkes model with CARMA(p,q) intensity**
Presenter: **Lorenzo Mercuri**, University of Milan, Italy
Co-authors: Andrea Perchiazzo, Edit Rroji

A new model named CARMA(p,q)-Hawkes process is introduced. The Hawkes model with exponential kernel implies a strictly decreasing behaviour of the autocorrelation function, and empirical evidence rejects the monotonicity assumption on the autocorrelation function. The proposed model is a Hawkes process where the intensity follows a Continuous Time Autoregressive Moving Average (CARMA) process and specifically is able to reproduce more realistic dependence structures. We also study the conditions of stationarity and positivity for the intensity and the strong

mixing property for the increments. Furthermore, we compute the likelihood, present a simulation method and discuss an estimation method based on the autocorrelation function. A simulation and estimation exercise highlights the main features of the CARMA(p,q)-Hawkes.

E0905: **Local asymptotic normality for discretely observed jump-diffusion processes**
Presenter: **Teppei Ogihara**, University of Tokyo, Japan
Co-authors: Yuma Uehara

When we try to show the local asymptotic normality (LAN) of jump-diffusion processes with discrete observations, there are two problems. The first one is to control transition density ratios between two different values of the parameter. To solve this, we use the scheme with the so-called L2 regularity condition. The original scheme cannot be applied for jump-diffusion processes because of their fat-tailed behaviors. Therefore, we extend the scheme so that it can be applied to jump-diffusion processes. The second problem is that the transition probability for no jump is quite different from that for the presence of jumps. This fact makes it difficult to identify the asymptotic behavior of the likelihood function. To deal with this problem, we approximate the original likelihood function by using a thresholding likelihood function that detects the existence of jumps. As a consequence of these techniques, we obtain LAN for jump-diffusion processes. Moreover, the quasi-maximum-likelihood and Bayes-type estimators are shown to be asymptotically efficient in this model.

E1834: **Convergence of mean field gradient Langevin dynamics for optimizing two-layer neural networks**
Presenter: **Taiji Suzuki**, University of Tokyo / RIKEN-AIP, Japan
Co-authors: Atsushi Nitanda, Denny Wu, Kazusato Oko

The optimization of two-layer neural networks is discussed via the gradient Langevin dynamics in the mean-field regime. For that purpose, we first establish a linear convergence guarantee of the mean-field gradient Langevin algorithm in the infinite width limit under a uniform log-Sobolev inequality condition. Next, we propose a few specific optimization methods for the finite width and discrete-time setting. In particular, we introduce an algorithm that enjoys linear convergence for a finite-sum loss function based on the stochastic dual coordinate ascent method. Finally, we discuss the linear convergence of the vanilla gradient Langevin dynamics without the infinite width assumption but under slightly different regularity conditions.

EO074 Room S0.13 PROJECTION PURSUIT: APPLICATIONS Chair: Nicola Loperfido

E1269: **Projection pursuit: An empirical tour**
Presenter: **Cinzia Franceschini**, Pollenzo University of Gastronomic Sciences, Italy
Co-authors: Nicola Loperfido

The applications of projection pursuit (PP) to some real data sets are described. Some applications have been published in subject-matter scientific journals and have a straightforward interpretation within the related fields. Other data sets are well-known in the statistical literature. For example, kurtosis minimization sequentially recovers the cluster structure of Fisher Iris Data. The results obtained with PP are compared with those obtained with other dimension reduction methods, for example, principal component analysis and invariant coordinate selection. In all the addressed applications, PP is based on either skewness or kurtosis optimization. The related algorithms are implemented in the R packages Kurt, MaxSkew and MultiSkew.

E0923: **Bayesian modelling of athletic performance**
Presenter: **Maria Zafeiria Spyropoulou**, University of Kent, United Kingdom
Co-authors: Jim Griffin, James Hopker

Publicly available databases of sporting competition results have made it possible to model athletic performance across a wide range of sports accurately. A Bayesian hierarchical model was developed to analyse athletic sporting performance in track and field athletics and weightlifting, accounting for confounding factors such as age, month and year effects, as well as environmental conditions in athletic events. Bayesian variable selection was used to fit a spline model separately to the performance results of each athlete within a performance database. The specific focus of this project was on the development of an approximate algorithm which addresses the issue of computational intensity and lack of processing speed associated with MCMC. This is a two-stage algorithm where the first step is to estimate parameters shared by all athletes, while the second step estimates individual athlete-specific parameters. For the first step, we implemented an EM algorithm in combination with variational Bayes. Then, for the second step, instead of using an MCMC algorithm for all parameters, we used an adaptively scaled individual (ASI) version of the MCMC algorithm to fit the athlete-specific spline model. This approach allows us to utilise parallel computing alongside the ASI algorithm to accelerate the processing speed further. We will illustrate the performance of the algorithm on several databases of sporting performance.

E0823: **Frequency estimation of irregularly sampled time series with red noise**
Presenter: **Efthymia Derezea**, University of Kent, United Kingdom
Co-authors: Alfred Kume

Irregularly sampled time series appear in many fields, such as Astronomy, Climatology, Economics etc. In many cases, it is of interest to estimate and identify possible periodic patterns. While many methods exist for the case of regularly sampled data, the problem is under-explored for time series with unequal time spacing. A simple harmonic model is considered with additional red noise for this type of data. The frequency estimate under this setting is proven to be consistent and asymptotically normal. This result is verified through an extensive simulation study which is reported along with an application to real data from Astronomy.

E1109: **Projection pursuit for bank data**
Presenter: **Alessandro Berti**, Urbino University Carlo Bo, Italy
Co-authors: Nicola Loperfido, Cinzia Franceschini

Projection pursuit for bank data.The world, continental and national crises faced by the financial system since 2008 have been addressed by both national and international authorities by favoring financial stability with respect to economic efficiency. This led to a more concentrated banking system. However, it is still a matter of debate whether more concentrated markets are also more stable markets. We use projection pursuit to investigate the structure of the Italian banking system, with special emphasis on the connections between bank performance and capital requirements. Our data include the balance sheet data collected from a large sample of Italian banks between the year 2008 and year 2021.

EO600 Room Safra Lecture Theatre ADVANCES IN FLEXIBLE REGRESSION MODELLING Chair: Helen Ogden

E0847: **Revisiting the mixture approach to mixed models: Thoughts on clustering and dimension reduction**
Presenter: **Jochen Einbeck**, Durham University, United Kingdom
Co-authors: Yingjuan Zhang

For generalized regression scenarios under various response distributions (Gaussian, Poisson, Binomial), the modern statistical methodology can deal with random effects on one or two levels quite easily. Such a methodology generally assumes a Gaussian distribution for the random effects, enabling access to tools such as the Laplace Approximation in order to integrate these out of the likelihood. An alternative, and less widely known approach, facilitates this integration step by means of a discrete mixture distribution based on a small number of mass points which can be estimated alongside with their masses and any regression parameters in a simple EM algorithm. We review this methodology with particular focus on its (not very widely appreciated) ability to cluster the units under investigation via the posterior probabilities of component membership resulting as a

by-product from the EM algorithm. Several examples will be provided, including a case study involving the analysis of Covid-19 mortality rates. The extension of the methodology to multivariate response scenarios, where the random effect takes on the additional functionality of identifying a one-dimensional latent subspace approximating the original data, is discussed.

E1275: **Elastic linear regression for curves in R^d**
Presenter: **Sonja Greven**, Humboldt University of Berlin, Germany
Co-authors: Lisa Steyer, Almond Stoecker

Regression models are proposed for curve-valued responses in two or more dimensions, where only the image but not the parametrisation of the curves is of interest. Examples of such data are handwritten letters, movement paths or outlines of objects. In the square-root-velocity framework, a parametrisation invariant distance for curves is obtained as the quotient space metric with respect to the action of re-parametrisation, which is by isometries. With this special case in mind, we discuss the generalisation of 'linear' regression to quotient spaces more generally, before illustrating the usefulness of our approach for curves modulo re-parametrisation. We test this model in simulations and apply it to human hippocampi data, obtained from MRI scans. Here we model how the shape of the hippocampus is related to age and Alzheimer's disease. We address the issue of irregularly sampled curves by using splines for modelling smooth predicted curves.

E1342: **Stochastic modelling and forecasting of mortality rates using a combination of semi-parametric and parametric models**
Presenter: **Erengul Dodd**, University of Southampton, United Kingdom
Co-authors: Jon Forster, Peter W F Smith, Jakub Bijak

S methodology is described for smoothing and forecasting mortality rates using a combination of generalised additive models (GAMs) and low-dimensional parametric models. GAMs are a flexible class of semi-parametric statistical models, and they allow us to differentially smooth model components (e.g. cohorts) in an integrated way. GAMs can provide a reasonable fit for the ages where there is adequate data. However, estimation and extrapolation of mortality rates using a GAM at higher ages can be problematic due to high variations in crude rates. At these ages, where exposure numbers are small and data are sparse, parametric models can enable a borrowing of strength across age groups and give a more robust fit. Our methodology assumes a smooth transition between a GAM at lower ages and a fully parametric model at higher ages, and acknowledges uncertainty, especially in the highest age groups.

E1613: **Spatial confounding and spatial+**
Presenter: **Emiko Dupont**, University of Bath, United Kingdom
Co-authors: Simon Wood, Nicole Augustin

Spatial confounding is an issue that can arise when regression models for spatially varying data are used for effect estimation. Such models include spatial random effects to account for the spatial correlation structure in the data. But as spatial random effects are not independent of spatially dependent covariates, they can interfere with the covariate effect estimates and make them unreliable. Traditional methods for dealing with this problem restrict spatial effects to the orthogonal complement of the covariates, however, recent results show that this approach can be problematic. Spatial+ is a novel method for dealing with spatial confounding when the covariate of interest is spatially dependent but not fully determined by spatial location. Theoretical analysis of estimates as well as simulations show that bias, in this case, arises as a direct result of spatial smoothing and, moreover, that it can be avoided by a simple adjustment to the model matrix in the spatial regression model.

EO430 Room Virtual R02 CLUSTERING APPROACHES FOR NOISY DATA Chair: Matthieu Marbac

E0776: **Frugal Gaussian clustering of huge imbalanced datasets through a bin-marginal approach**
Presenter: **Christine Keribin**, INRIA - Paris-Saclay University, France
Co-authors: Filippo Antonazzo, Christophe Biernacki

Clustering conceptually reveals all its interest when the dataset size considerably increases since there is the opportunity to discover tiny but possibly worthwhile clusters which were out of reach with more modest sample sizes. However, clustering is practically faced with computer limits with such high data volume, since it possibly requires extremely high memory and computation resources. In addition, the classical subsampling strategy, often adopted to overcome these limitations, is expected to fail heavily in discovering clusters in the highly imbalanced cluster case. Our proposal first consists in drastically compressing the data volume by just preserving its bin-marginal values, thus discarding the bin-cross ones. Despite this extreme information loss, we then prove the identifiability property for the diagonal mixture model and also introduce a specific EM-like algorithm associated with a composite likelihood approach. This latter is extremely more frugal than a regular but unfeasible EM algorithm expected to be used on our bin-marginal data, while preserving all consistency properties. Finally, numerical experiments highlight that this proposed method outperforms subsampling both in controlled simulations and in various real applications where imbalanced clusters may typically appear.

E1116: **Considering errors in ECG segmentation to improve racehorse arrhythmia detection**
Presenter: **Camille Meneur**, ENSAI, France
Co-authors: Guillaume Dubois, Sandrine Hanne-Poujade, Matthieu Marbac, Pauline Martin, Gilles Stupfler

Electrocardiograms (ECG) are signals that are composed of consecutive beat sequences. Arrhythmias are investigated with ECG and defined by an excessive length of a beat sequence. The real segmentation of the ECG being not observed, segmentation algorithms must be used but can produce errors in the segmentation that deteriorate the detection of arrhythmia. The aim is to take into account the errors made by the segmentation algorithm in order to improve the detection of arrhythmia. This problem can be interpreted as a problem of homogeneity testing made on the distribution of the length of the beat sequences, when this variable suffers from measurement errors. Indeed, the distribution of the length of a beat sequence depends on a latent variable that indicates whether the beat is arrhythmic. To circumvent the issue of measurement errors, we proposed to add, in the homogeneity test, other features extracted from the original signal. Results show that the proposed method permits to differentiate errors from arrhythmic and healthy sequences, and thus to reduce the number of type *I* and type *II* errors. This method is tested on racehorses ECG recorded during training.

E1149: **Swimming technical skills tracking using multivariate functional clustering of Inertial Measurement Unit data**
Presenter: **Antoine Bouvet**, University Rennes 2, France
Co-authors: Matthieu Marbac, Salima El Kolei, Nicolas Bideau

Tracking technical skills during training becomes a major duty for coaches to improve the swimmer's performance. This can be done using miniaturized sensors such as Inertial Measurement Units (IMU). IMU data are multivariate functional data composed of six coordinates describing the 3D accelerometer and gyroscopic temporal records. To investigate the technical skills of the swimmers, two clusterings are performed based on the IMU records. The first clustering aims to provide groups with homogeneous swimming patterns that measure the efficiency of the swimming technique. This clustering is achieved by decomposing the IMU data into Fourier basis and thus fitting a mixture model on the functional basis coefficients. Since the number of basis coefficients is large, a variable selection is performed during clustering to select which coordinates of the functional data are discriminative. The second clustering aims to provide groups with homogeneous variations around the mean swimming pattern. It is performed on the residuals obtained by the decomposition of the IMU data into the functional basis. Thus, it aims to investigate how swimmers can reproduce their mean swimming pattern. We will discuss how the combination of both clusterings can be used to investigate the technical skills of the swimmers.

E1249: **Function-on-function mixture of experts' regression models**
Presenter: **Jean Steve Tamo Tchomgui**, University of Lyon2 and Orange Innovation, France
Co-authors: Julien Jacques, Stephane Chretien, Guillaume FRAYSSE, Vincent BARRIAC
The relationship between the target variable and the predictors that one tries to estimate through a regression model is generally assumed to be identical for all the subjects. However, for unknown reasons or because of unobserved explanatory variables, this relationship may be heterogeneous. We introduce a method to relax this assumption with a regression structure by a group of individuals based on the framework of mixture models. In its original formulation (dedicated to unsupervised learning or explanatory modelling), the group membership probability of an individual is independent of its covariates. Knowing to which regression group a new individual belongs quickly proved difficult in the context of predictive modelling. To address this issue, the mixture model was modified to make this probability depend on the covariates: this is the mixture of experts model. The mixture model and its extension are well known, and implementation tools have been developed in the classical case, i.e. when the target and explanatory variables are both scalars. In the case where we are in the presence of functional observations for both variables, it would be relevant to develop these mixture models as well. This problem has already been tackled but, to the best of our knowledge, only when the predictors are functional. We plan to develop here the mixture of experts model in the case where both predictors and target variable are functional.

EO697 Room Virtual R04 DEPENDENCE MODELS FOR COMPOUND EVENTS Chair: Fabrizio Durante

E0333: **Rainfall interpolation in Canada by a smooth copula-based generalized extreme value approach**
Presenter: **Fatima Palacios Rodriguez**, Universidad de Sevilla, Spain
Co-authors: Elena Di Bernardino, Melina Mailhot
A statistical methodology is proposed based on a hybrid hierarchical smooth GEV copula model to map and predict extreme rainfall in Central Eastern Canada. The rainfall dataset contains a large portion of missing values. In addition, several non-concomitant record periods at different stations are observed in the considered dataset. The proposed approach combines GEV parameters' smooth functions in space through the use of spatial covariates and a flexible hierarchical copula-based model to take into account the dependence between the recording stations. The hierarchical copula structure is detected via a clustering algorithm implemented with a dissimilarity measure designed to handle missing data. Finally, the classical interpolation techniques are compared with our proposed approach.

E1503: **AMH copula-based clustering of variables with application to district heating demand**
Presenter: **Roberta Pappada**, University of Trieste, Italy
Co-authors: F Marta L Di Lascio, Andrea Menapace
Understanding thermal consumption in urban areas is a crucial need to increase the sustainability and efficiency of energy systems and reduce the impact of climate change. The focus is on district heating, which represents one of the key technologies involved in the ongoing process aimed at reducing the waste of energy in a flexible urban energy system. Motivated by the features of high-frequency district heating demand data in the Italian city of Bozen-Bolzano, we develop a clustering methodology grounded on a copula-based dissimilarity measure in the hierarchical framework. To this aim, we exploit the Ali-Mikhail-Haq copula to cluster residential users (buildings) based on the observed time series of heating consumption. The copula approach allows us to tackle both temporal and cross-sectional dependence while considering the effect of meteorological variables and spatial information. We investigate the proposed dissimilarity measure through Monte Carlo studies and compare it with its analogue based on Kendall's rank correlation. In particular, we evaluate the performance of the two measures in terms of overall dendrogram quality and agreement between the two partitions obtained. The application of the proposed measure to district heating demand yields clusters of buildings that are homogeneous with respect to their main characteristics, such as energy efficiency and heating surface, thus providing crucial information for the study of sustainable energy scenarios.

E0521: **The paradigmatic case of the Venice lagoon: A compound investigation and a stakeholder perspective**
Presenter: **Gianfausto Salvadori**, Universita Del Salento, Italy
The Venice lagoon represents a worldwide paradigm of coastal flood. In order to protect Venice, storm surge barriers (MoSE) are activated when specific water levels occur. When MoSE barriers are raised, the only access to the lagoon for ships is represented by the Malamocco lock gate, provided that suitable safety conditions (involving the significant wave height) are satisfied. The statistics of significant wave heights and water levels in the Venice lagoon are investigated: in particular, these variables turn out to be dependent, and their joint occurrence (statistically modeled under the typology of Multivariate Compound events via Copulas) can determine the stop of ship navigation, yielding significant economic losses. Bivariate Return Periods and Failure Probabilities are used to model the statistical behavior of these two relevant variables, in order to provide quantitative guidelines for the management of the tricky hydraulic, maritime and economic system of the Venice lagoon. The stakeholder perspective is illustrated by using available economic data concerning Venice port activities, as well as an interview with the Captain of a chemical ship tank frequently docking at Venice.

E1163: **Construction of generalized Pareto vectors for flexible peaks-over-threshold modeling**
Presenter: **Gwladys Toulemonde**, Universita de Montpellier, France
Co-authors: Jean-Noel Bacro, Carlo Gaetan, Thomas Opitz
A flexible multivariate threshold exceedances modeling is defined based on componentwise ratios between any two independent random vectors with exponential and Gamma marginal distributions. This construction allows flexibility in terms of extremal bivariate dependence. More precisely, asymptotic dependence and independence are possible, as well as hybrid situations. Two useful parametric model classes will be presented. One of the two, based on Gamma convolution models, will be illustrated through a simulation study. Good performance is shown for likelihood-based estimation of summaries of bivariate extremal dependence for several scenarii.

EO216 Room BH (SE) 1.02 HAWKES PROCESSES IN FINANCE Chair: Yoann Potiron

E0181: **Nonparametric estimation of Hawkes branching ratio with Ito semimartingale baseline**
Presenter: **Seunghyeon Yu**, KAIST, Korea, South
Co-authors: Yoann Potiron
In view of their tractability, Hawkes processes are widely employed in high-frequency data. However, even in the absence of kernel (i.e. Poisson case), it is well-documented empirically that the baseline is not constant, reproducing seasonalities from the financial market. We relax this constancy assumption and consider a more realistic nonparametric framework where Hawkes's self-exciting processes feature Ito semimartingale with possible jumps as a baseline. Based on local Poisson estimates and Two Scales truncated Realized Volatility of these estimates, we are able to jointly and consistently estimate the integrated baseline, the integrated volatility of the baseline and the branching ratio, i.e. the L1-norm of the kernel, together with its central limit theory and feasible statistics. As a byproduct of the central limit theory, we develop tests for the presence of Hawkes term, for criticality, for the presence of a time-varying baseline, and for Brownian baseline. A simulation study corroborates the theory and documents the superiority of our branching ratio estimator over two concurrent methods in realistic configurations. An empirical study illustrates that our estimator can capture endogeneity and explain the volatility well with the micro-foundation-based relation.

E0202: **Estimation of integrated intensity in Hawkes processes with time-varying baseline**
Presenter: **Yoann Potiron**, Keio University, Japan
Co-authors: Olivier Scaillet, Seunghyeon Yu

Transaction times are modeled as a Hawkes process with a time-varying baseline and a general kernel. The baseline is assumed to be the sum of a deterministic seasonal component and a stochastic Itô semimartingale with possible jumps. In *mixed* asymptotics, we provide a nonparametric estimation of the integrated intensity. In addition, we decompose the integrated intensity as a sum of contributions of the seasonal part, random part, etc.

E0253: **Alternative asymptotic inference theory for a nonstationary Hawkes process**
Presenter: **Jeffrey Kwan**, University of New South Wales (UNSW) Sydney, Australia
Co-authors: Feng Chen, William Dunsmuir

The Hawkes process is a popular point process model for events that exhibit a local clustering behaviour. It has been previously studied the asymptotic inference theory for a nonstationary Hawkes process where the baseline intensity is proportional to the sometime-varying function with the proportionality constant n tending to infinity. However, it was assumed the excitation kernel to be independent of n, and therefore, as n increases, the waiting times from a baseline event to its excited events are of order $O_P(1)$ while the waiting times between baseline events is $O_P(1/n)$, suggesting the excitation effect is not local any more, which defeats the purpose of choosing the exponential excitation kernel in the first place. To overcome this issue, we study the model in a more realistic setting where the excitation kernel also depends on the limit index n, so that the waiting times to excited events are of the same order ofmagnitude as those between baseline events. We establish consistency and asymptotic normality of the Maximum Likelihood Estimator, and derive the asymptotic properties of the score test. We will illustrate applications to ultra-high frequency financial data and verify the asymptotic results through simulation experiments.

E1954: **Modelling low latency**
Presenter: **Vladimir Volkov**, University of Tasmania, Australia
Co-authors: Yoann Potiron

A novel approach to measuring low latency, defined as the time it takes to learn about an event and generate a response to this event, is proposed. The measure of low latency is obtained from an intensity model, which is an extension of the Hawkes model, allowing a memory kernel to be dependent on an additional unobservable stochastic variable characterised by low latency. Detailed information about cancellation orders and identification of traders, normally used in the literature, is not required in this case, which makes our approach more flexible in applications. Low latency estimates for the US and Canadian stock markets vary between 2 and 9 milliseconds from 2020 to 2021. The US firms are found to be more involved in relative latency competition, implying different risk appetites for firms with different latencies. Changes in low latency have a significant impact on the level of volatility in the US and Canada.

EO154 Room BH (S) 2.02 GRAPHICAL MARKOV MODELS II Chair: Monia Lupparelli

E0597: **Graphical model inference with network-structured variables**
Presenter: **David Rossell**, Universitat Pompeu Fabra, Spain
Co-authors: Jack Jewson, Piotr Zwiernik, Laura Battaglia, Stephen Hansen

A main practical challenge to using graphical models in applications is that the sample size is often limited relative to the number of parameters to be learned. We discuss applications where one has access to external network data that provides valuable external information and effectively increases the sample size. The motivation stems from depicting the relation between COVID19 and social network data, and between the stock market and economic indicators extracted from text data. We propose a graphical LASSO framework where likelihood penalties are guided by external data, and a spike-and-slab prior framework that depicts how partial correlations depend on external network data. We develop computational schemes and software implementations in R and probabilistic programming languages. Our applications show how one may significantly improve interpretation, statistical accuracy and out-of-sample prediction, in some instances using significantly sparser graphical models than would otherwise be necessary.

E1235: **Hierarchical graphical modelling of count metagenomic data**
Presenter: **Veronica Vinciotti**, University of Trento, Italy
Co-authors: Ernst Wit

Unraveling interactions between microbial communities is of vital importance in understanding how microbes influence human health. Rich sources of microbiome data have been generated by the latest sequencing experiments, measuring microbial abundances under a variety of environmental conditions, such as at different body sites or across different time points. We model the complexity of these data, and of the underlying dependency structure, via a Gaussian copula graphical model, and we propose an efficient Bayesian structural learning procedure for inference. Heterogeneity in the data is accounted for both at the individual microbial level, via marginal distributions that are linked parametrically with external covariates, and at the dependency level, with a hierarchical prior on the graph that takes the form of a latent network model, capturing structural relatedness across multiple environments as well as dependencies of the microbial interactions from external covariates.

E0594: **Block structured Gaussian graphical models for spectrometric functional data**
Presenter: **Lucia Paci**, Universita Cattolica del Sacro Cuore, Italy
Co-authors: Alessandro Colombi, Raffaele Argiento, Alessia Pini

Within the framework of Gaussian graphical models, a prior distribution for the underlying graph is introduced to induce a block structure in the adjacency matrix of the graph and to learn relationships between fixed groups of variables. A novel sampling strategy named Double Reversible Jumps Markov chain Monte Carlo is developed for block-structured graph learning, under the conjugate G-Wishart prior. The algorithm proposes moves that add or remove not just a single link but an entire group of edges. The method is applied to smooth functional data, where the classical smoothing procedure is improved by placing a graphical model on the basis expansion coefficients, providing an estimate of their conditional independence structure. Since the elements of a B-Spline basis have compact support, the independence structure is reflected in well-defined portions of the domain. A known partition of the functional domain is exploited to investigate relationships among the substances within the compound.

E1741: **Bayesian structure learning in high-dimensional graphical models**
Presenter: **Reza Mohammadi**, University of Amsterdam, Netherlands

Graphical models have been used in many application areas for learning conditional independence structure among a (possibly large) collection of variables. For these models, Bayesian structure learning, while providing a natural and principled way for model uncertainty, often lag behind frequentist approaches in terms of computational efficiency and scalability. Hence, scalable approaches with theoretical and computational safeguards are critical to leveraging the advantages of posterior inference. We discuss the computational problems related to Bayesian structure learning, and we offer several solutions to cope with the computational issues. To show its empirical usefulness, we present the application of our approach to high-dimensional fMRI data for brain connectivity studies. In addition, we have implemented our method in the R packages BDgraph and ssgraph, which are available at CRAN.

EO464 Room BH (S) 2.05 ADVANCES IN VARIATIONAL APPROXIMATIONS Chair: Luca Maestrini

E0284: **Flexible variational Bayes based on a copula of a mixture**
Presenter: **David Gunawan**, University of Wollongong, Australia
Co-authors: Robert Kohn, David Nott

Variational Bayes methods approximate the posterior density by a family of tractable distributions and use optimisation to estimate the unknown parameters of the approximation. The variational approximation is useful when exact inference is intractable or very costly. A flexible variational approximation is developed based on a copula of a mixture, which is implemented using the natural gradient and a variance reduction method. The efficacy of the approach is illustrated by using simulated and real datasets to approximate multimodal, skewed and heavy-tailed posterior distributions, including application to Bayesian deep feedforward neural network regression models.

E0745: **Fitting structural equation models via variational approximations**
Presenter: **Khue-Dung Dang**, University of Melbourne, Australia
Co-authors: Luca Maestrini

Structural equation models are commonly used to capture the relationship between sets of observed and unobservable variables. Traditionally these models are fitted using frequentist approaches, but recently, researchers and practitioners have developed an increasing interest in Bayesian inference. In Bayesian settings, inference for these models is typically performed via Markov chain Monte Carlo methods, which may be computationally intensive for models with a large number of manifest variables or complex structures. Variational approximations can be a fast alternative; however, they have not been adequately explored for this class of models. We develop a mean field variational Bayes approach for fitting elemental structural equation models and demonstrate how bootstrap can considerably improve the variational approximation quality. We show that this variational approximation method can provide reliable inference while being significantly faster than Markov chain Monte Carlo methods.

E1038: **In mean and variance variable selection with variational approximations**
Presenter: **Mauro Bernardi**, University of Padova, Italy
Co-authors: Luca Maestrini, Giulia Livieri

Variable selection plays a key role in modern statistical research and learning. Major classes of variable selection approaches are implemented using Markov chain Monte Carlo methods. These methods may be computationally impractical for large-scale problems or complex models and faster approximations are desirable or necessary. We develop an approach to variable selection for heteroscedastic regression models based upon semiparametric mean field variational Bayes. The methodology we propose is suitable for models having linear mean and exponential variance functions with prior specifications that induce sparse solutions on the regression coefficients, namely Bayesian lasso, spike-and-slab, and adaptive spike-and-slab lasso. The use of classic mean field variational Bayes leads to the approximating densities having non-standard forms and challenging numerical problems arise in the determination of the optimal approximation. We achieve tractability by imposing a parametric assumption to the approximate marginal posterior densities of variance regression coefficients. Our iterative optimization of the log-likelihood lower bound includes Newton-type steps with analytical derivative expressions for the parametric component of the approximation. This optimization strategy uses new results that solve recurrent issues of constrained optimization involving multivariate skew-normal variational approximations.

E1019: **Stochastic variational inference for heteroskedastic time series models**
Presenter: **Hanwen Xuan**, The University of New South Wales, Australia
Co-authors: Luca Maestrini, Clara Grazian, Feng Chen

Stochastic variational inference algorithms are derived for fitting various heteroskedastic time series models using Gaussian approximating densities. Gaussian, t and skew-t response GARCH models are examined. We implement an efficient stochastic gradient ascent approach based upon the use of control variates or the reparameterization trick and show that the proposed approach offers a fast and accurate alternative to Markov chain Monte Carlo sampling. We also present a sequential updating implementation of our variational algorithms, which is suitable for the construction of an efficient portfolio optimization strategy.

EO310 Room K2.31 (Nash Lec. Theatre) MODERN DIRECTIONAL STATISTICS Chair: Andrea Meilan-Vila

E0790: **A copula model for multivariate circular data**
Presenter: **Christophe Ley**, University of Luxembourg, Luxembourg
Co-authors: Shogo Kato

A new family of distributions is proposed for multivariate circular data. The focus first lies on the trivariate case. Its density can be expressed in simple form without involving infinite sums or integrals. The univariate marginals of the proposed distributions are the uniform distributions on the circle, and therefore the presented family is considered a copula. The bivariate marginals of the proposed distributions are members of the Wehrly-Johnson family. The univariate and bivariate conditional distributions are also well-known in the literature. An efficient algorithm to generate random variates from our model is presented, trigonometric moments, as well as other appealing properties, are given, and maximum likelihood estimation for the presented distributions is considered. Finally, an extension of the proposed family for multivariate circular data is considered.

E0726: **Joint modeling of conditional mean and dispersion with a circular predictor**
Presenter: **Maria Alonso-Pena**, Universidade de Santiago de Compostela, Spain
Co-authors: Irene Gijbels, Rosa Crujeiras

The simultaneous and flexible estimation of the mean regression function and the dispersion function is considered in situations where the response is a count variable, and the predictor variable is circular. The estimation approach is based on the maximization of the circular local likelihood function, without the assumption of any parametric forms for the regression functions. The conditional distribution is assumed to belong to the double exponential family, which allows us to model overdispersion, underdispersion or a combination of both, where the amount of overdispersion (or underdispersion) may change with the value of the circular predictor. We apply this methodology to study how the number of neuronal spikes in a macaque monkey changes with the direction of a visual stimulus. The new approach allows the flexible estimation of not only the expected number of spikes, but also of the variability in the number of spikes, both as a function of the direction of the stimulus.

E1355: **Some undirected graphical models for circular variables**
Presenter: **Agnese Panzera**, University of Florence, Italy
Co-authors: Anna Gottard

Graphical models are a powerful probabilistic tool for studying the conditional independence structure of a set of random variables. This class of multivariate models expresses conditional independence by missing edges in a graph. The associated graph is undirected when all the variables are on equal footing. Then the model is called an undirected graphical model. We explore some multivariate circular distributions focusing on their properties about conditional independence and examine the corresponding classes of undirected graphical models. We discuss some related issues and present some applications in protein folding understanding.

E0766: **Nonparametric density estimation on the polysphere**
Presenter: **Andrea Meilan-Vila**, Universidad Carlos III de Madrid, Spain
Co-authors: Eduardo Garcia-Portugues

Polyspherical data refer to observations on $S_1^d \times \ldots \times S_r^d$, $d_1, \ldots, d_r \geq 1$, where S^d denotes the hypersphere of dimension $d \geq 1$. The poly sphere comprises the circle ($r = d_1 = 1$), sphere ($r = 1$, $d_1 = 2$), and torus ($d_1 = \ldots = d_r = 1$), as particular cases. The goal is to propose and study a kernel density estimator for this type of data. Using a tailored tangent-normal decomposition, the main asymptotic properties of the estimator,

such as bias, variance, pointwise normality, and optimal bandwidth, are obtained. Some guidelines, based on cross-validation and plug-in methods, to select the asymptotically optimal bandwidth parameter are also provided in practice. Moreover, the kernel efficiency with respect to a certain "Epanechnikov" kernel is studied. An application of the methodology to the hippocampus via s-reps on the polysphere $(S^2)^1 68$ is discussed.

EO576 Room K2.41 ADVANCES IN JOINT MEANCOVARIANCE MODELS FOR MULTIVARIATE DATA (VIRTUAL) Chair: Olcay Arslan

E0675: **Joint modelling of location, scatter matrix and skewness of multivariate skew normal distribution**
Presenter: **Yesim Guney**, Ankara University, Turkey

Assuming normality of response is practical from a computational point of view and common for location and scatter matrix models, but is rather restrictive. This assumption is relaxed by using a multivariate skew-normal distribution which includes the normal distribution as a special case and provides flexibility in capturing the asymmetric behavior presented. In this case, besides the location and scatter matrix, the skewness may also be expressed with a model involving some explanatory variables along with other unknown parameters. The objective is to extend the joint mean and covariance model by considering the outcomes to follow a multivariate skew-normal distribution. We propose simultaneous modeling location, scatter matrix, and skewness models of multivariate skew normal distribution by using Pourahmadi's modified Cholesky decomposition. Specifically, our joint model handles variance heterogeneity and skewness, which are typically observed in the collection of longitudinal data from many studies. The maximum likelihood estimation method is considered for the parameters of the proposed model. In addition, numerical studies are developed to show the flexibility and versatility of the proposed model.

E0822: **Joint modeling of mean and scale covariance using empirical likelihood**
Presenter: **Senay Ozdemir**, Afyon Kocatepe University, Turkey
Co-authors: Yesim Guney, Olcay Arslan

A joint mean and covariance model enables the determination of the mean and covariance structure, which has an important role in longitudinal studies that can be used for observation units that are re-measured over time. Some difficulties, including failure to achieve positive definiteness, are encountered in the estimation of the covariance structure in this model. To overcome these difficulties in obtaining the covariance matrix, the modified Cholesky decomposition can be used. The model parameters are widely estimated with the maximum likelihood estimation method, which requires a distribution assumption on error terms. Instead of making the distribution assumption, which causes some difficulties in practice, the usability of the empirical likelihood method, which offers a more flexible estimation by using the information in the sample, can be considered. The empirical likelihood method performs as maximizing empirical likelihood function consisting of probabilistic weights assigned to observations, under some constraints, which includes the weighted forms of estimating equations obtained from maximum likelihood. A small simulation study will be provided to assess the performance of the proposed estimator.

E0886: **Variable selection in robust heteroscedastic models with autoregressive covariance structures using EM-type algorithm**
Presenter: **Fulya Gokalp Yavuz**, Middle East Technical University, Turkey
Co-authors: Yesim Guney, Olcay Arslan

The joint modeling of location and scatter matrix with multivariate t-distribution provides a valuable extension to the classical approach with normal distribution when the data set under consideration involves outliers or heavy tail outcomes. Variable selection is essential in these models since the covariance model has three models built into it, and the number of unknown parameters grows quadratically with the size of the matrix. The first purpose is to obtain the maximum likelihood estimates of the parameters and provide an expectation-maximization-type algorithm as an alternative to the Fisher scoring algorithm widely used for these models in the literature. Parameter estimation and variable selection are achieved simultaneously in a multivariate t joint location and scatter matrix model using shrinkage approaches. To assess the performance of the considered methods, we conducted a simulation study and real data analysis.

E1252: **Parameter estimation of the partially linear models with skew heavy-tailed error distributions**
Presenter: **Olcay Arslan**, Ankara University, Turkey
Co-authors: Fatma Zehra Dogru

Partially linear models are considered an important flexible generalization of the linear model in applications for modelling economic and biometric datasets. Partially linear models contain a non-parametric component of some covariate besides the linear parametric part of the model. In general, the error term in a partially linear model is assumed to have a normal distribution. However, in applications, datasets may not have a normal distribution, so modelling a partially linear model under the assumption of normality may not be appropriate. Partially linear models under the skew-normal distribution were proposed to deal with the skewness in the data sets. However, the dataset may also have a heavy-tailedness problem along with the skewness. Therefore, we will propose modelling partial linear models with the skewed Laplace normal distribution. The skew Laplace normal distribution is a heavy-tailed alternative to the skew-normal distribution with the same number of parameters. We provide an expectation-maximization (EM) algorithm for the maximum likelihood estimation procedure of the proposed skew Laplace Normal partially linear models. We conduct a simulation study and a real data example to demonstrate the performance of the proposed model.

EC385 Room S-2.23 MULTIVARIATE STATISTICS Chair: Bettina Gruen

E0445: **Cluster-robust estimators for multivariate mixed-effects meta-regression**
Presenter: **Thilo Welz**, TU Dortmund University, Germany
Co-authors: Wolfgang Viechtbauer, Markus Pauly

Meta-analyses frequently include trials that report multiple outcomes based on a common set of study participants. These outcomes will generally be correlated. Cluster-robust variance-covariance estimators are a fruitful approach for synthesizing these dependent outcomes. However, when the number of studies is small, statistical tests regarding the model coefficients based on state-of-the-art robust estimators can have inflated type 1 error rates. Therefore, two new cluster-robust estimators are presented, in order to improve small sample performance. For both estimators, the idea is to transform the estimated variances of the residuals using only the diagonal entries of the hat matrix. The proposals are asymptotically equivalent to previously suggested cluster-robust estimators, such as the bias-reduced linearization approach. The methods are compared and contrasted based on empirical coverage of confidence regions for the coefficient vector β in a Monte Carlo simulation study. The focus is on bivariate meta-regression with a single covariate.

E1544: **Supervised dimensionality reduction method for heterogeneous sources data**
Presenter: **Kenta Sakamoto**, Doshisha University, Japan
Co-authors: Hiroshi Yadohisa

Numerous studies have been conducted on multiple multivariate datasets in biology and information retrieval. In these studies, multiple multivariate data were obtained from multiple sources for the same individual and simultaneously analyzed. Information potentially relevant to the individual is often obtained separately from multivariate data in these studies. This information is called label information. Especially in biology, studies have prioritized extracting label information. However, previous studies have not considered the heterogeneity of each information source. Consequently, determining which information sources are relevant to the label remains difficult. Additionally, biological studies have indicated that some information sources may not contribute to label identification, and using information sources that have little relevance to labels for learning is not the best option. Therefore, reducing the influence of information sources that contribute little to label identification is necessary. Therefore, a method is proposed for stable learning even when including information sources that are not very relevant for label identification. Addition-

ally, we quantitatively evaluated the importance of each information source on the label. This was expected to facilitate the interpretation of the characteristics of each information source.

E1560: **Analysis of multifactorial relative data**
Presenter: **Viktorie Nesrstova**, Palacky University, Olomouc, Czech Republic
Co-authors: Paulina Jaskova, Ivana Pavlu, Kamila Facevicova, Karel Hron

Data of a relative nature (also referred to as compositional data) frequently occur in a number of applications, such as geochemistry, metabolomics or time-use epidemiology. Due to their specific nature, which is expressed by their scale invariance property, a careful approach to their statistical processing is required. This is embedded in the framework of the logratio methodology. For the vector case of compositional data, to obtain simple information contained in pairwise logratios, so-called backwards pivot coordinates were introduced in order to set up an orthonormal coordinate system which enables reliable and interpretable statistical processing of compositions. This approach was already applied in principal components analysis and regression analysis. However, there is still a lack of suitable methods for data consisting of several factors (i.e. multifactorial data). Our aim is to extend the framework of backwards coordinates to the case of compositional tables, two-factorial compositions. For these structures, the elemental information lies in simple log odds ratios and pairwise row and column balances. The performance of this approach is demonstrated in data from time-use epidemiology depicting the relative structure of movement behaviour.

E1615: **Two-mode cluster elastic net**
Presenter: **Kaito Oi**, Doshisha University, Japan
Co-authors: Shintaro Yuki, Kensuke Tanioka, Hiroshi Yadohisa

When multiple regression analysis is applied to data with highly correlated groups of explanatory variables, the estimation of regression coefficients becomes unstable. Cluster elastic net (CEN) has been proposed as one of the methods to solve this problem. CEN infers clusters of features from the data based on the correlation among the variables and association with the response. As a result, CEN can predict the target variable with higher accuracy than multiple regression analysis when applied to such data. However, a drawback of CEN is that the prediction accuracy of the objective variable is reduced for data in which there are several unknown latent homogeneous groups. We propose a two-mode CEN to estimate clusters of individuals and partial regression coefficients for each cluster of individuals that improves the prediction accuracy of the target variable. This method improves the prediction accuracy of the target variable over CEN for data with several unknown latent homogeneous groups. We illustrate the performance of our approach through a simulation study and applying genetic data.

EC814 Room S-1.04 VARIABLE SELECTION Chair: Asaf Weinstein

E1717: **Distributional regression models with automatic variable selection**
Presenter: **Meadhbh ONeill**, University of Limerick, Ireland
Co-authors: Kevin Burke

Automatic variable selection and parameter estimation are performed simultaneously by our proposed distributional regression framework. This flexible method naturally adapts to heteroscedasticity in the data by allowing covariates to enter the model through multiple distributional parameters at the same time (e.g., location and scale). Automatic variable selection is performed using an information criterion that is optimized directly. The typical challenge of tuning parameter selection in lasso-type problems is circumvented, since the penalty parameter is known from the outset, e.g., it is log(n) for the BIC. As there are multiple regression components in the distributional regression setting (and, hence, each distributional parameter may have its own separate tuning parameter), our smooth information criterion is particularly computationally advantageous, since the tuning parameters are known and fixed from the start. This avoids the computationally demanding two-dimensional grid search that is typically carried out. Furthermore, the smooth (differentiable) penalty enables standard Newton Raphson optimization to be employed, making our approach more straightforward to implement than existing procedures. We will show that the method performs favourably in simulation studies and on real data.

E0514: **Simultaneous variable selection and fusion of categorical covariates levels in penalized logistic regression**
Presenter: **Lea Johanna Kaufmann**, RWTH Aachen University, Germany
Co-authors: Maria Kateri

In penalized logistic regression for high-dimensional data, including categorical covariates, dimension reduction of the parameter vector can be achieved not only through variable selection but also through the informative fusion of factor levels. For this purpose, a new regularization technique called L_0-fused group lasso, which simultaneously performs factors selection and fusion of their levels, is introduced. The factors selection procedure is enforced by a group lasso penalty while the levels fusion is based on the L_0 "norm" on the differences of the corresponding coefficients, suitably adjusted for nominal and ordinal covariates. Theoretical properties, such as existence and $\sqrt{n}$-consistency of estimators, along with oracle properties, are investigated for the fixed dimensional case. These results are extended to the case of a diverging number of parameters growing with the sample size. Further, algorithms for handling the associated non-convex optimization problem and obtaining the L_0-fused group lasso estimators are developed. The performance of the proposed procedure is investigated by simulation studies.

E1487: **Gene selection using generalized linear measurement error models**
Presenter: **Hajoung Lee**, Sungkyunkwan University, Korea, South
Co-authors: Jaejik Kim

Gene expression data is obtained by measuring the amount of DNA's genetic information expressed through each gene and its expression is involved in protein production, which is important in cell functioning. So far, many studies have been conducted to select significant genes from such data to understand disease causes and contribute to the development of medications and therapies. However, measurement errors caused by simultaneously measuring tens of thousands of genes with high-throughput equipment are inevitable, and gene selection considering them is uncommon. This is because it is practically difficult to quantify the measurement errors due to their unclear sources. However, if they are not considered in gene selection, it may cause an increase in the number of falsely discovered genes. To alleviate this problem, a gene selection method is proposed using generalized linear measurement error models (GLMEMs). Furthermore, to consider ultra-high dimensionality, we develop an iterative gene screening algorithm which repeats filtering and regularization in the GLMEM framework. The proposed method can reduce the number of falsely discovered genes, and it can also provide stable gene selection results under measurement errors. These results are verified through simulation studies and real data analysis.

E1847: **Sparse pairwise logratio variable selection for high-dimensional compositional data**
Presenter: **Paulina Jaskova**, Palacky University Olomouc, Czech Republic
Co-authors: Matthias Templ, Karel Hron, Javier Palarea-Albaladejo

In omics sciences, biomarker identification is of paramount importance. However, from a statistical perspective, this is a challenging task due to the high dimensionality of the data and the associated computational burden. Metabolomics data have been characterized as compositional data, relative data in which the relevant information is contained in (log)ratios between the variables/components that make up the observed metabolomic profile. Accordingly, it is possible to express biomarkers in terms of log-contrasts or any logratio coordinate representation. Alternatively, we can consider them directly in terms of their basic information provided by pairwise logratios. The main goal is to present and discuss a procedure for variable selection based on pairwise logratios from high-dimensional compositional data in the framework of the orthonormal (orthogonal) logratio

coordinate approach. After an initial dimensionality reduction aimed at filtering out noisy variables through univariate data processing, a selection algorithm is applied to obtain non-overlapping pairwise logratios, which are then used to effectively construct an orthonormal logratio coordinate system. This covers all possible pairwise logratios of a (sub)composition formed from such a set of initial pairwise logratios. Partial least squares regression is then applied to identify significant logratios. The properties of this new approach will be investigated using real, high-dimensional compositions.

EC789 Room K0.18 BIOSTATISTICS Chair: Shaun Seaman

E0426: **Transformation models for ROC analysis**
Presenter: **Ainesh Sewak**, University of Zurich, Switzerland
Co-authors: Torsten Hothorn

Receiver operating characteristic (ROC) analysis is one of the most popular approaches for evaluating and comparing the accuracy of medical diagnostic tests. Although various methodologies have been developed for estimating ROC curves and their associated summary indices, there is no consensus on a single framework that can provide consistent statistical inference whilst handling the complexities associated with medical data. Such complexities might include covariates that influence the diagnostic potential of a test, ordinal test data, censored data due to instrument detection limits or correlated biomarkers. We propose a regression model for the transformed test results, which exploits the invariance of ROC curves to monotonic transformations and naturally accommodates these features. Our use of maximum likelihood inference guarantees the asymptotic efficiency of the resulting estimators and associated confidence intervals. Simulation studies show that the estimates based on transformation models are unbiased and yield coverage at nominal levels. The methodology is applied to a cross-sectional study of metabolic syndrome where we investigate the covariate-specific performance of weight-to-height ratio as a non-invasive diagnostic test. Software implementations for all the methods described in the article are provided in the tram R package.

E1677: **Estimation of tissue profiles from blood RNA-seq based on latent Dirichlet allocation**
Presenter: **Shintaro Yuki**, Doshisha University, Japan
Co-authors: Yusuke Matsui, Yoshikazu Terada, Hiroshi Yadohisa

Disease prediction based on gene expression data from blood samples is clinically important. However, since the expression of the blood is a mixture of molecules from multiple tissues, it is necessary to estimate tissue-specific profiles to know the disease's source. To address this problem, we consider an estimation method using Latent Dirichlet Allocation (LDA), assuming that tissue-specific molecular markers are given as a priori information. Specifically, consider assuming the prior information described above as a topic-specific prior distribution for each topic of word frequency in the LDA. Another related method is penalized LDA, which addresses the effects of housekeeping genes corresponding to the "stop word". In particular, RNA-seq gene expression data is high-dimensional and contains an excess of zeros. We will discuss methods to achieve accuracy and robustness in such situations and to extract interpretable biological information.

E1666: **Comparing classification methods and their generalisability on antibody sequences**
Presenter: **Lutecia Servius**, King's College London, United Kingdom
Co-authors: Joseph Ng, Davide Pigoli, Franca Fraternali

Class Switch Recombination (CSR) is a biological process where antibodies change isotopes to adapt their function. The mechanism of CSR is not well understood, but high throughput sampling of antibody sequences from human samples offers an opportunity to build data-driven models to understand its determinants. The performance of logistic regression (LR), random forest (RF), and support vector machine (SVM) classifiers are compared on Respiratory Syncytial Virus (RSV) and hospitalised COVID 19 patient antibody sequence data sets. The results of the model performance show that the inclusion of data from a new set of patients in the test set affects the performance of all the models considered with, on average, 20% reduction in the model accuracy compared to when the observations from the same patient are split between the training and test set. The RF and SVM performance is more affected than the LR. Specifically for the RF, a further look into the hyperparameters of the model indicates that being allowed to grow to maximum depth, as is the default for many packages, results in a model that performs well in identifying patient-specific trends but is not open to generalisation. A generalised logistic mixed model was then trained to detect the features that contribute to CSR across datasets. These analyses suggest that there is a signal in the antibody sequences that indicates CSR thus, results should be generalisable.

E1757: **A compartmental model for smoking habits in Tuscany (Italy)**
Presenter: **Michela Baccini**, University of Florence, Italy
Co-authors: Alessio Lachi, Cecilia Viscardi, Giulia Cereda, Giulia Carreras

Investigating smoking habits in the population is crucial to plan appropriate tobacco control policies. We develop a compartmental model to describe the smoking dynamics in Tuscany (Italy) from 1993 to 2019 and forecast them until 2043. The model, which allows for mortality and new births, assumes that at each time, the population is divided into the following compartments: Never, Current, and Former smokers. Never smokers can become smokers, smokers can become ex-smokers and ex-smokers may relapse into smoking. Calibrating the size of the compartments on the estimated percentages of smokers, ex-smokers, and never-smokers arising from annual surveys, we estimate flexibly through regression splines the probability of starting and quitting smoking by age, and the probability of relapsing smoking as a negative exponential function of time from cessation. We also estimate the evolution of the number of individuals in the compartments in the past and we predict it in the future, under the assumption of unchanged transition parameters. We obtained the confidence intervals besides estimates through a parametric bootstrap procedure. Separate analyses were performed by gender. The predictive performance of the compartmental models, evaluated via cross-validation, was good. In a context of an overall decrease in smoking prevalence, the transition rates between compartments were quite different among males and females, with an evident change of them over the calendar period.

EC757 Room K0.19 HIGH-DIMENSIONAL STATISTICS Chair: Zeng Li

E1821: **Exponential bounds for regularized Hotelling statistics in high dimension**
Presenter: **Emmanuelle Gautherat**, University of Reims, France
Co-authors: Patrice Bertail, El Mehdi Issouani

In many applications (for instance, in genomics or natural language processes), the dimension of the parameter of interest q is large in comparison to the sample size n and sometimes increasing with n. Consider, for instance, the problem of estimating or testing a mean of variables in $\mathbb{R}^q$, with $q > n$; in that case, the empirical covariance matrix is not full rank and does not even converge to the true one when $n \to \infty$ so that the usual "studentized statistics" or Hotteling T^2 tests are no longer valid. It is thus important to construct estimators and testing procedures which take into account the high dimensional aspects of the problem. One relevant proposition which has been developed in the statistical literature is to use a penalized estimator of the covariance matrix, which is invertible and to use this matrix in tests. In that spirit, some authors have obtained asymptotically valid penalized Hotteling T^2 tests for the mean in the Gaussian case for high dimension framework, when n and $q = q(n)$ goes to ∞ at some specific rate. The purpose of that work is to further explore the finite sample properties of such tests by deriving the exponential bound of some correctly penalized Hotteling T^2.

E1927: **Inference on three-pass regression filter with high-dimensional target variables**
Presenter: **Shou-Yung Yin**, National Taipei University, Taiwan

A framework is considered for high-dimensional target variables using the three-pass regression filter (3PRF). We propose an estimator that involves two steps. First, we use the diversified projection to extract the information from the high-dimensional target variables. Then we adopt 3PRF to ensure that the factors from regressors can improve the forecast performance. The advantage of this approach is that we do not need to impose a number of factors, and the closed-form solution is easy to obtain. Consistency and asymptotic normality are then established. The simulation study shows that the proposed approach performs well when the number of factors is wrongly imposed while the results of using principal components analysis are sensitive. In the empirical study, we use the proposed method to extract the common components which can be used to predict the fundamentals of the dynamics of house prices in the U.S.

E1994: **Densely connected sub-gaussian linear structural equation model learning via l1- and l2-regularized regressions**
Presenter: **Semin Choi**, University of Seoul, Korea, South
Co-authors: Gunwoong Park
A new algorithm is developed for learning densely connected sub-Gaussian linear structural equation models (SEMs) in high-dimensional settings, where the number of nodes increases with an increasing number of samples. The proposed algorithm consists of two main steps: (i) the component-wise ordering estimation using L2-regularized regression and (ii) the presence of edge estimation using L1-regularized regression. Hence, the proposed algorithm can recover a large degree graph with a small degree constraint.

E0728: **A shrinkage likelihood ratio test for high-dimensional subgroup analysis with a logistic-normal mixture model**
Presenter: **Shota Takeishi**, University of Tokyo, Japan
The focus is on testing the existence of a subgroup with an enhanced treatment effect under the setting where the subgroup membership is potentially characterized by high-dimensional covariates. Using a logistic-normal mixture model, we propose a shrinkage likelihood ratio test built on a modified likelihood function that shrinks high-dimensional unidentified parameters towards zero when there exists no subgroup. This shrinkage helps handle the irregularity of the testing problem in the logistic-normal mixture model. It enables us to derive a tractable chi-squired-type asymptotic null distribution even under the high-dimensional regime.

EC811 Room K0.20 EXTREME VALUES Chair: Stephane Girard

E0358: **On runs estimator of the extremal index: Dealing with clusters of exceedances with atypical dimensions**
Presenter: **Manuela Souto de Miranda**, University of Aveiro, Portugal
Co-authors: Cristina Miranda, Ivette Gomes
In Extreme Value Theory, the extremal index can have an important role when the exceedances above high fixed thresholds present a dependence structure, as it happens when they occur in clusters of exceedances. Under general conditions, the extremal index exists and is related to the clusters dimension in the limit distribution; since then, it coincides with the reciprocal of the mean clusters dimension. Several estimators have been proposed for that parameter, namely the runs estimator, which counts the number of exceedances that occur before a non–exceedance observation. A robust version of the runs estimator in the presence of atypical cluster sizes is investigated. The procedure is based on robust methods for counting processes and on the use of negative binomial regression. A simulation study illustrates the performance of the estimator when atypical cluster sizes are induced by contamination, considering different types of contamination.

E1894: **Joint asymptotic behavior of maxima over subsets of concomitants in the extremal dependence framework**
Presenter: **Amir Khorrami Chokami**, University of Turin, Italy
Co-authors: Marie Kratz
The study of concomitants has recently met a renewed interest due to its applications in selection procedures. For instance, concomitants are used in ranked-set sampling, to achieve efficiency and reduce cost when compared to simple random sampling. In parallel, the search for new methods to provide a rich description of extremal dependence among multiple time series has rapidly grown, due also to its numerous practical implications and the lack of suitable models to assess it. The aim is to investigate extremal dependence when choosing the concomitants approach. We show how the extremal dependence of a vector (X,Y) impacts the asymptotic behavior of the maxima over subsets of concomitants. Discussing the various conditions and results, we point out the fundamental role played by the marginal distributions of X and Y.

E1877: **Time series prediction and extreme events**
Presenter: **Manuela Neves**, FCiencias.ID, Associacao para a Investigacao e Desenvolvimento de Ciencias, Portugal
Co-authors: Clara Cordeiro
Time series forecasting, i.e. making predictions from historical data available, is of major importance with a wide variety of applications, such as in finance, weather, the healthcare sector, etc. It is an intensively studied topic, but the existence of extreme events can result in weak performance and low accuracy in the results. Extreme events are rare but do play a critical role in many real applications. Whenever the focus is on large values, estimation is usually performed based on the largest k order statistics in the sample or on the excesses over a high level u. In Extreme Value Analysis and whenever dealing with large values, a few primordial parameters need to be adequately estimated. As we are interested in forecasting extremes in time series, procedures on time series and extreme value theory will come together. Resampling techniques and time series methods for modelling and predicting a time series are computational procedures proposed to improve the performance of the results. A simulation study and applications to real data sets are performed in the R software.

E2002: **Partial tail correlation coefficient applied to extremal network learning**
Presenter: **Yan Gong**, KAUST, Saudi Arabia
Co-authors: Peng Zhong, Thomas Opitz, Raphael Huser
A novel extremal dependence measure si proposed called the partial tail correlation coefficient (PTCC), which is an analogy of the partial correlation coefficient in the non-extreme setting of multivariate analysis, based on the framework of multivariate regular variation and transformed-linear algebra operations. Unlike other recently introduced asymptotic independence frameworks for extremes, our approach requires only minimal modeling assumptions and can thus be used generally in exploratory analyses to learn the structure of extremal graphical models. Thanks to representations similar to traditional graphical models where edges correspond to the non-zero entries of a precision matrix, we can exploit classical inference methods for high-dimensional data, such as the graphical LASSO with Laplacian spectral constraints, to efficiently learn the extremal network structure via the PTCC. The application of our new tools to study extreme risk networks for two datasets extracts meaningful extremal structures and allows for relevant interpretations. Specifically, our analysis of extreme river discharges observed at a set of monitoring stations in the upper Danube basin shows that our proposed method is able to recover the true river flow network quite accurately, and our analysis of historical global currency exchange rate data reveals interesting insights into the dynamical interactions between major economies during critical periods of stress.

EC810 Room S0.12 ROBUST STATISTICS AND DEPTH FUNCTIONS Chair: Sara Taskinen

E0543: **Robust, rank-based estimation of mixed effects models**
Presenter: **Barbara Brune**, TU Wien, Austria
Co-authors: Irene Ortner, Peter Filzmoser
Existing robust methods for the estimation of mixed effects models based on M-estimation and related concepts are often computationally very

expensive and rely on parameter tuning. Rank-based estimation methods offer an attractive alternative to classic M-estimation, as they are computationally cheap and robust. So far, the methodology published in this field only covers simple mixed effects models with random intercepts. We aim to close a gap in the literature regarding the estimation of more complex random effects structures, and develop an estimation framework for mixed effects models with random slopes. By modifying the norm used for estimation, the estimates can further be robustified against leverage points. The resulting residuals and weights can be used for diagnostic purposes, such as identifying unusual observations on both overall and group levels. The theoretical properties of the estimator are studied by means of simulation studies. The method is illustrated with an application to data from accelerated ageing experiments on photovoltaic modules.

E1985: **Smart initialisation and approximate loss function for robust regression**
Presenter: **Thomas Servotte**, University of Antwerp, Belgium
Co-authors: Tim Verdonck, Jakob Raymaekers

Two of the most common methods for robust regression are least trimmed squares (LTS) and least median squares (LMS) regression. Both of these methods require sorting the squared residuals. Because sorting is not a differentiable operation, end-to-end optimalisation with gradient-based methods is not stable. Furthermore, existing algorithms for estimating LTS and LMS regressors rely on multiple random initial starting points. We propose and investigate two potential improvements to LTS and LMS: (1) the use of soft differentiable sorting in the loss functions and (2) deterministic initialisation of the estimators using the wrapping transformation. We show that deterministic initialisation has significant benefits for LTS and LMS, both for predictive accuracy and computational speed. The soft loss function mostly benefits LMS, as it makes it possible to apply iterative optimisation schemes to the LMS loss function. We also demonstrate the potential application of the Soft LTS loss function to non-linear regression problems using neural networks.

E1763: **Anomaly component analysis (ACA)**
Presenter: **Romain Valla**, Telecom Paris, Institut Polytechnique de Paris, France
Co-authors: Pavlo Mozharovskyi, Florence d Alche-Buc

At the crossway of Machine Learning and Data Analysis, Anomaly Detection aims to identify observations that exhibit abnormal behaviour. Be it measurement errors, disease development, severe weather, production quality default(s) (items) or failed equipment, financial frauds or crisis events, their on-time identification and isolation constitute an important task in almost any area of industry and science. While a substantial body of literature is devoted to the detection of anomalies, little attention is paid to their explanation. This is the case mostly due to the intrinsically non-supervised nature of the task and the non-robustness of the exploratory methods like the principal component analysis (PCA). We introduce a new statistical tool dedicated to the exploratory analysis of abnormal observations using data depth as a score. Anomaly component analysis (shortly ACA) is a method that searches a low-dimensional data representation that best visualises and explains anomalies. Based on this, we further propose a procedure for finding clusters of anomalies in Euclidean space. In a comparative study, ACA proves advantageous for anomaly analysis with respect to methods present in the literature.

E1811: **On polynomial-time algorithms for data depths**
Presenter: **Jeremy Guerin**, LTCI, Telecom Paris, Institut Polytechnique de Paris, France
Co-authors: Pavlo Mozharovskyi

Data depth is a set of methods in non-parametric statistics that generalize to higher dimensions, such concepts as cumulative distribution function, median, and quantiles. Having undergone substantial theoretical developments and being known for its attractive properties that include affine invariance and robustness, statistical depth function became a universal methodology with numerous applications. Nevertheless, employing it in practice can be impeded by the high computational complexity of algorithms, in particular for larger data sets. In order to solve this computation problem, we define a new class of depths that can be (re)formulated as a polynomial optimization problem. We study the properties of this class and show that it includes several popular depth notions. To compute these depth functions, we suggest using a hierarchy of semi-definite programming relaxations based on the sum of squares certificates of positivity. The goal is to provide polynomial-time algorithms for depth functions, depending on their properties.

CO617 Room Virtual R01 BAYESIAN ANALYSIS OF FINANCE AND MACROECONOMICS Chair: Toshiaki Watanabe

C0867: **Posterior inferences on incomplete structural models: The minimal econometric interpretation**
Presenter: **Takashi Kano**, Hitotsubashi University, Japan

The minimal econometric interpretation (MEI) of DSGE models provides formal model evaluation and comparison of misspecified nonlinear SDE models in concert with atheoretical reference models. The MEI approach recognizes DSGE models as incomplete econometric tools that provide only prior distributions of targeted population moments but have no implications for actual data and sample moments. A Bayesian posterior inference method is developed based on the MEI approach. The prior distributions of targeted population moments simulated by the DSGE model impose restrictions on the hyperparameters of the Dirichlet distributions, which are natural conjugate priors for the multinomial distributions that the corresponding posterior distributions estimated by the reference model follow. The Polya marginal likelihood of the resulting restricted Dirichlet-multinomial model has a tractive approximated log-linear representation of the Jensen-Shannon divergence that the proposed distribution-matching posterior inference uses the limited information likelihood function. Monte Carlo experiments prove that the MEI posterior sampler rightly infers the calibrated structural parameters of an equilibrium nonlinear asset pricing model and detects the correctly specified model with the posterior odds ratios.

C0940: **Individual trend inflation**
Presenter: **Toshitaka Sekine**, Hitotsubashi University, Japan

Recent approaches to estimate trend inflation from the survey responses of individual forecasters are extended. It relies on a noisy information model to estimate the trend inflation of individual forecasters. Applying the model to the recent Japanese data, it reveals that the added noise term plays a crucial role, and there exists considerable heterogeneity among individual trend inflation forecasts that drives the dynamics of the mean trend inflation forecasts. Divergences in forecasts, as well as moves in estimates of trend inflation, are largely driven by an identifiable group of forecasters who see less noise in the inflationary process, expect the impact of transitory inflationary shocks to wane more quickly, and are more flexible in adjusting their forecasts of trend inflation in response to new information.

C1048: **Bayesian estimation of systemic risk in financial markets with dynamic networks**
Presenter: **Mike So**, The Hong Kong University of Science and Technology, Hong Kong

In financial markets, systemic risk is a kind of risk that the failure of one stock in the market triggers a sequence of failures. The study proposes a Bayesian decision scheme to dynamically keep track of the systemic risk under any preference and restriction in financial risk management. We begin with capturing the moving correlations of stock returns. The correlation represents the strength of the relationship among stocks. Then, we construct a dynamic financial network to link together those stocks with a strong relationship. Making use of the concept of financial space, which is related to the position of stocks on the network plot, we locate two stocks in the financial space at a closer distance when the relationship between these two stocks is strong. Using the distance between stocks in the financial space, together with the preference and restriction in financial risk management, we propose a systemic risk measure. We demonstrate the use of the proposed systemic risk measure in producing early warning signals to the global financial instabilities using financial in 2020 to 2022.

C1161: **High-frequency realized stochastic volatility model with the generalized hyperbolic skew Student's t-distribution**
Presenter: **Jouchi Nakajima**, Hitotsubashi University, Japan
Co-authors: Toshiaki Watanabe
The high-frequency stochastic volatility model is fit to intraday returns assuming that the intraday volatility consists of the autoregressive process, the seasonal component of the intraday volatility pattern, and the announcement component responding to macroeconomic announcements. The high-frequency realized stochastic volatility model augments the high-frequency stochastic volatility model with the daily realized volatility by taking account of the bias in the daily realized volatility caused by microstructure noise and non-trading hours. This article extends the high-frequency realized stochastic volatility model such that the return distribution follows the generalized hyperbolic (GH) skew Student's t-distribution. A Bayesian method using an efficient Markov chain Monte Carlo is developed for the analysis of the proposed model. The application to tail risk management such as Value-at-Risk (VaR) and expected shortfall (ES) is provided using the 5-minute returns of S&P 500 E-mini futures.

CO547 Room Virtual R03 NEW METHODS IN MULTIVARIATE BAYESIAN MODELING Chair: Annika Camehl

C1851: **Understanding trend inflation through the lens of the goods and services sectors**
Presenter: **Benjamin Wong**, Monash University, Australia
The goods and services sectors are distinguished in an otherwise standard unobserved components model of U.S. inflation. The main finding is that since the 1990s, overall variation in aggregate trend inflation has been predominantly driven by the services sector, while both sectors used to contribute to the overall variation in aggregate trend inflation before the 1990s. We pinpoint two key changes in sector-specific inflation dynamics which led to our main result: (i) a large fall in the volatility of trend goods inflation; and (ii) the disappearance of comovement between trend goods and trend services inflation. While our results appear to be robust to inflation developments associated with the recent Covid-19 pandemic, we caveat that estimates obtained during the pandemic period are associated with increased estimation uncertainty.

C1853: **A general Bayesian approach to multiple-output quantile regression**
Presenter: **Annika Camehl**, Erasmus University Rotterdam, Netherlands
Co-authors: Dennis Fok, Kathrin Gruber
Regression quantiles reveal relationships between variables outside the center of the distribution. Simultaneously studying multiple responses requires multivariate quantiles. Current definitions of multivariate quantiles can cover large parts of the domain with very low probability, and/or their covered probability does not equal the pre-set quantile level. We suggest superlevel sets of the multivariate density function as an alternative multivariate quantile definition. This quantile set contains all points in the domain for which the density exceeds a certain level. We show that such a quantile has a number of favorable mathematical and intuitive features. For empirical applications, we, first, use an overfitting Gaussian mixture model to fit the multivariate density and, next, calculate the multivariate quantile for a conditional or marginal density of interest. Operating on the same estimated density guarantees logically consistent quantiles. In particular, the quantiles are non-crossing. We use simulation to show that we recover the true quantiles for distributions with correlation, heteroskedasticity, or asymmetry in the disturbances, and we apply our method to study heterogeneity in household expenditures.

C1858: **Partial identification of heteroskedastic structural VARs: Theory and Bayesian inference**
Presenter: **Fei Shang**, The University of Sydney, Australia
Structural vector autoregressive models are proposed in which the structural parameters are identified via a stochastic volatility process for time-varying conditional variances. Our focus is on the question of how many and what shocks are identified via heteroskedasticity. Therefore, we derive a set of parametric restrictions under which the structural matrix is partially or globally unique, and Savage-Dickey density ratios are used to assess the validity of the identification conditions. We propose a shrinkage prior distribution for conditional log-volatilities and variances that is centred on a hypothesis of homoskedasticity, which assures that the evidence for the identification of the structural shocks is provided by the data. We apply identification through heteroskedasticity to estimate the dynamic output effects of unanticipated changes in tax policy that have been identified in previous studies by exclusion restrictions as well as by using narrative measures as proxies or time-varying volatility.

C1863: **Verifying sources of identification in structural VARs using the BETEL framework**
Presenter: **Tomasz Wozniak**, University of Melbourne, Australia
A spike'n'slab prior distribution is employed to discriminate between moment conditions identifying a fiscal policy structural vector autoregression. Various sources of identification, including instrumental variables as well as symmetric and asymmetric kurtosis, are presented as moment conditions informing the estimation of the structural parameters. Exclusion restrictions are also considered. The spike'n'slab prior is used to verify these conditions within a single MCMC run. We use a three-variable system for the US fiscal policy analysis to show that the structural parameters are identified thanks to the non-normality innovations and exclusion restrictions rather than via instruments or heteroskedasticity.

CO322 Room BH (S) 1.01 Lecture Theatre 1 CRYPTOCURRENCY PRICE DYNAMICS Chair: Julien Chevallier

C0257: **Are simple technical trading rules profitable in Bitcoin markets?**
Presenter: **Michael Froemmel**, Ghent University, Belgium
Co-authors: Niek Deprez
Since its conception in 2008, the academic literature about Bitcoin has been steadily growing. Even though the presence of speculative bubbles and market inefficiencies in the bitcoin market has been studied extensively, the profitability of technical trading rules remains relatively unexplored. Using high frequency trade data of the Bitstamp exchange, we employ 75,360 commonly used simple technical trading rules to test for predictability and profitability on daily and intraday frequencies. We use a multiple hypothesis procedure that takes into account data snooping, try to mimic realistic investor behaviour in our rule selection and take into account transaction costs. Our results show that technical trading rules can outperform a simple buy-and-hold strategy in the bitcoin market out-of-sample.

C0250: **Hedging the extreme risk of cryptocurrency**
Presenter: **Johnson Owusu-Amoako**, Fayetteville State University, United States
Co-authors: Kwamie Dunbar
The attractiveness of crypto investments is highlighted by their Sharpe ratios which are generally higher than that of similarly risky equity returns (MRP). However, this remarkable level of performance comes with significant risk. 1-week Value-at-Risk (VaR) losses indicate that cryptos' potential 1-week losses were far more significant than MRPs. New evidence is provided showing that MRP is a meaningful diversifier of crypto risks. We also document that MRP reduces the downside risk of risk-averse investors at exactly the time it is needed, such as during periods of elevated levels of economic uncertainty.

C0302: **Safe assets during Covid-19: A portfolio management perspective**
Presenter: **Julien Chevallier**, IPAG Business School, France
The pandemic crisis of Covid-19 hit the financial markets like a shockwave on March 19, 2020. The aim is to capture which "safe assets" asset managers could have fled during the first wave of the pandemic. From an investment manager's perspective, candidate assets are stocks, bonds, exchange rates, commodities, gold, and (gold-backed) cryptocurrencies. Empirical tests of the Safe-Haven hypothesis are conducted, upon which the selection of assets is performed. The methodological framework hinges on the Global Minimum Variance Portfolio with Monte Carlo

simulations. The portfolio optimization routine is performed under Python. The result will inform investors about the returns that could have been achieved during such a stressful event.

C0404: **Application of machine learning models and interpretability techniques to identify the determinants of bitcoin price**
Presenter: **Jose Manuel Carbo Martinez**, Bank of Spain, Spain
Co-authors: Sergio Gorjon Rivas

Historically, the price of bitcoin has been subject to large and abrupt fluctuations, as demonstrated once again by its sudden drop following the all-time high of 68,000 in November 2021 and, more recently, on the occasion of the crypto-asset market turmoil sparked by the likes of the Terra/Luna crash or the Celsius Networks collapse. Thus, a legit question arises as to which are the determinants that influence bitcoin the most. We attempt to answer that question, using a flexible machine learning model, specifically a Long Short Term Memory (LSTM) neural network, to establish the price of bitcoin as a function of a number of economic, technological and investor attention variables. Our LSTM model replicates reasonably well the behaviour of the price of bitcoin through different periods of time. We then use an interpretability technique called SHAP to understand with are the most important features of the LSTM outcome. We conclude that the importance of the different variables in the formation of the price of bitcoin changes substantially throughout the analysed period. What's more, we also find that not only does its influence vary, but that new explanatory factors seem to appear often over time that, at least, for the most part, remain initially unknown.

CO168 Room BH (SE) 1.01 STRUCTURAL INFORMATION IN ESTIMATING ASSET PRICING MODELS Chair: Diego Ronchetti

C0260: **Earnings extrapolation and predictable stock market returns**
Presenter: **Hongye Guo**, Hong Kong University, Hong Kong

The U.S. stock market's return during the first month of a quarter correlates strongly with returns in future months, but the correlation is negative if the future month is the first month of a quarter, and positive if it is not. These effects offset, leaving the market return with its weak unconditional predictive ability known to the literature. The pattern accords with a model in which investors extrapolate announced earnings to predict future earnings, not recognizing that earnings in the first month of a quarter are inherently less predictable than in other months. Survey data support this model, as does out-of-sample evidence across industries and international markets. These results challenge the Efficient Market Hypothesis and advance a novel mechanism of expectation formation.

C0774: **Asymmetric information, information life span, and lead-lag effects on a multi-asset lagged adjustment price dynamics**
Presenter: **Iordanis Kalaitzoglou**, Audencia Business School, France
Co-authors: Hai Dang NGO, Emilios GALARIOTIS

The classic martingale-plus-noise model is extended for high-frequency price formation dynamics to accommodate pricing errors and a lagged price adjustment mechanism. The generalized price formation dynamics are motivated by (i) asymmetric information is an inherent part of the market microstructure, triggering mispricing of true underlying values; (ii) residual information has a life span, decaying over time before fully incorporated into the price process. We identify the pricing of asymmetric information and residual information as separable sources driving the price formation process. The gradual discovery of the latter gives rise to temporal lead-lag effects, which, due to Epps effects, vanish in contemporaneous time of asynchronous trading. Econometric inference on the model is built on the bridge between microstructure models and Hawkes point process theory, and enables the estimation of price lagged adjustment to be computed at every point in time. This property allows for (i) statistical measures for the efficiency of the financial market; (ii) separable estimation of lead and lag correlations at different time scales; and (iii) an efficient estimator of locally integrated covariance over any time interval. Our empirical application to a set of selected DJIA stocks supports these theoretical postulations.

C1247: **Structural estimation of nonlinear rational expectations models with recursive preferences**
Presenter: **Bart Claassen**, University of Groningen, Netherlands
Co-authors: Diego Ronchetti

The risk and the timing premia for the U.S. market from 1952 to 2019 are measured separately by consistently estimating the risk aversion and the endurance to cope with uncertainty over time of a representative agent with Epstein-Zin preferences in a Markovian setting. To minimize the risk of misspecification, we employ a novel Local Generalized Method of Moments built upon conditional moment restrictions for a nonparametric time series model, with moment functions identified by a functional equation in contraction mapping. The point estimate for the relative risk-aversion parameter is substantially lower and more plausible than previously empirically obtained through other empirical estimation techniques. We assess the finite sample performance of our estimator in a Monte Carlo study of the Bansal-Kiku-Yaron long-run risks model. We find that the method performs well in finite samples. Furthermore, we argue that this method can be applied to any functional form of recursive preferences that belongs to the Chew-Dekel class.

C1722: **GMM estimation of the long run risks model**
Presenter: **Jules Tinang**, University of Groningen, Netherlands
Co-authors: Nour Meddahi

A GMM estimation of the structural parameters of the Long Run Risk model is proposed that allows for the separation between the consumer optimal decision frequency and the frequency by which the econometrician observes the data. Our inference procedure is also robust to weak identification. The key finding is that the Long Run Risk model adapts well to the data, and the use of the estimated parameters to simulate the model enables us to improve some quantitative predictions of the model. We also show that the commonly used methods of statistical inference, such as the bootstrap (parametric or block bootstrap), might be misleading in this case since they imply an undercoverage of the true confidence interval.

CO460 Room BH (SE) 1.05 TOPICS IN APPLIED ECONOMETRICS Chair: Emese Lazar

C1860: **Copula-based multivariate mixture normal GARCH**
Presenter: **Alexander Georges Gretener**, University of Kiel, Germany
Co-authors: Markus Haas, Marc Paolella

Univariate mixture normal GARCH models have been shown to provide accurate density and risk forecasts for financial returns. Current multivariate extensions of this model class are only applicable to low-dimensional return vectors. To overcome this limitation, a novel model coupling a univariate mixture of normal GARCH specifications for the conditional marginals with a mixture of Gaussian copulas for the dependence structure is proposed, resulting in a highly flexible multivariate return process which is also applicable to high-dimensional portfolios. The properties of the model and estimation issues are discussed. An application to the returns of the Dow Jones Industrial Average stocks shows that the model provides plausible disaggregation of the conditional multivariate distribution and delivers competitive risk forecasts and risk-based portfolio allocations.

C1688: **Cross-sectional variation of forward equity risk premium**
Presenter: **Shuyuan Qi**, Central University of Finance and Economics, China
Co-authors: Emese Lazar, Radu Tunaru

The forward equity risk premium (FERP) is constructed, which is a function of investors' risk aversion coefficient and the forward-looking return moments, for stocks listed in S&P 100 from 2006 to 2020. The risk aversion coefficient of investors and forward-looking return moments are

estimated based on the stochastic volatility model with price jumps. The initial portfolio analysis shows that the companies with higher/lower FERP also reveal higher/lower returns in the future. In addition, we also find that the aggregate FERP is significantly and positively linked to future stock market returns. We then investigate the power of the FERP in predicting real economic activity growth.

C0634: **Option implied discount rates**
Presenter: **Andrei Stancu**, Newcastle University Business School, United Kingdom
Co-authors: Chardin Wese Simen, Davide Avino
A methodology is developed to recover the term structure of discount rates implied by S&P 500 index option prices. On average, implied discount rates are 36 basis points higher than the corresponding Treasury rates. This spread varies over time and is significantly related to variables that proxy for the business cycle, credit and liquidity risks. We investigate the information content of the term structure of discount rates and find that long-term implied discount rates are mainly informative about future short-term discount rates (rather than the term premium).

C0338: **Measuring climate risk in finance**
Presenter: **Emese Lazar**, University of Reading, United Kingdom
Co-authors: Jingqi Pan, Shixuan Wang
A framework is provided to estimate the climate downside risk (value-at-risk and expected shortfall) that could be attributed to climate risk factors for equity portfolios. We compare the ratio of climate risk to total risk in various equity sectors, identifying the sectors where the climate risk factors contribute most to the total risk.

CO104 Room BH (S) 2.03 TIME SERIES ECONOMETRICS Chair: Josu Arteche

C1035: **Local Whittle estimation in time varying long memory series**
Presenter: **Josu Arteche**, University of the Basque Country, Spain
The memory parameter is usually assumed to be constant in traditional long memory time series. We relax this restriction by considering the memory of the series to be a function that depends on a finite number of parameters, allowing it to vary smoothly with time. A Local Whittle-type estimator of these parameters is proposed. Its consistency and asymptotic normality are shown for locally stationary and non-stationary long memory processes, where the spectral behaviour is only locally restricted around the origin. Its good finite sample performance is shown in a Monte Carlo and two empirical applications, supporting its benefits over the fully parametric Whittle estimator.

C1101: **Directional predictability tests**
Presenter: **Carlos Velasco**, Universidad Carlos III de Madrid, Spain
Co-authors: Weifeng Jin
New tests of predictability for non-Gaussian sequences are proposed that may display general nonlinear dependence in higher-order properties. We test the null of martingale difference against parametric alternatives, which can introduce linear or nonlinear dependence as generated by ARMA and all-pass restricted ARMA models, respectively. We also develop tests to check for linear predictability under the white noise null hypothesis parameterized by an all-pass model driven by martingale difference innovations and tests of non-linear predictability on ARMA residuals. Our Lagrange Multiplier tests are developed from a loss function based on pairwise dependence measures that identify the predictability of levels. We provide asymptotic and finite sample analysis of the properties of the new tests and investigate the predictability of different series of financial returns.

C0304: **Modelling and forecasting minimum and maximum temperatures in the Iberian Peninsula**
Presenter: **Esther Ruiz**, Universidad Carlos III de Madrid, Spain
Co-authors: Carlos Vladimir Rodriguez Caballero, Gloria Gonzalez-Rivera
A novel methodology is proposed to model and forecast intervals of minimum and maximum temperature based on fitting state space models to center and log-range temperature. In doing so, we allow the center and log-range temperature to be related and to obtain measures of the uncertainty associated with estimates of the temperature trend and dispersion. The methodology is first implemented separately to intervals of minimum and maximum temperature observed monthly in four locations in the Iberian peninsula chosen to represent different climate conditions. Namely, we consider temperatures in Barcelona, Coruna, Madrid and Seville. Second, given that, at each location, center and log-range temperatures are shown to be unrelated, we fit a multivariate dynamic factor model to extract potential commonalities among center (log-range) temperatures observed at a large number of locations in the Iberian peninsula.

C1576: **Seasonal data and global warming**
Presenter: **Antonio Montanes**, University of Zaragoza, Spain
The evolution of temperature worldwide is analyzed using data supplied by the Climate Research Unit. We employ information from 165 stations, constituting a complete monthly database for 1900-2020. The use of a recent methodology proposed leads us to reject the null hypothesis of convergence for most of the months, which can be understood as the lack of a single pattern of behavior. The exception is the case of January. For the rest of the months, we can find the existence of several convergence clubs. The number and the composition of these estimated convergence clubs greatly vary. If we examine the evolution of the average temperature of each club, we can observe the presence of a positive trend in most of the estimated convergence clubs. The magnitude of the trend again varies across the convergence clubs. We should also note that we have also found downward trend behaviors for some estimated clubs of October. All these results confirm once more the heterogeneity of the warming process, questioning its globality and suggesting the existence of several Local Warming processes.

CO528 Room BH (SE) 2.05 ADVANCES IN QUANTITATIVE FINANCE AND INSURANCE Chair: Asmerilda Hitaj

C1262: **Lambda quantile regression and its application to finance**
Presenter: **Ilaria Peri**, Birkbeck-University of London, United Kingdom
A rigorous foundation of Lambda-quantile regression is provided. Lambda-quantiles have been introduced as a generalization of the notion of quantiles which maintain the basic structure of the usual definition but add more flexibility in the choice of the threshold lambda. We generalize the classical quantile regression in an extended framework that encompasses both conditional quantiles and Lambda quantiles. In particular, we focus on the linear case in analogy with the usual quantile regression. We conduct an empirical exercise where we compare two different strategies for forecasting Lambda quantiles of the Standard & Poor's 500 index. Finally, we present the results of the backtesting experiment.

C0467: **Evaluating compound sums through Panjer's formula and discretization of continuous random distributions**
Presenter: **Alessandro Barbiero**, Universita degli Studi di Milano, Italy
The accurate evaluation of compound sums is an important task in actuarial science and operational risk management. The total claims amount that a non-life insurance company has to pay in a specific period of time can be modeled as $S_N = X_1 + ... + X_N$, with N being the number of occurring claims and X_i the $i-$th claim size, $i = 1, ..., N$; the X_i are assumed to be iid random variables, typically continuous, and N is assumed to be independent of the X_i. While determining the first integer moments of S_N (when they exist) is easy, the evaluation of the whole distribution based on convolution is, in general, not analytically viable and is computationally demanding if addressed numerically. Approximations to the Gaussian and translated Gamma distributions can be employed. Alternatively, a recursive approach for the determination of the single probabilities of S_N is available (Panjer's recursive formula); however, a special type of distribution for the number of claims N and, more importantly, for the claim size

X_i is required, which limits the range of application in practice. We show how discretizing the (continuous) claim size X_i on a lattice, according to some criterion for determining the relevant probabilities, and then applying Panjer's formula, can lead to acceptable and computationally feasible approximations of the distribution of S_N.

C0522: **Qualitative robustness of set-valued value-at-risk**
Presenter: **Elisa Mastrogiacomo**, Insubria University, Italy
Co-authors: Giovanni Crespi

Risk measures are defined as functionals of the portfolio loss distribution, thus implicitly assuming the knowledge of such a distribution. However, in practical applications, the need for estimation arises and with it, the need to study the effects of misspecification errors, as well as estimation errors on the final conclusion. We focus on the qualitative robustness of a sequence of estimators for set-valued risk measures. These properties are studied in detail for two well-known examples of set-valued risk measures: the value-at-risk and the maximum average value-at-risk. Our results illustrate, in particular, that estimation of set-valued value-at-risk can be given in terms of random sets. Moreover, we observe that historical set-valued value-at-risk, while failing to be sub-additive, leads to a more robust procedure than alternatives such as the maximum likelihood average value-at-risk.

C0386: **ALM under distributional uncertainty: From DSP to DRO optimal pension fund management**
Presenter: **Asmerilda Hitaj**, University of Insubria, Italy

The focus is on a second pillar defined benefit (DB) occupational pension fund (PF) asset-liability management (ALM) problem from the perspective of a PF manager delegated to pay benefits to the employees, the PF members by a company, and the sponsor who is also funding the pension plans. The pension fund collects the contributions from the sponsor and pays the benefits to the passive members. We do not consider the possibility of covering the fund through an insurance company (which certain systems, is compulsory). The PF manager's objective is to determine an investment strategy that allows the fund to cover its liabilities while minimizing the cost of funding, given by the contributions paid by the sponsor and the deficit between liabilities and total asset value at the end of the time horizon. We take into account uncertainty over members' lifetime and asset returns. In particular, the LeeCarter model is considered for the survival probabilities and the Nelson/Siegel one for the yield curve. The ALM problem is formulated in constant monetary values to immunize the impact of inflation. To solve the problem, we propose a distributionally robust stochastic optimization (DRSO) approach and analyze how the choice of metric affects the worst-case distribution and the out-of-sample performance of the solution.

CO132 Room BH (SE) 2.09 MIDAS AND ZOMBIES IN MACROECONOMICS Chair: Etsuro Shioji

C0765: **Forecasting GDP growth using stock returns in Japan: A factor-augmented MIDAS approach**
Presenter: **Hiroshi Morita**, Hosei University, Japan

Asset prices reflect expectations of future economic conditions. We use the property of asset prices, especially stock prices, to forecast the GDP growth rate in Japan. For optimal use of the rich time-series and cross-sectional information of stock prices, we combine MIDAS (mixed-data sampling) regression and factor analysis to examine which dimensions of information contribute to the accuracy of the GDP growth rate forecast. Our results show that the use of factors significantly improves forecast accuracy and that extracting factors from a broader set of stock prices further improves accuracy. This highlights the important role of cross-sectional stock market information in forecasting macroeconomic activity.

C0875: **Nowcasting Japanese GDP using text data and machine learning**
Presenter: **Mototsugu Shintani**, University of Tokyo, Japan

A nowcasting analysis of Japanese GDP using text data and machine learning is conducted. We employ a machine learning approach because the estimation of mixed-data sampling (MIDAS) models without parameter restriction, as well as models with text information, involves high dimensional data. Based on the unrestricted MIDAS model with 15 monthly hard macroeconomic indicators and newspaper articles, we find that the model with news data outperforms the model using only hard data, especially during the period of COVID-19 crisis.

C1118: **Structural determinants of credit market tightness and the zombie firm share**
Presenter: **Philip Schnattinger**, Bank of England, United Kingdom
Co-authors: Masahige Hamano, Francesco Zanetti

A tractable model linking aggregate bank lending and heterogeneous production lines of non-financial firms is presented. Building on the dynamic stochastic general equilibrium product variety models, a frictional financial market supplying non-financial firms with credit is integrated into the framework. This extension renders credit market tightness, on which the probability of a firm successfully entering and producing positively depends, a non-trivial function of interest rates, productivity, competition, love for varieties, and the current states of active and inactive production lines producing varieties. Credit market tightness, wages, as well as the costs of entering and exiting productive and lending activities, are, in turn, shown to determine the share of zombie firms in an economy. The tractable model provides an explanation for the variation of the share of zombie firms in Europe and has wide applicability describing not only the determinants of capital-flushed economies exhibiting firm-zombiefication, but also capital-starved economies with insufficient credit market lending.

C0768: **Responses of households' expected inflation to oil prices and the exchange rate: Evidence from daily data**
Presenter: **Etsuro Shioji**, Hitotsubashi University, Japan

Our recent experiences suggest that the household sector's inflation expectations can shift very quickly, in an environment of rapidly developing inflation. The question is if utilizing daily indicators could help predict where those expectations are heading, using Japanese data. For that purpose, we consider several indicators that might affect people's perceptions of the current state of inflation. They include the exchange rate, world oil prices, domestic gasoline prices, daily data on grocery prices at supermarkets and drug stores (called CPINOW), and the index of geopolitical risks. Those variables are incorporated into a Mixed Data Sampling (MIDAS) model that is designed to predict monthly changes in inflation expectations of households captured in the Consumer Confidence Survey. As a result, CPINOW and, to some extent, world oil prices are found to be significant. The magnitudes of their coefficients, however, are rather small.

CO138 Room BH (SE) 2.10 RECENT ADVANCES IN FINANCIAL ECONOMETRICS AND EMPIRICAL ASSET PRICING Chair: Yuqian Zhao

C1148: **Monitoring the housing market in real-time with high-dimensional predictors**
Presenter: **Shanglin Lu**, University of International Business and Economics, China
Co-authors: Lajos Horvath, Zhenya Liu, Vincent Yao

A sequential detection procedure for dynamic linear models is employed to monitor the structural breaks in the housing market in real time. We incorporate 127 macroeconomic variables as the driving factors in modeling the changes in the log housing prices to mitigate the estimation bias caused by omitted variables. The principal components are used to shrink the high-dimensional macroeconomic factor space when the training sample size of the monitoring procedure is small. We find that those superstar cities in the U.S. experienced another amplification mechanism phase during the COVID-19 period.

C0537: **Change point detection in the distribution of errors in dynamic linear models**
Presenter: **Shixuan Wang**, University of Reading, United Kingdom
Co-authors: Lajos Horvath, Zhenya Liu, Yaosong Zhan

A new test procedure is developed for detecting changes in the distribution of errors in dynamic linear models. Under the null hypothesis, the distribution of errors remains the same, while there are multiple changes in the distribution of errors under the alternative. Our procedure is based on the cumulative sum (CUSUM) process that compares the empirical distribution functions of the residuals in the first part observations and the whole sample. We derive the asymptotic properties of the proposed test statistics. Monte Carlo simulations show that the proposed test has good size control and high power. We provide two empirical applications for GDP and inflation forecasting.

C0550: **The dynamics of storage costs**
Presenter: **Lazaros Symeonidis**, University of Essex, United Kingdom
Co-authors: Andrei Stancu, Chardin Wese Simen, Lei Zhao

It is documented that the monthly storage cost of oil averages 0.50% of the spot price and varies over time. We decompose the basis, defined as the ratio of the spread between the futures and spot prices over the spot price, into the storage cost (scc) and the adjusted convenience yield (acyc) channels. The scc dominates the mean of the basis and accounts for nearly half of its variations. We show that the scc predicts future inventory growth and is the main conduit through which the predictive power of the basis for oil spot returns arises.

C1089: **A Knockoff filter approach to asset allocation**
Presenter: **Arman Hassanniakalager**, University of Bath, United Kingdom

The purpose is to investigate using the Knockoff filter as a novel to optimise long-only equity portfolios by controlling false discoveries and cross-sectional dependence among US equities. Three conventional strategies of value investing, growth investing, and investing in companies by largest market capitalisation are considered benchmark strategies. Using the FamaFrenchCarhart factor analysis, the benefits of the Knockoff regression in generating excess returns alpha and mitigating factor-related risks embedded in common investment strategies are quantified.

CO742 Room BH (SE) 2.12 CLIMATE CHANGE ECONOMETRICS AND FINANCIAL MARKETS Chair: Luca De Angelis

C1609: **Bayesian augmentation for financial network stability and climate stress testing**
Presenter: **Regis Gourdel**, WU Vienna, Austria

Financial networks with granular portfolio information have become a staple of the academic and regulatory literature, with financial interlinkages recognised for their importance in financial stability. However, in applications, network-based methods present inherent flaws that are seldom addressed. This includes unequal data coverage, an insufficient estimation of uncertainty around the results, or unrealistic simulations of alternative and future states of these networks. The latter point is especially important for climate change-related simulations, where the evolution of portfolios is key in assessing the likely impact of future shocks. Building on previous data completion techniques, a framework is designed to perform Bayesian sampling of financial networks, which allows for data augmentation when the network can be partially recovered at some points in time. This contributes to addressing the uncertainty issues currently observed and allows determination by experts of stronger priors that can help the model perform better in stress testing future states of the network.

C1987: **Temperature and growth: A panel mixed frequency VAR analysis using NUTS2 data**
Presenter: **Fabio Parla**, University of Palermo, Italy
Co-authors: Andrea Cipollini

The effects of an increase in temperature levels on regional economic activity are studied using data observed for 225 EU NUTS2 regions over the period 1981-2019. The regional economic growth is proxied by the real GVA growth rate (available at an annual frequency), while we use the level of temperature recorded during four seasons as a meteorological variable. The joint interaction between economic growth and seasonal temperature is modelled through a Panel VAR fitted to data sampled at a different frequency. The structural form of temperature shocks is identified by imposing a recursive ordering of the variables, with the temperature ordered before the proxy of economic activity. The empirical results show that an increase of 1 degree Celsius in the level of annual temperature is associated with a reduction in regional economic growth (-0.3 percent is the cumulative response of GVA after a two-year horizon). Moreover, we find that on impact, summer and, to a lesser extent, winter contribute the most to the decrease in economic activity. Results from a subsample analysis (i.e., 2000-2019) reveal a more detrimental effect of an increase in temperature on economic activity. Finally, the heterogeneity in the response across NUTS2 is investigated by splitting the sample into climate and income groups.

C1789: **Identification of climate risk shocks: A proxy-SVAR approach**
Presenter: **Giovanni Angelini**, University of Bologna, Italy
Co-authors: Luca De Angelis

The evaluation and measurement of the effects of climate change on macroeconomic outcomes is one of the current key challenges in modern economic literature and has recently received considerable attention. Climate change is largely a macroeconomic issue, for it requires the analysis of the global and long-run consequences of the dynamic exchange between natural systems and human activity to be given center stage. We investigate the macroeconomic effects of changes in climate physical risk by using the distribution of extreme weather events (extreme temperatures, heavy rainfall, drought, high wind, and sea level) as a proxy variable in a (proxy) structural vector autoregressive analysis.

C1861: **Now you see it, now you do not: Pricing of climate risk and confusion in stock markets**
Presenter: **Luca De Angelis**, University of Bologna, Italy
Co-authors: Irene Monasterolo

Achieving the climate targets requires both massive new low-carbon investments and divestments from carbon-intensive assets. There is a growing investors' appetite for sustainable assets (e.g. ESG, green bonds), but the lack of standardized classification generates uncertainty on the definition of what is "green". Indeed, the existing literature does not agree on whether markets react to climate risk and whether and how they price green/brown factors. This represents the main knowledge gap for investors and financial regulators. We investigate how European stock markets price climate risk by evaluating different taxonomies for defining "green" and "brown" portfolios. In particular, we compare more traditional taxonomies, such as industry-based or greenhouse emissions disclosure, with more sustainability-oriented taxonomies, such as ESG, Climate Policy Relevant Sectors (CPRS Granular), and the EU taxonomy. Results confirm that different taxonomies lead to very different evidence on asset financial performance.

CC748 Room BH (SE) 1.06 FINANCIAL ECONOMETRICS II Chair: Vincenzo Candila

C0356: **A comparison of probabilities of default inferred from option prices and credit default swaps**
Presenter: **Ana Monteiro**, University of Coimbra, Portugal
Co-authors: Antonio Santos

Credit default swaps and option prices could reveal market expectations about stock price behavior and the probability of default. In order to estimate this probability of default, parametric models are applied, namely a delta-lognormal density and a mixture of two lognormal densities augmented with a probability of default. The resulting optimization problems incorporate various restrictions in order to guarantee proper results. The performance of the Merton distance to default (DD) model, which is based on Merton's option pricing model, to compute the probability of default is also analyzed. The models are tested by calibrating them to credit default swap and option prices from technological firms.

C0411: **Regime-specific exchange rate predictability**

Presenter: **Marco Kerkemeier**, University of Hagen, Germany
Co-authors: Joscha Beckmann, Robinson Kruse-Becher

Explaining exchange rate behaviour is a long-standing puzzle in international finance. The link between exchange rates and corresponding economic fundamentals has been elusive. Predictability highly depends on the choice of predictors, forecast horizon, sample period, chosen models and the applied evaluation metrics. We consider the most established benchmarks in the exchange rate literature with recent upcoming research on the relationship between uncertainty and exchange rates. We implement a threshold predictive regression framework with transition variables capturing uncertainty and a broad range of other variables. This setting is applied to nine major currencies against the US Dollar. The predictors we consider cover a wide range of measures, including uncovered interest rate parity, purchasing power parity, monetary fundamentals, Taylor rules, yield curve factors, stock markets and order flow data. In addition, expectation and uncertainty measures derived from the financial market and survey data, as well as sentiment measures, are adopted. Our results demonstrate regime dependence in the sense that predictability is only observed during times of high uncertainty, while there is no predictability under medium and low levels of uncertainty. Additionally, we observe strong co-movement between the predictability of different currencies.

C0300: **Explaining long-term bond yields synchronization dynamics in Europe**

Presenter: **Oscar Fernandez**, Vienna University of Economics and Business, Austria
Co-authors: Jesus Crespo Cuaresma

The aim is to examine the factors that contribute robustly to explaining sovereign yield synchronization dynamics in the European Monetary Union. Using a time-varying measure of (long-term) government bond yields synchronization rates and Bayesian Model Averaging (BMA) methods, we show that turbulent economic times (recession and ZLB periods specifically) shape the association between lagged values of the synchronization rates and current ones. This result also holds when comparing the commonly named PI(I)GS countries relative to the others. Overall, synchronization rates are found to be highly persistent. Contrary to the yield spreads literature, we find that economic fundamentals describing fiscal positions (in levels) are not able to predict synchronization rates robustly. When analyzing how synchronization rates of economic fundamentals are related to the dependent variable of interest, we find that inflation synchronization rates are robustly associated with the sovereign yield synchronization rates. This effect is also dependent on whether the economy finds itself in the ZLB period or not, and whether the country belongs to the GIIPS category.

C1836: **Nonlinear scalar BEKK**

Presenter: **Bilel Sanhaji**, University Paris 8, France

A nonlinear conditional covariance model with five scalars is proposed and the asymptotic theory of the quasi-maximum likelihood estimator is developed. We propose Lagrange Multiplier and Likelihood Ratio tests for nonlinearity in conditional covariances in multivariate GARCH models. We also show asymptotic properties through Monte Carlo simulations, and provide empirical illustrations.

Sunday 18.12.2022 10:25 - 12:05 Parallel Session H – CFE-CMStatistics

EI013 Room Safra Lecture Theatre GRAND CHALLENGES AND ADVANCES IN BAYESIAN COMPUTATION Chair: David Rossell

E0170: **Bayesian inference with R-INLA: The road ahead**
Presenter: **Haavard Rue**, KAUST, Saudi Arabia

Recent methodology progress is discussed, which concerns INLA and its R-package R-INLA. First, the use of the variational form of a Bayes theorem is discussed. This result frames the variational inference scheme methodologically within approximate Bayesian inference, and allows us to do a highly accurate correction to improve the current estimates. We will show how to do a low-rank mean and variance correction within the R-INLA framework. Secondly, we will discuss our effort to improve the parallel performance for HPC, using both OpenMPand SIP. This includes new algorithms for improving numerical gradients, a parallel line-search algorithm in the BFGS optimisation, and a new dense-matrix re-implementation of INLA.

E0171: **Monte Carlo and variational methods: Bridging the gap**
Presenter: **Christian Andersson Naesseth**, University of Amsterdam, Netherlands

The goal of inference is to reach conclusions based on evidence and reasoning; to use data and models to find patterns and answer questions. Practical problems often result in situations where exact inference is intractable and we must resort to approximations. The two main paradigms for approximate inference are sampling-based methods, also known as Monte Carlo methods, and optimisation-based methods, also known as variational methods. We will explore the interplay between Monte Carlo and variational methods, focusing on recent work combining MCMC and variational inference using the forward Kullback-Leibler divergence.

E0188: **Robust generalised Bayesian inference for intractable likelihoods**
Presenter: **Chris Oates**, Newcastle University, United Kingdom
Co-authors: Francois-Xavier Briol, Takuo Matsubara, Jeremias Knoblauch

Generalised Bayesian inference updates prior beliefs using a loss function, rather than a likelihood, and can therefore be used to confer robustness against possible mis-specification of the likelihood. We consider generalised Bayesian inference with a Stein discrepancy as a loss function, motivated by applications in which the likelihood contains an intractable normalisation constant. In this context, the Stein discrepancy circumvents evaluation of the normalisation constant and produces generalised posteriors that are either closed form or accessible using the standard Markov chain Monte Carlo. On a theoretical level, we show consistency, asymptotic normality, and bias-robustness of the generalised posterior, highlighting how these properties are impacted by the choice of Stein discrepancy. Then, we provide numerical experiments on a range of intractable distributions, including applications to kernel-based exponential family models and non-Gaussian graphical models.

EO701 Room S-2.25 THE STEIN METHOD AND STATISTICS Chair: Robert Gaunt

E1243: **Normal approximation for the posterior in exponential families**
Presenter: **Adrian Fischer**, Universita libre de Bruxelles, Belgium
Co-authors: Gesine Reinert, Robert Gaunt, Yvik Swan

Under suitable regularity conditions, the asymptotic normality of the posterior distribution is a fundamental result in Bayesian statistics, often referred to as the Bernstein-von Mises Theorem. In particular, it follows that the contribution of the prior distribution to the posterior distribution becomes negligible for large sample sizes n. We use Stein's method to obtain explicit bounds to quantify the multivariate normal approximation of the posterior distribution in multi-parameter exponential family models. We provide bounds of order $n^{-1/2}$ in the total variation (and thus Kolmogorov) and Wasserstein distances. Moreover, we apply our general bounds to several examples from exponential families, including Poisson likelihood with gamma prior, multinomial likelihood with Dirichlet prior, and the normal distribution with unknown mean and variance with normal-inverse gamma prior. These bounds are of the expected order $n^{-1/2}$ and have an explicit dependence on the parameters of the prior distribution and sufficient statistics of the data from the sample, and thus provide insight into how these factors affect the quality of the normal approximation. The performance of the bounds is also assessed with simulations.

E0658: **Wasserstein distance error bounds for the normal approximation of the maximum likelihood estimator**
Presenter: **Robert Gaunt**, The University of Manchester, United Kingdom

Explicit Wasserstein distance error bounds between the distribution of the multi-parameter MLE and the multivariate normal distribution are presented. Our general bounds are given for possibly high-dimensional, independent and identically distributed random vectors, and are of the optimal order with respect to the sample size n. In deriving these bounds, we make use of recent advances from Stein's method literature concerning optimal order Wasserstein distance bounds in the multivariate central limit theorem. As concrete examples, we use our general bounds to obtain Wasserstein distance error bounds for the normal/multivariate normal approximation of the MLE for the exponential distribution under canonical parameterisation and the normal distribution under canonical parametrisation.

E1962: **Stein method, algebra and statistics**
Presenter: **Ehsan Azmoodeh**, University of Liverpool, United Kingdom

Let $d \geq 1$. Consider target probability distributions of the form $Y = h(N_1,...,N_d)$ where here $(N_1,...,N_d)$ is a d-dimensional standard Gaussian vector and, h is a polynomial in d variables. We introduce the novel notion of an algebraic polynomial Stein operator and show that in dimension $d = 1$, any polynomial Stein operator is, in fact, algebraic. Furthermore, we discuss – from a non-commutative algebra viewpoint – in details, the class of polynomial Stein operator associated with the standard Gaussian distribution, denoted by $PSO(N)$. N stands for the one-dimensional standard Gaussian random variable. We show, among many other findings, that $PSO(N)$ is a principal right-ideal of the first Weyl algebra generated by the so-called divergence operator. We will discuss applications in mathematical statistics.

E0401: **Tuning-free Stein variational gradient descent**
Presenter: **Christopher Nemeth**, Lancaster University, United Kingdom

Stein variational gradient descent (SVGD) has become a popular inference technique in statistics and machine learning to sample from intractable distributions. Using a kernelised version of Stein's method, SVGD provides a deterministic sampling algorithm that iteratively transports a set of particles to progressively approximate a given distribution, usually a Bayesian posterior distribution. This is achieved through gradient-based updates constructed to decrease the Kullback-Leibler divergence within a function space optimally. Like many gradient-based algorithms, the efficiency of SVGD is tied to the choice of step-size parameter. We will introduce a tuning-free version of SVGD based on parameter-free convex optimisation and show that this new algorithm is competitive against vanilla SVGD and enjoys many of the same theoretical properties.

EO130 Room S-1.06 RECENT ADVANCES IN CLUSTERING OF MIXED-TYPE DATA Chair: Marta Nai Ruscone

E1491: **Mixed data distances**
Presenter: **Michel van de Velden**, Erasmus University Rotterdam, Netherlands
Co-authors: Carlo Cavicchia, Alfonso Iodice D Enza, Angelos Markos

In many statistical methods, distance plays an important role. For instance, data visualization, classification and clustering methods require

quantification of distances among objects. How to define such distances depends on the nature of the data and the problem at hand. For the distance between numerical variables, in particular in multivariate contexts, there exist many definitions that depend on the actual observed differences between values. For categorical data, defining a distance is more complex as the nature of such data prohibits straightforward arithmetic operations. However, various specific measures have been introduced that can be used to quantify observed differences in categorical data. For mixed data, aggregate distances can be constructed by taking a (weighted) sum of the distances. We consider several definitions for mixed variable distances and show how to implement them efficiently.

E1708: **Recent results in validating and benchmarking mixed-type clustering**
Presenter: **Gero Szepannek**, Stralsund University of Applied Sciences, Germany
Co-authors: Rabea Aschenbruck

A straightforward extension of the well-known k means clustering algorithm for mixed-type data is given by k prototypes as implemented in the clustMixType R package. Nonetheless, in the recent past, several other algorithms have been proposed. An important challenge (not only in clustering) consists in model selection. For clustering, this covers, in particular, the number of clusters but also the selection of variables or the appropriate algorithm. Benchmarking studies may help to provide guidance on these decisions. For the purpose of clustering mixed-type data, so far, only a few benchmarking results are available. Challenges are given by setting up appropriate simulation designs or cluster validation. The aim is to give an overview and discussion of existing work as well as recent results, which may help to explore the landscape of mixed-type clustering algorithms.

E0930: **MDGMM and MIAMI: Towards flexible and interpretable models for mixed data**
Presenter: **Robin Fuchs**, CNRS, France
Co-authors: Denys Pommeret, Samuel Stocksieker, Cinzia Viroli

Modeling mixed data remains a challenging task due to the heterogeneous nature of the variables (continuous, ordinal, categorical, binary, count). Taking advantage of the recently introduced Deep Gaussian Mixture Models and Generalized Linear Latent Variable Models, we propose two models able to cluster and generate synthetic mixed data. The resulting multi-layer models, namely the Mixed Deep Gaussian Mixture Model (MDGMM) and MIxed data Augmentation MIxture (MIAMI), explicitly handle the different variable types and learn a continuous and low-dimensional latent representation of the data. This latent representation captures the dependence structure in the data and is used to perform clustering and generate synthetic data. The models are completed by visual diagnostic tools, architecture selection methods, and dedicated initialization procedures. Benchmarking the MDGMM and MIAMI with state-of-the-art models on several UCI datasets, the approaches have proven to deliver solid performance, model flexibility, and result interpretability.

E1542: **Clustering mixed-type data via the KAMILA algorithm**
Presenter: **Marianthi Markatou**, University at Buffalo, United States

Despite the existence of a large number of clustering algorithms, clustering mixed measurement scale data, that is, interval (continuous) and categorical (nominal and/or ordinal) scale data, remains a challenging problem. We first review the literature on this topic and show that most of the current clustering methods for mixed-scale data suffer from at least one of two central challenges: 1) they are unable to equitably balance the contribution of continuous and categorical scale variables without strong parametric assumptions; 2) they are unable to properly handle data sets in which only a subset of variables are related to the underlying cluster structure of interest. We then develop KAMILA (KAY-means for MIxed LArge data), a clustering method that addresses (1) and, in many situations, (2) without requiring strong assumptions. We next discuss MEDEA (Multivariate Eigenvalue Decomposition Error Adjustment), a weighting scheme that addresses (2) even in the face of a large number of uninformative variables. We study the theoretical aspects of our methods and demonstrate their performance using Monte Carlo simulations and real data sets.

EO665 Room S-1.22 ADVANCES AND CHALLENGES IN ACCELERATED LIFE TESTING Chair: Maria Kateri

E0621: **Maximum likelihood inference based on censored samples from the geometric distribution**
Presenter: **Anna Dembinska**, Warsaw University of Technology, Faculty of Mathematics and Information Science, Poland
Co-authors: Krzysztof Jasinski

When high-quality and long-life products are tested, we need special time-effective methods to gain knowledge about their reliability. One method to accelerate life testing is to use Type-II right censoring. During an experiment in which Type-II right censoring is applied, n items with independent and identically distributed lifetimes are placed on a test and the experiment is terminated at the moment of the rth failure, where $r < n$ is fixed in advance. Maximum likelihood estimation for the geometric distribution based on Type-II right censored sample is considered. First, general conditions for discrete distributions guaranteeing the almost sure existence of a strongly consistent sequence of maximum likelihood estimators (MLE's) will be given. Then, the discussion will be limited to the case of the geometric distribution, and a closed-form formula for the MLE will be presented. Moreover, some finite-sample properties of the MLE of the geometric parameter will be shown in the special case when $r = 1$. In particular, its bias and mean squared error will be obtained. Finally, the results will be generalized to the situation when more than one censored sample is observed.

E0222: **Imprecise statistical inference based on the log-rank test for step stress accelerated life testing data**
Presenter: **Sultan Albalwy**, Durham University, United Kingdom
Co-authors: Frank Coolen, Jonathan Cumming

A new imprecise nonparametric statistical method is developed for step stress accelerated life testing data, where the Arrhenius link function is implemented for the data analysis. This function expresses the relationship between the lifetime and the applied stresses in terms of temperature to link the scale parameters of different stress levels. This method consists of three steps. First, it transforms failure times that occurred by different strategies of experimental settings at higher stress levels to the normal stress level. Second, it creates imprecision based on the log-rank test on the accelerating parameter for which the null hypothesis, that all failure times come from the same distribution, is not rejected. This imprecision allows for the transformation of failure times into interval values at the normal stress level, where it is assumed that these transformed failure times are not distinguishable from failure times occurring at the normal stress level. Third, nonparametric predictive inference is applied to the transformed data to provide robust predictive inference. This method leads to more imprecision if data are used from higher stress levels or in case of model misspecification. The performance of this method is evaluated by simulation studies.

E0778: **Maximum product of spacings estimator for simple step-stress model with Weibull lifetimes**
Presenter: **Nikolay Nikolov**, RWTH Aachen University, Germany
Co-authors: Maria Kateri

Accelerated life testing (ALT) experiments are widely used in reliability studies on extremely durable products having large mean times to failure. A simple step-stress ALT (SSALT) is a special class of ALT experiments that tests the units under investigation on two different conditions by changing the stress factor (e.g., temperature) at a predetermined time point of the experiment. The maximum product of spacings (MPS) technique is considered for estimating the unknown lifetime parameters of an SSALT model under censoring. The MPS estimator is defined under Type-II censoring and proved to be asymptotically equivalent to the corresponding maximum likelihood (ML) estimator. The specific case of Weibull lifetimes sharing a common shape parameter on both stress levels is studied in more detail. The MPS estimator is given as a solution of a non-linear system of equations and can be evaluated numerically, e.g., by using the Newton-Raphson algorithm. The suggested statistical procedure is applied

to an illustrative data example, while the MPS and ML approaches are compared via an extensive simulation study. In particular, approximations of the estimation bias and mean squared error are presented for various combinations of the scale and shape parameter values.

E0712: **Simple heterogeneous step-stress accelerated life testing model for lithium-ion battery aging**
Presenter: **Yao Lu**, RWTH Aachen University, Germany
Co-authors: Maria Kateri

Accelerated life testing (ALT) is widely implemented to investigate lifetime performance within a comparatively short time period in the lithium-ion battery (LIB) field. With higher-than-usual levels of stress factors, e.g., temperature, charge and discharge voltages, the aging of the tested LIB cells is sped up, inducing failures much earlier than under the user level. Step-stress ALT (SSALT) is a special case of ALT, where the stress level imposed on a unit changes gradually at pre-specified time points during the experiment. Statistical models for SSALT experiments assuming a homogeneous population and a variety of lifetime distributions, under different censoring schemes, have been extensively discussed in the literature. However, SSALT for a heterogeneous population received little attention, especially in the case of unknown group membership. Motivated by heterogeneous LIB cell aging patterns observed in practice, a simple heterogeneous SSALT model for exponential distributed lifetimes is introduced. Heterogeneity is captured through a mixture model approach. An EM algorithm is developed to derive the maximum likelihood estimates of the model's parameters under Type-II censoring and corresponding bootstrap confidence intervals are provided. In case of heterogeneity, the validity of the proposed model and its advantage over the classical SSALT model are demonstrated via simulation studies.

EO660 Room S-1.27 LATENT VARIABLE MODELS FOR COMPLEX DATA STRUCTURES Chair: Silvia Cagnone

E0800: **Generalized linear factor score regression with different methods**
Presenter: **Fan Wallentin**, Uppsala University, Sweden

Factor score regression has recently received growing interest as an alternative to structural equation modeling. However, many applications are left without guidance because of the literature's focus on normally distributed outcomes. We perform a simulation study to examine how a selection of factor scoring methods compare when estimating regression coefficients in generalized linear factor score regression. The current study evaluates the regression and correlation-preserving methods as well as two sum score methods in ordinary, logistic, and Poisson factor score regression. Our results show that scoring method performance can differ notably across the considered regression models. In addition, the results indicate that the choice of scoring method can substantially influence research conclusions. The regression method generally performs the best in terms of coefficient and standard error bias, accuracy, and empirical Type I error rates. Moreover, the regression and correlation-preserving methods mostly outperform the sum score methods.

E1276: **Modelling "don't know" responses in multi Item Latent Trait Models**
Presenter: **Maria Iannario**, University of Naples Federico II, Italy

The focus is on a general mixed item response theory (IRT) framework that allows for the differentiation of respondents with respect to the type of processes underlying item responses. The study analyses the presence of response styles in answering and don't know responses. The latter is considered qualitatively different and is modelled using an additional latent variable that captures respondents' willingness to respond. Naive approaches, based on the evaluation of the option as incorrect answer or missing value, can lead to biased measures of the latent trait(s). The mixed model is studied with simulation studies and applied to an empirical example concerning the effect of financial technology on financial inclusion.

E1314: **Alternative methods for parameter estimation in discrete latent variable models**
Presenter: **Luca Brusa**, Universita di Milano Bicocca, Italy
Co-authors: Francesco Bartolucci, Fulvia Pennoni

The Expectation-Maximization (EM) algorithm is undoubtedly one of the most widely used techniques to estimate a discrete latent variable (DLV) model. However, while it is possible to prove that this algorithm converges to a local maximum of the log-likelihood function, there is no guarantee of convergence to the global maximum of this function. We propose two modifications to the EM algorithm to tackle this serious problem. The first one incorporates a tempering scheme into the EM algorithm: the log-likelihood is initially flattened to escape local maxima and then warped back to its original shape in a gradual way. The second uses evolutionary computation to encourage more accurate parameter space exploration. The performance of the resulting tempered EM (T-EM) and evolutionary EM (E-EM) algorithms is assessed for latent class and hidden Markov models in terms of both ability to reach the global maximum and computational time; a comparison with the standard EM algorithm is carried out through an extensive Monte Carlo simulation study. We show that the proposed algorithms outperform the standard EM, significantly increasing the chance of reaching the global maximum in almost all the examined cases. This improvement remains considerable, even accounting for the inflated overall computing time.

E1369: **Markov switching stereotype logit models with two latent indicators for longitudinal ordered data**
Presenter: **Sabrina Giordano**, University of Calabria, Italy
Co-authors: Roberto Colombi

Longitudinal ordered categorical data are affected by response styles when respondents are asked to evaluate, on Likert scales, items at different time occasions and decide to use only a few of the given options of the rating scale irrespectively of the content of the item. The novelty, in the context of longitudinal ordered categorical data, is in considering simultaneously the temporal dynamics of observable ordered responses and unobservable answering behaviors, possibly influenced by response styles (RS), through a Markov switching logit model with two latent components. One component accommodates serial dependence and respondent's unobserved heterogeneity, and the other component determines the responding attitude (due to RS or no-RS). The dependence of the observable variables on covariates is modelled by a stereotype logit model with parameters varying according to the two latent indicators. Unobserved heterogeneity, serial dependence and tendency to response style are modelled through our approach on real longitudinal data collected by the Bank of Italy.

EO082 Room K0.16 STATISTICAL MODELLING OF NETWORK DATA Chair: Goeran Kauermann

E0968: **All that glitters is not gold: Relational events models with spurious events**
Presenter: **Cornelius Fritz**, LMU Munich, Germany
Co-authors: Goeran Kauermann, Marius Mehrl, Paul Thurner

As relational event models are an increasingly popular model for studying relational structures, the reliability of large-scale event data collection becomes more and more important. Automated or human-coded events often suffer from non-negligible false-discovery rates in event identification. And most sensor data is primarily based on actors' spatial proximity for predefined time windows; hence, the observed events could relate either to a social relationship or random co-location. Both examples imply spurious events that may bias estimates and inference. We propose the Relational Event Model for Spurious Events (REMSE), an extension to existing approaches for interaction data. The model provides a flexible solution for modeling data while controlling for spurious events. Estimation of our model is carried out in an empirical Bayesian approach via data augmentation. Based on a simulation study, we investigate the properties of the estimation procedure. To demonstrate its usefulness in two distinct applications, we employ this model to combat events from the Syrian civil war and student co-location data. Results from the simulation and the applications identify the REMSE as a suitable approach to modeling relational event data in the presence of spurious events.

E1039: **Disentangling homophily, community structure and triadic closure in networks**
Presenter: **Tiago Peixoto**, Central European University, Austria

One of the most typical properties of network data is the presence of homophily, i.e. the increased tendency of an edge to exist between two nodes if they share the same underlying characteristic, such as a social parameter, location, etc. Another pervasive pattern encountered is transitivity, i.e. the increased tendency to observe an edge between two nodes if they share a neighbor in common. Although these patterns are indicative of two distinct mechanisms of network formation, namely homophily and triadic closure, respectively, they are generically conflated in non-longitudinal data. This is because both processes can result in the same kinds of observation: 1. the presence of triangles, and 2. the formation of community structure. This conflation means we cannot reliably interpret the underlying mechanisms of network formation merely from the abundance of triangles or observed community structure in network data. We present a solution to this problem, consisting in a principled method to disentangle homophily and community structure from triadic closure in network data. This is achieved by formulating a generative model that includes community structure in a first instance, and an iterated process of triadic closure in a second. Based on this model, we develop a Bayesian inference algorithm that is capable of identifying which edges are more likely to be due to community structure or triadic closure, in addition to the underlying community structure itself.

E1296: **Assessing competitive balance in the English Premier League for over forty seasons using a stochastic block model**
Presenter: **Nial Friel**, University College Dublin, Ireland

Competitive balance is a desirable feature in any professional sports league and encapsulates the notion that there is unpredictability in the outcome of games as opposed to an imbalanced league in which the outcome of some games is more predictable than others, for example, when an apparent strong team plays against a weak team. We develop a model-based clustering approach to provide an assessment of the balance between teams in a league. We propose a novel Bayesian model to represent the results of a football season as a dense network with nodes identified by teams and categorical edges representing the outcome of each game. The resulting stochastic block model facilitates the probabilistic clustering of teams to assess whether there are competitive imbalances in a league. A key question, then is to assess the uncertainty around the number of clusters or blocks and consequently estimation of the partition or allocation of teams to blocks. We apply our model to each season in the English premier league from 1978/79 to 2019/20. A key finding of this analysis is evidence which suggests a structural change from a reasonably balanced league to a two-tier league which occurred around the early 2000's.

E1790: **Sampling from weighted network models when aggregate information is available**
Presenter: **Axel Gandy**, Imperial College London, United Kingdom

Many networks have weights attached to their edges, such as the number of liabilities in a financial network. One might be interested in sampling from such weighted networks given information about the aggregate amount of in-/out weights of the nodes. This could be because only such partial information about the network is available or because a conditional statistical test should be performed. We discuss several of these settings and present algorithms that can be used to sample from appropriate conditional distributions.

EO599 Room K0.18 QUANTILE METHODS AND APPLICATIONS Chair: Jayeeta Bhattacharya

E0256: **Uniform inference for conditional deconvolution estimation**
Presenter: **Stefan Hubner**, University of Bristol, United Kingdom

A uniform inference method is developed for conditional deconvolution estimators. For this, we rewrite Kotlarskis identity as a system of linear conditional moment inequalities by approximating the space of conditional distributions of the mismeasured function of interest by a multi-dimensional separable Hermite sieve basis. Separability has the advantage that partial application of the Fourier transform again forms a separable, orthonormal basis of the resulting space of conditional characteristic functions, which are implicitly defined by the conditional moments. By appropriately choosing instrument functions, we can then rewrite the conditional moments as unconditional ones without losing identifying power. Based on the latter, using generalised moment selection, we obtain uniform confidence bands for the mismeasured function of interest.

E0265: **Renewable quantile regression analysis of streaming data**
Presenter: **Keming Yu**, Brunel University, United Kingdom

Online updating is an important statistical method for the analysis of big data arriving in streams due to its ability to break the storage barrier and the computational barrier under certain circumstances. The quantile regression, as a widely used regression model in many fields, faces challenges in model fitting and variable selection with big data arriving in streams. Renewable-optimized objective functions for regression parameter estimation and variable selection in a quantile regression are proposed. The proposed methods are illustrated using current data and the summary statistics of historical data. Theoretically, the proposed statistics are shown to have the same asymptotic distributions as the standard version computed on an entire data stream with the data batches pooled into one data set, without additional conditions. Both simulations and data analysis are conducted to illustrate the finite sample performance of the proposed methods.

E0348: **Elicitation of elasticity of intertemporal substitution, risk and time preferences**
Presenter: **Jose Olmo**, Universidad de Zaragoza, Spain
Co-authors: Gabriel Montes-Rojas, Antonio Galvao, Luciano De Castro

The elicitation of the elasticity of intertemporal substitution (EIS), discount factor, and risk attitude parameters in dynamic models is of central importance to economics, finance and public policy. An alternative method is suggested to elicit and estimate these three parameters using experimental data jointly. We employ a new model based on dynamic quantile preferences, where individuals maximize the stream of future tau-quantile utilities, for tau in (0, 1). These preferences are simple, dynamically consistent, and monotonic. In the quantile model, the risk attitude is captured by the quantile of the payoff distribution, while the EIS and the discount factor are related to the utility function describing individuals' intertemporal behavior, hence allowing for complete separability between risk, EIS and discount factor. The estimation of the parameters of interest uses a structural maximum likelihood method. Individuals' risk aversion is estimated below the median. The discount factor is marginally smaller than estimates reported in the literature, and the EIS is slightly larger than one, which suggests that utility over time is concave. The estimates for the elasticity contrast with those reported by the existing studies using observational disaggregated data, which in general, find an elasticity smaller than one.

E0890: **Asymptotic normality of quantile regression with generated variables**
Presenter: **Jayeeta Bhattacharya**, University of Southampton, United Kingdom

Linear quantile regression models are studied when regressors and/or dependent variables are not directly observed, but estimated in an initial first step and used in the second step quantile regression for estimating the quantile parameters. This general class of generated quantile regression (GQR) covers various statistical applications, for instance, the estimation of endogenous quantile regression models and triangular structural equation models, and some new relevant applications are discussed. We study the asymptotic distribution of the two-step estimator, which is challenging because of the presence of generated covariates and/or dependent variable in the non-smooth quantile regression estimator. We employ techniques from empirical process theory to find uniform Bahadur expansion for the two-step estimator, which is used to establish its functional central limit theorem. We illustrate the performance of the GQR estimator through simulations and an empirical application

EO056 Room K0.19 GRAPHICAL MARKOV MODELS I Chair: Monia Lupparelli

E0928: **On model selection of colured Gaussian graphical models for paired data**
Presenter: **Alberto Roverato**, University of Padova, Italy
Co-authors: Dung Ngoc Nguyen

The problem of learning a graphical model is considered when the observations come from two groups sharing the same variables but, unlike the usual approach to the joint learning of graphical models, the two groups do not correspond to different populations and therefore produce dependent samples. A Gaussian graphical model for paired data may be implemented by applying the methodology developed for the family of graphical models with edge and vertex symmetries, also known as coloured graphical models. Many model search algorithms require the exploration of the search space that is typically carried out by means of local moves between neighbouring models. It is, therefore, crucial to be able to rely on procedures that allow us to explore the space of models efficiently. However, the exploration of the space of coloured graphical models is much more challenging than for classical graphical models. We identify a family of coloured graphical models suited for the paired data problem and investigate the structure of the corresponding model space. More specifically, we provide a comprehensive description of the lattice structure formed by this family of models both under the model inclusion order and under a novel order that we call the twin order. We show that our novel order allows a more efficient exploration of the search space. This is then used to implement a stepwise model search procedure and an application to the identification of a brain network from fMRI data is given.

E1349: **Foundations of structural causal models with cycles and latent variables**
Presenter: **Stephan Bongers**, Delft University of Technology, Netherlands
Co-authors: Patrick Forre, Jonas Peters, Joris Mooij

Structural causal models (SCMs), also known as (nonparametric) structural equation models (SEMs), are widely used for causal modeling purposes. In particular, acyclic SCMs, also known as recursive SEMs, form a well-studied subclass of SCMs that generalize causal Bayesian networks to allow for latent confounders. We will investigate SCMs in a more general setting, allowing for the presence of both latent confounders and cycles. We show that in the presence of cycles, many of the convenient properties of acyclic SCMs do not hold in general: they do not always have a solution; they do not always induce unique observational, interventional and counterfactual distributions; a marginalization does not always exist, and if it exists the marginal model does not always respect the latent projection; they do not always satisfy a Markov property; and their graphs are not always consistent with their causal semantics. We prove that for SCMs in general each of these properties does hold under certain solvability conditions. We generalize results for SCMs with cycles that were only known for certain special cases so far. We introduce the class of simple SCMs that extends the class of acyclic SCMs to the cyclic setting, while preserving many of the convenient properties of acyclic SCMs. The aim is to provide the foundations for a general theory of statistical causal modeling with SCMs.

E0390: **Structure learning of undirected graphical models for count data**
Presenter: **Thi Kim Hue Nguyen**, University of Padova, Italy
Co-authors: Monica Chiogna

Mainly motivated by the problem of modelling biological processes underlying the basic functions of a cell -that typically involve complex interactions between genes- we present a new algorithm, called PC-LPGM, for learning the structure of undirected graphical models over discrete variables. We prove the theoretical consistency of PC-LPGM in the limit of infinite observations and discuss its robustness to model misspecification. To evaluate the performance of PC-LPGM in recovering the true structure of the graphs in situations where relatively moderate sample sizes are available, extensive simulation studies are conducted that also allow us to compare our proposal with its main competitors. Biological validation of the algorithm is presented through the analysis of two real data sets.

E1738: **Algebraic and combinatorial methods for causal model representation and selection**
Presenter: **Liam Solus**, KTH Royal Institute of Technology, Sweden

When a statistical model is defined by a family of polynomial constraints, such as a graphical model or a more general conditional independence model, tools from algebraic geometry and combinatorics can allow us to prove theorems relevant in model representation and selection. The resulting insights from these theorems can also direct us towards methods for model selection, such as algorithms for data-driven learning of causal models under assumptions weaker than the classic conditions, such as faithfulness. We will present examples of such algebraic techniques and discuss their consequences in regard to the problem of causal model selection.

EO490 Room K0.20 STATISTICAL ANALYSIS IN NON-EUCLIDEAN SPACES Chair: Xianzheng Huang

E0427: **Handling errors in circular data**
Presenter: **Charles C Taylor**, University of Leeds, United Kingdom
Co-authors: Marco Di Marzio, Stefania Fensore

Circular data are often recorded imprecisely. This may be due to instrument error, in which case we can treat the errors as being i.i.d.., and seek to recover the density function of the error-free distribution. Data which include times can also be considered as circular (for example, minutes past the hour), and these are often rounded - to the nearest minute, nearest five minutes, nearest quarter of an hour etc. Although the error-free values may be i.i.d., in this case of digit preference, the errors will no longer have the same distribution. We give examples of data which exhibit these different characteristics and investigate corresponding approaches, including deconvolution, to estimate the underlying density.

E0623: **Elliptically symmetric distributions for directional data of arbitrary dimension**
Presenter: **Zehao Yu**, University of South Carolina, United States

A class of angular Gaussian distributions is formulated that allows different degrees of isotropy for directional random variables of arbitrary dimension. Through a series of novel reparameterization, this distribution family is indexed by parameters with meaningful statistical interpretations that can range over the entire real space of an adequate dimension. The new parameterization greatly simplifies the maximum likelihood estimation of all model parameters, which in turn leads to theoretically sound and numerically stable inference procedures to infer key features of the distribution. Byproducts from the likelihood-based inference are used to develop graphical and numerical diagnostic tools for assessing the goodness of fit of this distribution in a data application. Simulation study and application to data from a hydrogeology study are used to demonstrate the implementation and performance of the inference procedures and diagnostics methods.

E0878: **Directions old and new: Palaeomagnetism and Fisher meet modern statistics**
Presenter: **Janice Scealy**, Australian National University, Australia

Most modern articles in the palaeomagnetism literature are based on statistics developed by Fisher's 1953 paper Dispersion on a sphere, which assumes independent and identically distributed (iid) spherical data. However, palaeomagnetic sample designs are usually hierarchical, where specimens are collected within sites and the data are then combined across sites to calculate an overall mean direction for a geological formation. The specimens within sites are typically more similar than specimens between different sites, and so the iid assumptions fail. We will first review, contrast and compare both the statistics and geophysics literature on the topic of analysis methods for clustered data on spheres. We will then present a new hierarchical parametric model, which avoids the unrealistic assumption of rotational symmetry in Fisher's 1953 paper Dispersion on a sphere and may be broadly useful in the analysis of many palaeomagnetic datasets. To help develop the model, we use publicly available data as a case study collected from the Golan Heights volcanic plateau. Next, we will explore different methods for constructing confidence regions for

the overall mean direction based on clustered data. Two bootstrap confidence regions that we propose perform well and will be especially useful to geophysics practitioners.

E1672: **Principal nested shape subspace analysis of molecular data**
Presenter: **Ian Dryden**, Florida International University, United States

Molecular dynamics simulations produce huge datasets of temporal sequences of molecules. It is of interest to summarize the shape evolution of the molecules in a succinct, low-dimensional representation. However, Euclidean techniques such as principal components analysis (PCA) can be problematic as the data may lie far from a flat manifold. Principal nested spheres give a fundamentally different decomposition of data from the usual Euclidean subspace-based PCA. Subspaces of successively lower dimensions are fitted to the data in a backwards manner with the aim of retaining signal and dispensing with noise at each stage. We adapt the methodology to 3D shape subspaces and provide some practical fitting algorithms. The methodology is applied to cluster analysis of peptides, where different states of the molecules can be identified. Also, the temporal transitions between cluster states are explored. Further molecular modelling tasks include resolution matching, where coarse-resolution models are back-mapped into high-resolution (atomistic) structures.

EO636 Room K0.50 HIGH DIMENSIONAL DATA ANALYTICS: TOOLS, TRICKS, TIPS AND PITFALLS Chair: Farrukh Javed

E0419: **The penalized instrumental variables methods for many invalid instruments**
Presenter: **Muhammad Qasim**, Jonkoping University, Sweden
Co-authors: Kristofer Mansson

The valid instrumental variables must not directly affect the outcome variable and not be correlated to unmeasured variables. But practically, instrumental variables (IV) are likely to be invalid. We derive a LASSO procedure for the k-class IV estimation methods in the linear IV model. The proposed method is robust for estimating the causal effect in the presence of many/weak invalid and valid instruments, with theoretical assurances on its execution. In addition, a two-step numerical algorithm is developed to estimate causal effects. The performance of the proposed penalized k-class IV methods is assessed by Monte Carlo simulation.

E1238: **Some tests of hypotheses for high-dimensional linear models**
Presenter: **Rauf Ahmad**, Uppsala University, Sweden

A statistic for testing the parameter vector in a general linear model is presented when the dimension of the parameter vector is large, i.e., when the number of columns of the design matrix may exceed the number of independent rows. The distribution of the proposed statistic, obtained under a few mild assumptions, depends on a simple function of the eigenvalues of the known, fixed design matrix. Simulations are used to show the accuracy of the proposed theory. Applications on real data are demonstrated. Some extensions of the test for other models are also considered.

E1805: **Efficient inversion of sparse positive definite matrices based on the dyadic orthogonalization algorithm**
Presenter: **Hanqing Wu**, Lund University, Sweden
Co-authors: Krzysztof Podgorski

The problem of inverting large sparse positive definite matrices is at the center of many high-dimensional statistical methods. Recently, a highly efficient dyadic algorithm has been proposed for diagonalization and inversion of the band matrices, i.e. matrices that are zero outside of a narrow band of the entries around the main diagonal. Random permutation of a band matrix creates a sparse matrix for which it is easy to find the permutation that reverses it back to the band form. Using this simple observation, we propose an algorithm that allows efficient diagonalization and thus also inversion of a class of sparse matrices that can be decomposed into a small dimensional block and a number of not connected blocks of permuted band matrices. We formulate and prove mathematical results as well as compare the efficiency of our algorithms with other existing methods of inverting high-dimensional sparse matrices.

E1938: **Singular conditional autoregressive Wishart model for realized covariance matrices**
Presenter: **Farrukh Javed**, Lund University, Sweden
Co-authors: Taras Bodnar, Gustav Alfelt, Joanna Tyrcha

Realized covariance matrices are often constructed under the assumption that the richness of intra-day return data is greater than the portfolio size, resulting in nonsingular matrix measures. However, when for example, the portfolio size is large, assets suffer from illiquidity issues, or market microstructure noise deters sampling on very high frequencies, this relation is not guaranteed. Under these common conditions, realized covariance matrices may obtain as singular by construction. Motivated by this situation, we introduce the Singular Conditional Autoregressive Wishart (SCAW) model to capture the temporal dynamics of time series of singular realized covariance matrices, extending the literature on econometric Wishart time series models to the singular case. This model is furthermore developed by covariance targeting adapted to matrices and a sector-wise BEKK-specification, allowing excellent scalability to large and extremely large portfolio sizes. Finally, the model is estimated to a 20-year long time series containing 50 stocks and to a 10-year long time series containing 300 stocks, and evaluated using out-of-sample forecast accuracy. It outperforms the benchmark models with high statistical significance and the parsimonious specifications perform better than the baseline SCAW model, while using considerably fewer parameters.

EO446 Room S0.03 BRANCHING PROCESSES: THEORY, COMPUTATION AND APPLICATIONS II Chair: Ines M del Puerto

E1387: **Martingale method for studying branching random walks**
Presenter: **Elena Yarovaya**, Lomonosov Moscow State University, Russia

A continuous-time branching random walk is considered on a multidimensional lattice, in which particles can die and produce offspring at any point on the lattice. Let the transport of particles over the lattice be given by a symmetric, homogeneous and irreducible random walk. A branching intensity at a point x at the lattice tends to zero as the norm $|x|$ tends to infinity. Moreover, an additional condition is satisfied for the parameters of the branching random walk, which guarantees exponential growth in the time of the average number of particles at each point of the lattice. Under these assumptions, we prove the limit theorem about the mean square convergence of the normalized number of particles at an arbitrary fixed point x, as the time t tends to infinity. The proof is based on the approximation of the normalized number of particles by a nonnegative martingale.

E1856: **Weak convergence results in the class of controlled branching processes**
Presenter: **Pedro Martin-Chavez**, University of Extremadura, Spain
Co-authors: Miguel Gonzalez Velasco, Ines M del Puerto

The study of functional weak limit theorems for branching processes aroused a lot of interest since 1951, when the convergence of standard Bienayme Galton Watson processes (BGWPs), suitably scaled, to a class of diffusion processes, was established. The aim is to contribute, focusing the attention on a more general class of branching processes. Concretely, we consider controlled branching processes. It is a wide family of branching processes that add the novelty with respect to BGWPs that the number of progenitors in each generation is determined by a random mechanism. Besides the theoretical motivation by itself, from a practical viewpoint, the interest in developing these results stems from the usefulness of this kind of limit theorems for sequences of branching processes in determining the asymptotic distributions of the weighted least squares estimators of the main parameters of the model. We will establish a set of sufficient conditions for the weak convergence of (suitably scaled) discrete-time, discrete-state controlled branching processes.

E1855: **Sequential MCMC methods for controlled branching processes**
Presenter: **Miguel Gonzalez Velasco**, University of Extremadura, Spain
Co-authors: Pedro Martin-Chavez, Ines M del Puerto

Controlled branching processes (CBPs) are stochastic growth population models in which the number of individuals with reproductive capacity in each generation is determined by random control functions. This kind of process is flexible enough to model the evolution of different kinds of populations, including logistic growth populations or epidemic outbreaks (at least in its exponential growth phase). We deal with the estimation of the main parameters of CBPs from a Bayesian perspective. We consider different sampling schemes, including partially observed CBPs. In all situations, we use sequential MCMC methodologies. We show the accuracy of the proposed methodology via simulated examples making use of the statistical software R.

E1258: **Modeling for coronavirus pandemic: A comparative research**
Presenter: **Maroussia Slavtchova-Bojkova**, Sofia University, Bulgaria
Co-authors: Valeriya Simeonova

Comparative analysis for the ongoing pandemics caused by Coronavirus SARS-CoV-2 will be presented based on the model of the general branching processes vs artificial neural networks models. We have developed and maintained three possible scenarios: main, optimistic and pessimistic. They are depending on the time-varying reproduction number of the epidemic. It is considered as a measure of the effectiveness of the public health interventions changing to a greater or lesser degree throughout the pandemic waves observed. The focus is given to the possibility of using different models on such data with attention to their reliability, time consuming, accuracy and aggregation/differentiation by scenarios in prognosis.

EO436 Room S0.11 ASYMPTOTIC THEORY APPLIED TO STATISTICAL COMPUTATION AND SIMULATION Chair: Nakahiro Yoshida

E0644: **Quasi-likelihood inference for Student-Levy regression**
Presenter: **Hiroki Masuda**, Kyushu University, Japan
Co-authors: Yuma Uehara

The quasi-likelihood analysis is considered for a linear regression model driven by a Student Levy process with constant scale and arbitrary degrees of freedom. The model is observed at a high frequency over an extended period, which quantitatively clarifies how the sampling frequency affects estimation accuracy. In that setting, however, joint estimation of trend, scale, and degrees of freedom does not seem to have been investigated as yet. The bottleneck is that the Student distribution is not closed under convolution, preventing us from estimating all the parameters fully based on the high-frequency time scale. To efficiently deal with the intricate nature from both theoretical and computational points of view, we propose a two-step quasi-likelihood analysis: first, we make use of the Cauchy quasi-likelihood for estimating the regression-coefficient vector and the scale parameter; then, we construct the sequence of the unit-period cumulative residuals to estimate the remaining degrees of freedom.

E0989: **Markov-switching Hawkes processes for high-frequency trade data**
Presenter: **Ioane Muni Toke**, CentraleSupelec, France

The focus is on a multidimensional point process defined by multiple Hawkes-like intensities and a switching mechanism based on a hidden Markov chain. Previous works in such a setting assume constant intensities between consecutive events. We extend the model to general piecewise-constant Hawkes excitation kernels and develop an expectation-maximization algorithm for the statistical inference of the parameters of the multiple Hawkes intensities as well as the state transition probabilities. The numerical convergence of the estimators is extensively tested on simulated data. The model is finally applied to equity transaction data on multiple financial markets in an attempt to identify meaningful market states.

E1218: **Parameter inference for partially observed linear SDEs with discrete observations**
Presenter: **Masahiro Kurisaki**, University of Tokyo, Japan

A problem of parameter estimation is considered for the state space model described by linear stochastic differential equations. We assume that an unobservable Ornstein-Uhlenbeck process drives another observable process by the linear stochastic differential equation, and these two processes depend on some unknown parameters. We construct the quasi-likelihood estimator (QMLE) of the unknown parameters and show the asymptotic properties of the estimator. Moreover, the function of YUIMA to execute our estimation on R will be discussed.

E1192: **Order estimate of functionals related to fBm and asymptotic expansion of variation of fractional SDE**
Presenter: **Hayate Yamagishi**, University of Tokyo, Japan
Co-authors: Nakahiro Yoshida

An asymptotic expansion is derived for the quadratic variation of a stochastic process satisfying a stochastic differential equation driven by a fractional Brownian motion, based on the theory of asymptotic expansion of Skorohod integrals converging to a mixed normal limit. In order to apply the general theory, it is necessary to estimate functionals that are a randomly weighted sum of products of multiple integrals of the fractional Brownian motion, in expanding the quadratic variation and identifying the limit random symbols. To overcome the difficulty, we introduce exponents by means of the weighted graphs capturing the structure of the sum in the functional, and investigate how the exponents change by the action of the Malliavin derivative and its projection.

EO580 Room S0.12 DISTRIBUTIONAL MODEL VALIDATION Chair: Bruno Ebner

E1374: **Stein goodness-of-fit tests for a new family of distributions**
Presenter: **Yvik Swan**, Universite libre de Bruxelles, Belgium
Co-authors: Bruno Ebner

A new family of distributions is introduced for which we develop bespoke goodness-of-fit tests. We illustrate our results on various family members, including many familiar distributions such as Poisson, gamma, compound Poisson, and compound geometric, as well as lesser-known distributions such as the Dickman distribution.

E1169: **Testing multivariate normality in the presence of missing data**
Presenter: **Bojana Milosevic**, University of Belgrade, Serbia
Co-authors: Danijel Aleksic

Missing data are a very common problem in practice. Therefore providing an adequate methodology for statistical inference is of great importance for a wide scientific community. We focus on the problem of testing the hypothesis of multivariate normality in the presence of different missingness mechanisms. In particular, the focus will be on the impact of different imputation algorithms on some of the popular tailor-made normality tests - their size and powers under commonly used alternatives, and their comparison with the complete-case approach. Finally, potential directions for future research will be discussed.

E1475: **Nonparametric distribution function estimation and goodness-of-fit testing**
Presenter: **James Allison**, Northwest University, South Africa
Co-authors: Jaco Visagie, Elzanie Bothma

When analysing lifetime data in the presence of censoring, one is often required to estimate the distribution function of the lifetimes non-parametrically. The most popular estimator used for this purpose is the Kaplan-Meier estimator. For values larger than the sample maximum,

two different assumptions are commonly used for this estimator in the statistical literature. The first is to set the value of the estimate to one, while the second is to use the value of the estimate at the sample maximum when estimating the tail of the distribution function. We illustrate the profound effect of these assumptions on the sizes and powers of goodness-of-fit tests in both the i.i.d. case and mixture cure model.

E1482: **Stein characterizations of non-normalized discrete probability distributions and their applications in statistics**
Presenter: **Steffen Betsch**, Karlsruhe Institute of Technology, Germany
Co-authors: Bruno Ebner, Franz Nestmann

From the distributional characterizations that lie at the heart of Stein's method, explicit formulae are derived for the mass functions of discrete probability laws that identify those distributions. These identities are used to develop tools for the solution of statistical problems. The characterizations, and hence the applications built on them, do not require any knowledge about normalization constants of the probability laws. To demonstrate that our statistical methods are sound, we provide comparative simulation studies in goodness-of-fit and parameter estimation problems. In particular, we discuss parameter estimation for discrete exponential-polynomial models, which, generally, are non-normalized.

EO498 Room S0.13 STATISTICS OF HIGH-FREQUENCY DATA I Chair: Carsten Chong

E0605: **How and when are high-frequency stock returns predictable**
Presenter: **Yacine Ait-Sahalia**, Princeton University, United States
Co-authors: Jianqing Fan, Lirong Xue, Yifeng Zhou

The predictability of ultra-high-frequency stock returns and durations to relevant price, volume and transactions events is studied using machine learning methods. We find that contrary to low-frequency and long-horizon returns, where predictability is rare and inconsistent, predictability in high-frequency returns and durations is large, systematic and pervasive over short horizons. We identify the relevant predictors constructed from trades and quotes data and examine what determines the variation in predictability across different stocks' own characteristics and market environments. Next, we compute how the predictability improves with the timeliness of the data on a scale of milliseconds, providing a valuation of each millisecond gained. Finally, we simulate the impact of getting an (imperfect) peek at the incoming order flow, a look-ahead ability that is often attributed to the fastest high-frequency traders, in terms of improving the predictability of the following returns and durations.

E0448: **Stock co-jump networks**
Presenter: **Yi Ding**, The University of Macau, China
Co-authors: Yingying Li, Guoli Liu, Xinghua Zheng

A Degree-Corrected Block Model with Dependent Multivariate Poisson edges (DCBM-DMP) is proposed to study stock co-jump dependency. To estimate the community structure, we extend the SCORE algorithm and develop a Spectral Clustering On Ratios-of-Eigenvectors for networks with Dependent Multivariate Poisson edges (SCORE-DMP) algorithm. We prove that SCORE-DMP enjoys strong consistency in community detection. Empirically, using high-frequency data of S&P 500 constituents, we construct two co-jump networks according to whether the market jumps and find that they exhibit different community features than GICS. We further show that the co-jump networks help in stock return prediction.

E1474: **Systematic jump risk**
Presenter: **Jean Jacod**, Sorbonne université, France
Co-authors: Huidi Lin, Viktor Todorov

In a factor model for a large panel of N asset prices, a random time S is called a "systematic jump time" if it is not a jump time of any of the factors, but nevertheless is a jump time for a significant number of prices: one might, for example, think that those S's are jump times of some hidden or unspecified factors. The aim is to test whether such systematic jumps exist and, if they do, to estimate a suitably defined "aggregated measure" of their sizes. The setting is the usual high-frequency setting with a finite time horizon T and observations of all prices and factors at times iT/n for $i = 0, \ldots, n$. We suppose that both n and N are large, and the asymptotic results (including feasible estimation of the above aggregate measure) are given when both go to infinity, without imposing restrictions on their relative size. In an empirical application, we document the existence of systematic jumps and further show that the associated risk commands a nontrivial risk premium.

EO220 Room Virtual R01 ADVANCES IN FUNCTIONAL DATA ANALYSIS AND NONPARAMETRIC REGRESSION Chair: Yuko Araki

E1520: **Fast multilevel functional principal component analysis**
Presenter: **Erjia Cui**, Johns Hopkins University, United States
Co-authors: Luo Xiao, Ciprian Crainiceanu, Ruonan Li

Fast multilevel functional principal component analysis (fast MFPCA) is introduced, which scales up to high dimensional functional data measured at multiple visits. The new approach is orders of magnitude faster than and achieves comparable estimation accuracy with the original MFPCA. Methods are motivated by the National Health and Nutritional Examination Survey (NHANES), which contains minute-level physical activity information of more than 10000 participants over multiple days and 1440 observations per day. While MFPCA takes more than five days to analyze these data, fast MFPCA takes less than five minutes. A theoretical study of the proposed method is also provided. The associated function mfpca.face() is available in the R package refund.

E1521: **Sup-norm convergence of deep network estimator for nonparametric regression with corrected adversarial training**
Presenter: **Masaaki Imaizumi**, The University of Tokyo, Japan

The purpose is to study the stability of an estimator for the nonparametric regression problem by deep neural networks and adversarial training. Several studies show that deep neural networks give estimators for the nonparametric problem which theoretically outperform conventional estimators in a specific setting. A limitation of the estimator by deep networks is stability: its convergence is measured by a restricted class of norms. We consider adversarial training for deep networks and develop an estimator for the nonparametric regression problem. We investigate its efficiency by the minimax optimization scheme and derive several convergence rates with different norms. We also discuss an application based on the result.

E1553: **On weak convergence of recovered functional data**
Presenter: **Yoshikazu Terada**, Osaka University; RIKEN, Japan
Co-authors: Masaki Sasaki

In functional data analysis, subjects are represented as smooth curves, and observed data consist of observations of random curves at discrete time points. In some cases, such as classification problems, we need to recover individual smooth curves from discretely observed data, and the properties of the recovered curves play important roles. We focus on the weak convergence of the empirical distribution of the recovered smooth curves. The existing results require the independence of recovered individual curves. When the data are observed on a dense grid of time points for each subject, we may recover the smooth curves independently. However, when data are observed not so densely, we often use the reconstruction method based on functional principal component analysis (FPCA). In this case, the recovered curves are not independent anymore. Thus, we establish the weak convergence of the empirical measure of the individual curves recovered by FPCA.

E1888: **Functional mediation analysis with model selection**
Presenter: **Yuko Araki**, Tohoku University, Japan

Mediation analysis has been developed when data are curves or images. We first estimate a set of functions to represent data using basis expansions. We have selected which basis function should be used among several candidate functions and how many basis functions should be used. For the

mediation analysis model, a few types of functional regression models are used to represent the direct and indirect effects. Further, the proposed model selection criterion provides a powerful approach to identifying causal relationships. We conducted a simulation study and real data examples to examine the performance of the proposed model.

EO597 Room Virtual R02 INNOVATIVE APPROACHES FOR UNSUPERVISED CLASSIFICATION METHODS Chair: Carlo Cavicchia

E0791: **An EM algorithm for penalized mixed-effects multitask learning: A general framework for regularizing MLM Models**
Presenter: **Andrea Cappozzo**, Politecnico di Milano, Italy
Co-authors: Francesca Ieva, Giovanni Fiorito

Linear mixed modeling is a well-established technique widely employed when observations possess a grouping structure. Nevertheless, this standard methodology is no longer applicable when the learning framework encompasses a multivariate response and high-dimensional predictors. To overcome these issues, a penalized estimation scheme based on an expectation-maximization (EM) algorithm is devised. Any penalty criteria for fixed-effects models can be conveniently incorporated into the fitting process. We employ the novel methodology for creating surrogate biomarkers of cardiovascular risk factors, such as lipids and blood pressure, from whole-genome DNA methylation data in a multi-center study. The described method performs better than state-of-the-art alternatives, both in terms of predictive power and bio-molecular interpretation of the results.

E1366: **Density-peak clustering of graphs**
Presenter: **Riccardo Giubilei**, Luiss Guido Carli, Italy

Graph clustering, intended as the task of grouping observations that are in the form of graphs, is attracting increasing attention thanks to its various applications. These include identifying similar brain networks for ability assessment or disease prevention, as well as clustering different snapshots of the same network evolving over time to identify similar patterns or abrupt changes. However, there are no well-established procedures for performing this task. The method proposed here builds upon the density-peak algorithm (DP), which is a mode-based clustering approach that identifies cluster centers as data points being surrounded by neighbors with lower density and far away from points with higher density. The new method: 1) inherits the favorable properties of the DP; 2) overcomes two main limitations of the DP, namely, the unstable density estimation and the absence of an automatic procedure for selecting cluster centers; 3) can be applied to graphs of any type, provided that a sensible distance between observations is selected. Numeric applications, including an empirical analysis whose goal is clustering brain connectomes to distinguish between patients affected by schizophrenia and healthy controls, show the adequate performance of the proposed approach.

E1297: **Mixed-type data spectral clustering with variable specific distances**
Presenter: **Alfonso Iodice D Enza**, Universita di Napoli Federico II, Italy
Co-authors: Cristina Tortora, Francesco Palumbo

At the core of the spectral clustering approach is the decomposition of the graph Laplacian matrix, a weighted kernel transformation of the pairwise distances/dissimilarities between the observations at hand. It follows that the definition of the distance/dissimilarity matrix is crucial and, in the case of non-continuous and/or mixed datasets, non-obvious nor trivial. A straightforward solution is: to compute pairwise Euclidean distances for the continuous variables, and Hamming distances for the non-continuous variables; to define a general distance matrix via a convex combination of the two matrices previously obtained. The weight of the convex combination dictates the influence of the continuous and categorical variables on the clustering solution. Using Euclidean distances on standardised continuous variables is a reasonable choice; instead, considering the simple matching for the categorical variables is simplistic. We consider a set of association-based, variable-specific, distances and dissimilarities, to define a custom Laplacian matrix suitable for the spectral clustering of mixed data. In particular, we propose a data-driven approach to select, for the considered variable, the most appropriate distance/dissimilarity: the combination of distance/dissimilarity of choice is the one providing the best spectral clustering solution according to a suitable metric.

E0893: **Mixed data convex clustering**
Presenter: **Carlo Cavicchia**, Erasmus University Rotterdam, Netherlands

Clustering analysis is an unsupervised learning technique widely used for information extraction. Current clustering algorithms often face instabilities due to the non-convex nature of their objective function. The class of convex clustering methods does not suffer from such instabilities and finds a global optimum for the clustering objective. Whereas convex clustering has previously been established for single-type data, real-life data sets usually comprise both numerical and categorical, or mixed, data. Therefore, we introduce the mixed data convex clustering (MIDACC) framework. We implement this framework by developing a dedicated subgradient descent algorithm. Through numerical experiments, we show that, in contrast to baseline methods, MIDACC achieves near-perfect recovery of both spherical and non-spherical clusters, is able to capture information from mixed data while distinguishing signal from noise, and has the ability to recover the true number of clusters present in the data. Furthermore, MIDACC outperforms all baseline methods on a real-life data set.

EO558 Room BH (S) 2.02 DEPENDENCY IN BAYESIAN MIXTURE MODELS AND BEYOND Chair: Jim Griffin

E0485: **A marginalization approach to local regression and clustering with variable-dimension covariates**
Presenter: **Fernando Quintana**, Pontificia Universidad Catolica de Chile, Chile
Co-authors: Garritt Page, Matthew Heiner

Incomplete covariate vectors are known to be problematic for estimation and inferences on model parameters, but their impact on prediction performance is less understood. We develop an imputation-free method that builds on a random partition model admitting variable-dimension covariates. Cluster-specific response models further incorporate covariates via linear predictors, facilitating the estimation of smooth prediction surfaces with relatively few clusters. The response models are analytically marginalized according to the pattern of missing covariates, yielding a local regression with internally consistent uncertainty propagation that utilizes only one set of coefficients per cluster. Aggressive shrinkage of these coefficients crucially regulates uncertainty due to missing covariates. The method allows in- and out-of-sample prediction for any missingness pattern, even if the pattern in a new subject's incomplete covariate vector was not seen in the training data. We demonstrate the model's effectiveness for nonlinear prediction under various circumstances, including non-random missingness mechanisms, by comparing it with other recent methods for variable-dimension regression on synthetic and real data.

E0576: **Capturing correlated clusters using mixtures of latent class models**
Presenter: **Gertraud Malsiner-Walli**, WU Vienna University of Economics and Business, Austria

Latent class models are useful for model-based clustering of multivariate categorical data. These models rely on the conditional independence assumption, i.e., it is assumed that the categorical variables are independent given the cluster membership. In case this assumption is violated, the latent class model will lead to more components being fitted than there are clusters in the data, as each cluster distribution will be approximated by several components. A crucial issue is then the identification of the clusters from the components as the likelihood is completely invariant to the assignment of components to a cluster. Within a Bayesian framework, we propose a suitable specification of priors for the latent class model to identify the clusters in multivariate categorical data where the independence assumption is not fulfilled. Each cluster distribution is approximated by a latent class model, leading overall to a mixture of latent class models. The Bayesian approach allows to identify the clusters and fits their cluster distributions using a one-step procedure, thus not relying on two-step procedures usually pursued in frequentist analysis where first a semi-parametric approximation using a latent class model with many components is performed and then components are combined to form clusters. We

provide suitable estimation and inference methods for the mixture of latent class models and illustrate the performance of this approach on artificial and real data.

E1179: **Mixtures of normalized nested compound random measures and their application**
Presenter: **Riccardo Corradin**, University of Nottingham, United Kingdom
Co-authors: Federico Camerlenghi, Andrea Ongaro

Dependent random measures have been studied extensively over the past decades. Among the possible choices, a remarkable strategy to define a vector of dependent random measures is given by the family of compound random measures. We first provide a posterior characterization for vectors of normalized compound random measures, which allows us to perform efficient conditional posterior inference. Further, we embed a vector of compound random measures in a nested structure, obtaining a model which induces ties among different dimensions of the vector of random measures. As a byproduct, by convoluting a kernel function with these nested models, we can perform clustering of distributions and observations simultaneously. Upon a posterior representation of compound random measures, we can derive a conditional sampling strategy to perform conditional inference also for the nested case. Our studies are motivated by an ecological problem, where we aim to cluster provinces in Lombardy based on their distributions of the daily concentration of particulate matter, but also to properly quantify the risk of exceeding a threshold imposed by the European Union's regulations.

E1404: **Transformed scaled process priors for generalized Indian Buffet processes**
Presenter: **Mario Beraha**, Universita di Torino, Italy

In trait allocation models, each observation displays a collection of traits corresponding to a (usually nonnegative integer) association level. In the Bayesian nonparametric framework, data are modeled as conditionally i.i.d. stochastic processes (termed generalized Indian buffet processes) whose law depends on a random measure. Traditionally, completely random measures are employed as prior, but, as shown previously, this leads to a rather simplistic predictive structure. Namely, the probability of a new observation displaying new (unobserved) traits depends on the observed sample only through its cardinality. In the context of latent feature models, this problem was faced recently by proposing to use scaled processes as priors instead of completely random measures. We extend this framework to the more general latent trait models. The proposal is based on a suitable transformation of SPs and which are recovered as a special case. We characterize the marginal, posterior, and predictive distribution induced by the proposed class of prior processes in trait allocation models, showing in particular that this choice leads to a richer predictive structure. We consider the case of Bernoulli, Poisson, and negative binomial distributed traits as examples.

EO677 Room BH (S) 2.05 DEVELOPMENTS AND APPLICATIONS OF APPROXIMATE BAYESIAN COMPUTATION Chair: Veronica Ballerini

E1657: **ABC in a compartmental model for smoking habit dynamics**
Presenter: **Alessio Lachi**, University of Florence, Italy
Co-authors: Cecilia Viscardi, Michela Baccini

Smoking is the main risk factor for many common chronic diseases. In order to describe the evolution of the population's smoking habits in Tuscany (Italy), we have developed a compartmental model. Compartmental models assume that at any given time, the population is divided into groups called "compartments" and that individuals can move from one to the other following simple probabilistic rules described by a system of differential equations. The population is divided into Never, Current, and Former smokers, with ex-smokers allowed to relapse smoking. The likelihood function of the model is complex to evaluate analytically, and the model requires specific estimation methods. We investigated the use of approximate Bayesian computation (ABC), a class of likelihood-free algorithms, as a tool to perform inference in both a frequentist and a fully Bayesian context. From a frequentist perspective, we used ABC as a method for the "stochastic search" for optimal parameters, using the deterministic version of the model and compared the results with those obtained from standard optimization algorithms. From a fully Bayesian perspective, we used ABC to sample from the joint posterior distribution of model parameters. Our results suggest that ABC is a powerful method to provide solutions in complex compartmental models.

E1679: **Species abundance estimation in the presence of record linkage errors: An ABC approach**
Presenter: **Davide Di Cecco**, University of Rome La Sapienza, Italy
Co-authors: Andrea Tancredi

The phenomenon of one-inflation in zero-truncated count data has been receiving increasing attention in capture-recapture and species abundance literature. The phenomenon manifests itself as an abundance of singletons (units captured exactly once), which suggests the necessity of explicitly modeling a mechanism for this deviation. We distinguish two possible causes for one-inflation: the erroneous inclusion of spurious units, and missed links in a preliminary record linkage step. Note that we do not have access to the record linkage procedure, but only to the aggregated count data. While the first mechanism can easily be estimated both in frequentist and Bayesian context (via simple Gibbs-based MCMC), we found record linkage errors to be hard to investigate outside a Bayesian ABC approach. As a matter of fact, missing links errors are sometimes tacitly treated as spurious data. We implemented an ABC algorithm for various count distributions and applied it to the estimation of the number of microbial species in literature data.

E1718: **Likelihood-free transport Monte Carlo**
Presenter: **Cecilia Viscardi**, Univeristy of Florence, Italy
Co-authors: Dennis Prangle

Approximate Bayesian computation (ABC) is a class of methods for drawing inferences when the likelihood function is unavailable or computationally demanding to evaluate. Importance sampling and other algorithms using sequential importance sampling steps are state-of-art methods in ABC. Most of them get samples from tempered approximate posterior distributions defined by considering a decreasing sequence of ABC tolerance thresholds. Their efficiency is sensitive to the choice of an adequate proposal distribution and/or forward kernel function. We present a novel ABC method addressing this problem by combining importance sampling steps and optimization procedures. We resort to Normalising Flows (NFs) to optimize proposal distributions over a family of densities to transport particles drawn at each step towards the next tempered target. Therefore, the combination of sampling and optimization steps allows tempered distributions to get efficiently closer to the target posterior. Finally, we show the performance of our method on examples that are a common benchmark for likelihood-free inference.

E1724: **Box-ABC for likelihood-free inference**
Presenter: **Elena Bortolato**, University of Padova, Italy
Co-authors: Laura Ventura

Among the determining factors for accurately approximating posterior distributions in Approximate Bayesian Computation (ABC), the choice of a suitably small threshold ε, for bounding the distance between observed and simulated data, plays a crucial role. This decision also prescribes the computational effort of the procedure. In fact, the value of ε is inversely related to the necessary number of simulations to be performed for obtaining sufficiently large Monte Carlo draws from the posterior. Furthermore, the process of tuning the tolerance may be time demanding. We propose to modify ABC algorithms, by defining acceptance rules that circumvent the use of distance functions and the choice of threshold parameters. The method implicitly makes use of a pseudo-likelihood that inherits some desirable properties from confidence distributions. We study the asymptotic behaviour of the methodology and the computational efficiency in different regimes.

EO086 Room BH (SE) 2.05 ADVANCES IN BAYESIAN COMPUTATION TECHNIQUES I Chair: Siew Li Linda Tan

E0917: **Bayesian inference for partial orders**
Presenter: **Kate Lee**, The University of Auckland, New Zealand
The focus is on the underlying social rank-order which constrains the ranks within lists and the order is not necessarily a complete order. The social rank order is not necessarily homogeneous in time either. The goal is to estimate the evidence for evolving social order by evolving partial order over time. Lists are uniform linear extensions of the underlying partial order in the model. We specify a marginally consistent prior stochastic process over partial orders, driven by a multivariate latent stochastic process with a hyperparameter controlling the partial order depth distribution.

E1022: **Fast and accurate variational inference in mixed models**
Presenter: **Luca Maestrini**, The Australian National University, Australia
Mixed models with fixed and random effects are widely used to analyse longitudinal and multilevel data that can potentially be high dimensional and have a variety of measurement scales. In these complex settings, variational approximations may facilitate fast approximate inference for the parameters of mixed models. We explain how streamlined solutions to sparse matrix problems can be used for making fast variational Bayes inference for models with a high number of random effects, where Bayesian computation is typically hindered by the size of design matrices. Accuracy is also a crucial aspect in variational approximations, especially for models with non-Gaussian responses. We show that resampling methods can offer a valid remedy to the potential inaccuracy of variational approximations and illustrate the use of bootstrap for variational inference in mixed models.

E0738: **Adversarial Bayesian simulation**
Presenter: **Yuexi Wang**, University of Chicago, United States
Co-authors: Veronika Rockova
In the absence of explicit or tractable likelihoods, Bayesians often resort to approximate Bayesian computation (ABC) for inference. ABC s bridged with deep neural implicit samplers based on generative adversarial networks (GANs) and adversarial variational Bayes. Both ABC and GANs compare aspects of observed and fake data to simulate from posteriors and likelihoods, respectively. We develop a Bayesian GAN (B-GAN) sampler that directly targets the posterior by solving an adversarial optimization problem. B-GAN is driven by a deterministic mapping learned on the ABC reference by conditional GANs. Once the mapping has been trained, iid posterior samples are obtained by filtering noise at a negligible additional cost. We propose two post-processing local refinements using (1) data-driven proposals with importance reweighing, and (2) variational Bayes. We support our findings with frequentist-Bayesian results, showing that the typical total variation distance between the true and approximate posteriors converges to zero for certain neural network generators and discriminators. Our findings on simulated data show highly competitive performance relative to some of the most recent likelihood-free posterior simulators.

E1849: **The development of "variationalDCM" an R package performing variational Bayesian estimation for DCMs**
Presenter: **Keiichiro Hijikata**, The University of Tokyo, Japan
Co-authors: Motonori Oka, Kazuhiro Yamaguchi, Kensuke Okada
"variationalDCM" is provided, an R package that performs recently developed variational Bayesian (VB) estimation methods for Diagnostic Classification Models (DCMs). DCMs are a class of latent variable models used to reveal students' current knowledge status and applied to various educational tests. Despite increasing attention to DCMs, there are few software programs available on the Internet for DCMs, and, to the best of our knowledge, there do not seem to be any programs that estimate parameters by VB methods. VB methods are techniques for approximating the posterior distribution of parameters in the Bayesian estimation framework. They are characterized by their fast calculation time compared to Markov Chain Monte Carlo methods which are usually used for Bayesian estimation. This package enables fast estimations by VB methods for various DCMs and can be applied to large-scale data. We implement five functions that estimate model parameters for 1) deterministic input noisy AND gate (DINA) model, 2) saturated DCM, 3) multiple-choice DINA model, and 4) hidden Markov type longitudinal general DCM and estimates 5) Q-matrix for DINA model.

EO110 Room K2.31 (Nash Lec. Theatre) ADVANCEMENTS IN SPATIAL AND SPATIO-TEMPORAL MODELS Chair: Maria Michela Dickson

E0965: **Identifying geographical configurations of industries by regional concentration and spatial polarization**
Presenter: **Diego Giuliani**, University of Trento, Italy
Co-authors: Maria Michela Dickson, Giuseppe Espa, Flavio Santi
The most popular areal-based indices of geographical concentration of industries (such as Gini and Herfindahl/Hirschman indices) do not take into account the spatial positions of areas. As a consequence, they are insensitive to the spatial order of areas and do not control for neighboring effects. In order to deal with this issue, different indices that allow measuring the level of geographical concentration of industry while controlling for the spatial distance among areas have been recently developed. According to the principle that a single index is not able to depict an industry entirely with respect to both geographical concentration and spatial connections, an alternative approach is proposed to measure both aspects jointly and classify industries into relevant geographical configurations.

E1181: **Spatial quantile regression for modelling the impact of digital transformation on European regional economic growth**
Presenter: **Alfredo Cartone**, University of Chieti-Pescara, Italy
Co-authors: Luca Di Battista, Paolo Postiglione
The study of regional economic growth still raises many issues, and, in the last years, a strain focused on quantile regression, showing the advantages of modelling the entire conditional distribution and going beyond average. However, the modelling of spatial effects in quantile regression is still to explore. We add on the literature on economic growth on two sides. First, we consider quantile regression at the regional level by comparing different spatial quantile specifications, particularly Spatial Lag and Spatial Durbin Model. Second, we measure the effect of digital transformation by considering IT variables on the growth level of European NUTS2. We focus on the higher level of the conditional distribution meant as the conditional quantile useful to measure regions' performance. Finally, we estimate direct and indirect impacts to observe spatial spillovers across the whole conditional distribution.

E1284: **Joint geostatistical modelling of lymphatic filariasis antigenaemia and microfilariae prevalence**
Presenter: **Claudio Fronterre**, Lancaster University, United Kingdom
Co-authors: Emanuele Giorgi
Lymphatic filariasis (LF) is a mosquito-borne neglected tropical disease targeted for global elimination by 2020. In recent years, the mapping of LF has been greatly facilitated by the use of simple and rapid detection tests based on the immuno-chromatographic test (ICT), which avoids the need to collect blood at night and the time-consuming preparation and examination of blood slides. Even if the scientific output of interest is the prevalence of microfilaraemia, the number of mapping surveys that measure MF is low, and it is decreasing due to the diffusion and cost-effectiveness of ICT tests. We develop a geostatistical model that exploits the abundance of ICT prevalence surveys and the relationship between ICT and MF prevalence to predict microfilaraemia prevalence at unobserved locations. We use LF data from West-Africa to show how this modelling framework can be used to produce relevant output for control and elimination programmes.

E1385: **Dynamic probit models with network interdependence and unobserved heterogeneity**
Presenter: **Michaela Kesina**, University of Groningen, Netherlands
Co-authors: Peter Egger
A Bayesian estimation framework is proposed for panel-data sets with binary dependent variables where a large number of cross-sectional units is observed over a short period of time, and cross-sectional units are interdependent. Our estimation approach enables accounting for various forms of dynamic relationships and different types of cross-sectional dependence. These features should make the approach interesting for applications in many empirical contexts. The estimation approach is outlined. Its suitability is illustrated through simulation examples. An application is provided to study dynamic exporting patterns among Chinese firms.

EO062 Room K2.40 RECENT ADVANCES IN FUNCTIONAL DATA ANALYSIS Chair: Michelle Carey

E1947: **Functional covariance estimation for partially observed functional data**
Presenter: **Uche Mbaka**, University College Dublin, Ireland
Co-authors: Michelle Carey
Most approaches for reconstructing curves from partially observed functional samples depend on estimating the covariance surface. We propose an approach for estimating the covariance surface from sparse and fragmented functional observations using a finite element method called finite element sigma (FE-Sigma). When applied to densely observed functional data, the finite element sigma produces a covariance surface with the desirable property of being positive definite, which leads to an accurate estimation of the inverse covariance surface (i.e., the precision surface). By including this improved estimate of the covariance surface in the principal analysis by conditional expectation (PACE) approach, we can obtain improved estimates of the principal functional components and use these functions to estimate full curves. We compare the results of curve estimation accuracy using a simulated dataset and apply the new approach to a real dataset (somatic cell scores of dairy cows from selected Irish farms).

E1966: **Sparse estimation within the Pearson system, with an application to financial market risk**
Presenter: **Michelle Carey**, Univerity College Dublin, Ireland
Co-authors: Christian Genest, James Ramsay
Pearson's system is a rich class of models that includes many classical univariate distributions. It comprises all continuous densities arising as solutions to a differential equation involving a vector of parameters. Estimating a Pearson density is challenging as small parameter variations can induce wild changes in the shape of the corresponding density. It is shown how to estimate the parameters and the corresponding density effectively through a penalized likelihood procedure involving differential regularization. The approach combines a penalized regression method and a profiled estimation technique. Simulations and an illustration with S&P 500 data suggest that this method can improve market risk assessment substantially through value-at-risk and expected shortfall estimates that outperform those currently used by financial institutions and regulators.

E0456: **Principal differential analysis: A review with extensions**
Presenter: **Edward Gunning**, University of Limerick, Ireland
Co-authors: Giles Hooker
Principal Differential Analysis (PDA) is a technique used to estimate time-varying linear ordinary differential equations (ODEs) from functional data. We review PDA and investigate some extensions. First, we extend the PDA model to include stochasticity in the form of smooth random disturbances to the ODE and examine the implications of this additional source of variation. Second, we investigate whether PDA can be used to approximate nonlinear time-invariant ODE models.

E1896: **A functional logistic regression model with non-independent functional variables**
Presenter: **Cristhian Leonardo Urbano Leon**, Universidad de Granada, Spain
Co-authors: Manuel Escabias, Ana Maria Aguilera
A proposal is presented to extend the functional logistic regression models, which model a binary scalar response variable from a functional predictor to the case when the functional variables are not independent but paired, i.e., the same functional variable under different experimental conditions. We assume that the curves of the functional predictor and the parameter function of the model belong to the same finite-dimensional subspace of the space L_2 of square-integrable functions over the same closed real interval. This approach allows the use of basis expansion methods for the treatment of functional data. The proposal is contextualized with an application to biomechanical functional data that records the position of joints (angles they form with certain axes) with respect to the total walking cycle.

EO646 Room K2.41 ADVANCES IN STATISTICAL MODELING WITH NEURAL NETWORKS Chair: David Ruegamer

E0327: **Approximating spatial extreme value processes with deep learning**
Presenter: **Reetam Majumder**, North Carolina State University, United States
Co-authors: Brian Reich, Benjamin Shaby
The Intergovernmental Panel on Climate Change has projected an increased frequency of hydroclimatic extremes in its Sixth Assessment released in 2021. Quantifying how the probability and magnitude of extreme flooding events are changing is key to mitigating their impacts. While climate data are inherently spatially dependent, spatial models such as Gaussian processes (GP) do not adequately model extreme events, and theoretically-justified extreme value analysis models like the max-stable process (MSP) give intractable likelihoods. We propose a process mixture model which specifies spatial dependence in extreme values as a convex combination of a GP and an MSP, using a deep learning model to approximate the likelihood. We propose a unique computational strategy where a feed-forward neural network is embedded in a density regression model to approximate the conditional distribution at one spatial location given a set of neighbors. We then use this univariate density function to approximate the joint likelihood for all locations by way of a Vecchia approximation. The process mixture model is used to analyze changes in annual maximum streamflow within the US over the last 50 years, and is able to detect areas which show increases in extreme streamflow over time.

E1128: **Sparse regularization of neural networks using a Hadamard parametrization-based optimization transfer approach**
Presenter: **Chris Kolb**, LMU Munich, Germany
Co-authors: David Ruegamer
Neural networks are becoming an increasingly popular framework for estimating complex or high-dimensional regression models, allowing scaling up models to very large data sets using stochastic gradient descent (SGD). Incorporating sparsity into neural networks has shown to be difficult due to the non-smooth nature of the added penalty term, typically requiring specialized optimization routines such as projected gradient or coordinate descent methods. Instead, a method for inducing sparsity in neural networks with ℓ_p regularization ($0 < p \leq 1$) is presented that is amenable to conventional first-order optimizers such as SGD or Adam. This is achieved by solving an equivalent surrogate problem, obtained by applying a Hadamard product reparametrization to the model parameters, under which smooth and strongly convex ℓ_2 regularization (or weight decay) induce non-smooth and potentially non-convex ℓ_p regularization in the original parametrization. This optimization transfer approach can be readily extended to structured sparsity problems, yielding $\ell_{p,q}$ regularization of the original parameters for $0 < p < q < 2$.

E1363: **Uncertainty quantification in semi-structured distributional regression models**
Presenter: **Daniel Dold**, Institute for Optical Systems (IOS) Hochschule Konstanz, Germany

Co-authors: Oliver Duerr, Beate Sick, David Ruegamer

Many applications require predicting a conditional outcome distribution based on semi-structured input data, such as the combination of images and tabular data. Recent deep semi-structured distributional regression models combine deep neural networks employed for complex data and statistical regression models for structured data. This approach combines the high predictive power of neural networks and the interpretability of statistical models. While deep semi-structured distributional regression models enable to account for the aleatoric uncertainty, it is still not clear how to capture the model uncertainty (epistemic uncertainty), which is of pivotal interest in many applications like out-of-distribution detection or active learning. A natural way of capturing epistemic uncertainty is to use Bayesian approaches. Many Bayesian methods exist for the two individual parts of semi-structured deep regression models. For statistical models, variants of Markov Chain Monte Carlo (MCMC) are usually successfully used. However, MCMC is not efficient for a deep neural network because of the high dimensional parameter space, and therefore, an approximate approach like variational inference needs to be used. We propose an efficient Bayesian approach for the whole semi-structured model.

E0694: **Ensembling deep transformation models**

Presenter: **Lucas Kook**, University of Zurich, Zurich University of Applied Sciences, Switzerland

Co-authors: Andrea Goetschi, Philipp FM Baumann, Torsten Hothorn, Beate Sick

Aggregating predictions from several models is a well-known and popular approach to forecasting in many scientific domains, such as machine learning and meteorology. The models may be as simple as decision trees or as complex as deep neural networks. Combining probablistic predictions from deep neural networks is referred to as deep ensembling. It has been shown to lead to better and more robust predictions and uncertainty quantification than the individual members. However, even if individual ensemble members are partially interpretable, the ensemble itself is no longer interpretable in general. We present transformation ensembles which guarantee improved prediction performance and preserve their members' interpretable model structure. The key idea of transformation ensembles is to specify a latent random variable with a simple distribution, and to estimate the model and aggregate its predictions on this latent scale. We demonstrate how to build and fit deep and partially interpretable transformation ensembles and use them to quantify both aleatoric and epistemic uncertainty.

EC788 Room S-2.23 MACHINE LEARNING Chair: Andriette Bekker

E1511: **Accelerated componentwise gradient boosting using efficient data representation and momentum-based optimization**

Presenter: **Daniel Schalk**, LMU Munich, Germany

Co-authors: Bernd Bischl, David Ruegamer

Model-based or componentwise boosting (CWB) is an interpretable gradient-boosting variant that builds on additive models as base learners. Using statistical models as base learners induces an additive model estimation, inherent feature selection, applicability in high-dimensional settings, and favorable scaling w.r.t. the number of features. One downside of CWB is its computational complexity in terms of memory and runtime. We present two novel approaches that (1) reduce the memory load of CWB by applying a discretization technique to numerical features and (2) incorporate Nesterov momentum to speed up the fitting process of the model. Our adaptation of CWB not only drastically reduces memory consumption but also allows the use of specialized matrix operations that further speed up its runtime. Our incorporation of Nesterov momentum preserves well-known advantages of CWB while notably speeding up the algorithm's convergence. We demonstrate its effectiveness in simulation studies and a large real-world benchmark experiment.

E1557: **Temporal event boosting: Gradient boosting applied to conditional intensity models**

Presenter: **Fredrik Lundvall Wollbraaten**, University of Oslo, Norway

Temporal point process data arise in many real-world settings due to increased focus on capturing and storing data across many fields. We consider Marked Temporal Point Processes (MTPP), where each arrival time $t_i \in \mathbb{R}^+$ has a discrete mark $m_i \in \{1,...,M\}$, which we refer to as an event type. Typical examples are system log messages, neural spiking activity, online customer behavior data, events in football matches, or any other setting where discrete events occur in continuous time. Considering the MTPP as a multivariate point process, we propose Temporal Event Boosting (TEB) for estimating mark-specific conditional intensities depending on the history. Despite the success of gradient boosting, its extension to MTPP has not been considered. TEB is a gradient boosting approach for MTPP based on discretizing time, encoding the history of the process using counts of events in different intervals, whereafter gradient boosting is applied. The method is simple to implement using existing software. Using both simulated and real data (trading and football data), we show that TEB performs very well compared to parametric and recurrent neural network-based alternatives.

E1880: **Updating weights using shrinkage methods in artificial neural networks**

Presenter: **Theodor Loots**, University of Pretoria, South Africa

Co-authors: Mohammad Arashi, Andriette Bekker

It is well-known that updating the weights in the gradient descent stage of backpropagation is usually conducted by an unbiased method. On the other hand, in shrinkage estimation, unbiasedness is sacrificed for improved mean square error (MSE) of the estimation process. Therefore, a biased estimation can improve MSE of prediction in the artificial neural network if the weights are updated with a biased shrinkage method. This strategy of updating the weights of the input data is implemented, and findings are illustrated with simulation studies and analysis of the MNIST data.

E1910: **Understanding deep neural network via statistical regression modelling approaches**

Presenter: **Il Do Ha**, Pukyong National University, Korea, South

Co-authors: Kevin Burke, Youngjo Lee

Deep learning (DL) has recently provided breakthrough results for prediction problems, including classification for a wide variety of applications. In particular, the core architectures that currently dominate the DL are deep feed-forward neural networks (DNN), CNN, RNN, LSTM, AE, GAN and Transformer, etc. The DNN models are represented as structured neural networks consisting of three layers (input, hidden and output layers) for constructing (or modelling) the functional relationship (mainly nonlinear) between input and output variables. The main goal is to find a nonlinear predictor of the output Y given the input X. The output models of DNN can be expressed as structured mean models, leading that the estimation of such a mean provides the prediction of Y. It is thus interesting to study the DNN from a statistical perspective. The DNN models can be viewed as a highly nonlinear and semi-parametric generalization of statistical regression models such as the generalized linear model (GLM). The fitting (i.e. learning or training) of DNN models based on train data is usually implemented using likelihood-based methods, which are very useful for the construction of loss function or regularization. We present how to understand the DNN models via the GLM framework, and then extend this perspective to survival models allowing for censoring and also to random-effect models, with simulations and practical examples.

EC771 Room S-1.04 COMPUTATIONAL STATISTICS I Chair: Jonathan Crook

E1551: **Approximate maximum likelihood estimation of the lognormal-GPD model with dynamic weights**

Presenter: **Marco Bee**, University of Trento, Italy

Mixture distributions with dynamic weights are an efficient way of modeling loss data characterized by heavy tails. However, maximum likelihood estimation of this family of models is a difficult problem, mostly because of the need to evaluate an intractable normalizing constant numerically. In such a setup, simulation-based estimation methods are likely to work well. Accordingly, we employ Approximate maximum likelihood estimation (AMLE). This is a general approach that can be applied to a mixture of any component; we focus on the dynamic lognormal-GPD distribution,

and use the empirical characteristic function as a summary statistics. In particular, we develop a hybrid procedure where the standard maximum likelihood is first employed to determine the bounds of the uniform priors of the parameters required as input for the AMLE method. Simulation experiments and a real-data application suggest that this approach yields a major improvement with respect to standard maximum likelihood estimation.

E1569: **Numerical schemes for effective calibration of elliptic and hypo-elliptic diffusions**
Presenter: **Yuga Iguchi**, University College London, United Kingdom
Co-authors: Alexandros Beskos, Matthew Graham

Parametric inference for multi-variate diffusion processes requires using a numerical discretisation scheme as a proxy for the underlying intractable model. The choice of numerical schemes is critical in both likelihood-based inference and computational MCMC methods. Two weak second-order schemes are proposed as effective sampling tools for elliptic and hypo-elliptic diffusions in conjunction with a new closed-form approximation formula for the transition density of the underlying model. Due to consideration of higher-order weak approximation, the proposed schemes are, in general, conditionally non-Gaussian, as opposed to classical Gaussian-type schemes such as the Euler-Maruyama scheme that achieves a weak first-order convergence. The closed-form density approximation is derived by making use of Malliavin calculus in elliptic and hypo-elliptic settings and enables us to construct a (log) likelihood linked to the proposed schemes. Under both the high and low-frequency observations regime, analytical results associated with the weak second-order schemes showcase the effectiveness of the use of proposed schemes in the statistical calibration of diffusion processes compared to earlier works based upon Gaussian-type discretisation schemes.

E1660: **Correlation-adjusted simultaneous testing among small-sized groups in high-dimensional DNA methylation data**
Presenter: **Patrick Wincy Reyes**, University of the Philippines, Philippines
Co-authors: Iris Ivy Gauran, Erniel Barrios, Hernando Ombao

Epigenetics plays a crucial role in understanding the underlying molecular processes of Type 2 Diabetes (T2D) and determining therapeutic targets. In a natural experiment of life conducted on T2D to investigate the complex interplay of genetic variation compared to environmental exposures, we aim to identify differentially methylated probes (DMPs) between controls and cases. Statistically, this is a high-dimensional testing problem with sparse signals and correlated variables across an inherent grouping structure. We propose a class of multiple testing procedures that utilizes the correlation within the genes to control the False Discovery Rate (FDR). Simulation studies show that the proposed methods have superior empirical power while controlling the FDR compared to the benchmark procedures such as Group Benjamini-Hochberg and Group Benjamini-Yekutieli methods. These existing methods fail to control the FDR when the data is grouped with correlated probes. We applied the methods to the data containing a sample of 346 Filipinos enrolled in either Manila or California. Using p-values from the analysis with covariates, we identified a much lower number of significant DMPs, which may facilitate more cost-efficient experimental studies for scientists in identifying novel therapeutic tools for the treatment of Diabetes Mellitus.

E1739: **RandomVariables.jl: A Julia package for random variables, transformations and probabilities**
Presenter: **Manuel Stapper**, WWU Muenster, Germany

When computing probabilities that involve random variables in the Julia language, the Distributions.jl package provides methods for more than 50 univariate distributions. The new package RandomVariables.jl is based on the Distributions.jl package and offers to calculate probabilities of events in such a way that code notation matches formula notation. Single events can be combined by logical connectives, such as conjunction and disjunction as well as by defining conditional events. Random variables can furthermore be transformed, so that the user can simply run $P(exp(X) > 3)$, for instance. Implemented transformations include shifting and scaling, inversion, logarithmic and exponential transformations, absolute values, and powers of random variables.

EC829 Room BH (S) 2.03 MCMC ALGORITHMS Chair: Radu Craiu

E0944: **Bayesian parameter inference estimation for partially observed fractional Brownian motion**
Presenter: **Mohamed Maama**, KAUST University, Saudi Arabia

State space models are widely used in several branches of science, including statistics, applied mathematics, biology, and economics. We consider static bayesian parameter estimation for partially observed diffusions with fractional Brownian motion (fBm). We elaborate on adaptive Markov chain Monte Carlo algorithms that permit us to infer static parameters of stochastic processes based on the Euler-Maruyama approximation. We simulate our algorithms on two models, the first is an Ornstein Uhlenbeck process driven by fBm, and the second is the CoxIngersollRoss (CIR) model with real data. By using numerical results, we compare the efficiency of our algorithms by using the mean square error versus cost.

E0376: **A Markov chain Monte Carlo algorithm for change-point detection in nanopore sequencing data**
Presenter: **Georgy Sofronov**, Macquarie University, Australia
Co-authors: Sophia Shen

Understanding the genetic makeup of organisms is a very important goal in bioinformatics. DNA sequencing, the process of determining the order of the nucleotide bases in DNA, is now able to be performed quickly and cheaply with commercially available devices no bigger than a USB stick. These third-generation nanopore sequencers are capable of capturing long, repetitive DNA structures; however, the reported reading accuracy needs improving. One main source of error occurs when the raw nanopore signals are being translated into genetic alphabets (A, C, G and T). This process is called base-calling. We present a novel base-calling algorithm using Bayesian methodologies and Markov chain Monte Carlo (MCMC) sampling techniques that allow transitions between different models. Since each base transition could be thought of as a change-point in the raw signals, change-point detection or segmentation methods are adopted. We use real and artificially simulated data to illustrate the usefulness of the proposed approach.

E1513: **MCMC approach on Bayesian image analysis in Fourier space**
Presenter: **Konstantinos Bakas**, King Abdullah University of Science and Technology, Saudi Arabia
Co-authors: John Kornak, Hernando Ombao

Bayesian methods are commonly applied to solve image analysis problems such as noise reduction, feature enhancement and object detection. A primary limitation of these approaches is the computational complexity due to the interdependence of neighboring pixels, which limits the ability to perform full posterior sampling through Markov Chain Monte Carlo (MCMC). To alleviate this problem, we will develop a new posterior sampling method that is based on modeling the prior and likelihood in the space of the Fourier transform of the image. One advantage of Fourier-based methods is that spatially correlated processes in image space can be modeled via independent processes in Fourier space. A recent approach known as Bayesian Image analysis in Fourier Space (or BIFS), has introduced parameter functions to describe prior expectations about image properties in Fourier Space. To date, BIFS has relied on Maximum a Posteriori (MAP) estimation for generating posterior estimates, i.e. providing just a single point estimate. A posterior sampling approach for BIFS is presented that can explore the full posterior distribution while continuing to take advantage of the independence modeling over Fourier space. As a result, computational efficiency is improved and mixing concerns that commonly have to be dealt with in high dimensional Markov Chain Monte Carlo sampling problems are avoided.

E1866: **Bayesian semiparametric copula estimation using an infinite mixture of Gaussian copulas.**
Presenter: **Jichan Park**, Korea University, Korea, South
Co-authors: Taeryon Choi

A Bayesian semiparametric approach is presented for estimating an unknown bivariate copula using an infinite mixture of Gaussian copulas. To accomplish this, we use a copula model that can separately build the dependence structure and marginal distributions of a bivariate distribution. The dependence structure is induced by different parametric bivariate copulas such as Gaussian, Student t, Clayton and Gumbel copula and specifically, they are sampled through the reversible jump Markov chain Monte Carlo (RJMCMC) algorithm. Furthermore, we consider Dirichlet process priors to estimate a more flexible dependence structure in the nonparametric way. Marginal distributions are estimated via the Bayesian spectral analysis density estimation (BSAD) method with and without shape restrictions. Finally, we study the performance of the RJMCMC algorithm and compare estimates in each model with simulations of various settings and several real data sets.

EP001 Room Posters Virtual Room 1 POSTER SESSION I (ONLY VIRTUAL) Chair: Cristian Gatu

E0480: **Construction of random forest based on parametric bootstrapping**
Presenter: **Asanao Shimokawa**, Tokyo University of Science, Japan
Co-authors: Etsuo Miyaoka

Ensemble learning improves prediction accuracy by combining multiple learners. Random forest is one of the typical methods of ensemble learning, and it combines the tree-structured models as the individual learners. In many studies, individual learners are trained based on samples obtained by sampling with replacement of supervised data. Although this is a non-parametric bootstrap-based learning method, if there is additional information about the population distribution, it seems more plausible to apply the parametric bootstrap method. Therefore, we focus on random forest based on parametric bootstrap method and compare it with the conventional ensemble learning method through theoretical and simulation studies. In addition to this, we show the results of the forest obtained from actual data.

E0667: **Multiple clustering based Wishart mixture models and its application to seismic data analysis**
Presenter: **Tomoki Tokuda**, The University of Tokyo, Japan
Co-authors: Hiromichi Nagao

For high-dimensional data, it is not straightforward to cluster objects because all features are not always relevant to a particular cluster solution. Some features may be relevant for one cluster solution, but irrelevant for another. In general, in high-dimensional cases, one may assume multiple cluster solutions depending on a specific subset of features. In this situation, a conventional clustering method would not be able to reveal the underlying cluster structure, which is characterized by multiple cluster solutions. So far, effective methods to find such a cluster structure have been less developed. A novel clustering method is discussed, which is useful for revealing the underlying multiple cluster structure. The method is based on Wishart mixture models, which apply to correlation matrices of connectivity data without vectorization. The uniqueness of this method is that multiple cluster solutions are based on particular networks of nodes, optimized in a data-driven manner. Hence, it can identify the underlying pairs of associations between a cluster solution and a node sub-network. The method is applied to seismic data, and it is shown that the method can potentially capture weak tremor, which is otherwise difficult to identify.

E1518: **Matrix decomposition factor analysis with variable selection directly constrained by the number of zero columns**
Presenter: **Shimada Naoya**, Osaka University, Japan

It is very important to select the variables to be used in factor analysis when there are many observed variables. We propose a new method for variable selection in factor analysis. The proposed method is based on matrix decomposition factor analysis, which treats all factor analysis model parameters (common and unique factors, loadings, and unique variances) as a fixed unknown matrix. This method treats parameters as grouped parameters by column (variable) and then uses the L_0 norm to constrain the number of variables used in the analysis. Three features distinguish this approach from previous studies. First, the tuning parameters that determine the number of variables to be selected are finite positive integers, and the optimal parameter values can be easily estimated by considering all possible values. Second, the number of variables to be used in the analysis can be ascertained prior to conducting the analysis. Third, the proposed method can derive a residual matrix that indicates the degree not accounted for in the model equation. In addition, we propose a method of constraining the L_0 norm for both the parameters of the individual variables and the parameters grouped by columns to facilitate the interpretation of the factor loading matrix. This method can simultaneously achieve variable selection and estimation of interpretable sparse factor loadings. Simulation and real data experiments were conducted to investigate the effectiveness of the proposed method.

E1713: **Optimization methodologies for big data analysis**
Presenter: **Carisa Kwok Wai Yu**, The Hang Seng University of Hong Kong, Hong Kong
Co-authors: Siu Kai Choy, Wai Leong Ng, Chi Chung Siu

The explosive growth of big data brings opportunities and has benefited the development and the revolution of various disciplines. However, the curse of missing data, gross error and dimensionality is an arduous challenge for big data analysis. The low-rank matrix optimization is an important technique to deal with the curse of missing data and gross error in big data analysis in various disciplines. Sparse optimization is a popular and practical technique to deal with dimensionality for big data problems and has been successfully implemented in various fields. We will discuss new optimization methodologies with continuation techniques to solve the problems in big data analysis.

E0676: **Bartlett correction of T^2 type test statistic with two-step monotone missing data in two-sample problem**
Presenter: **Tamae Kawasaki**, Aoyama Gakuin University, Japan
Co-authors: Takashi Seo

The problem of testing the equality of mean vectors in a two-sample problem with two-step monotone missing data is discussed. Under the assumption that the population covariance matrices are equal, we derive the stochastic expansion of Hotelling's T^2 type statistic for the case where sample sizes are large. The asymptotic first two moments of T^2 type statistic are obtained by its stochastic expansion. We also propose the Bartlett corrected statistic for two-step monotone missing data. Finally, the accuracy and asymptotic behavior of the approximation are investigated by Monte Carlo simulation.

E1643: **A performance evaluation of randomization methods in small-size clinical trials with binary or survival outcomes**
Presenter: **Kanae Takahashi**, Hyogo Medical University, Japan
Co-authors: Kouji Yamamoto

Clinical trials are the most definitive method for determining the treatment effect. The important component of clinical trials is randomization which is a technique used for assigning patients to either the experimental treatment(s) group or the control group. The selection of a randomization method is important at the time of protocol planning. There are several methods of randomization. Complete randomization, permuted block design, stratified blocked randomization, and minimization are relatively commonly used. The performance of randomization methods may be related to trial sample size, the setting of prognostic factors, and statistical analysis methods. However, for small-size randomized controlled trials with binary or survival outcomes, the impact of using a prognostic factor that has an interaction or a baseline characteristic that has no effect on the outcome for randomization and statistical analysis was not considered well in previous research. Thus, the objective is to compare the performance of the existing randomization methods when using a prognostic factor that has an interaction or a baseline characteristic that has no effect on the outcome by simulations and to suggest the recommended randomization method for small-size randomized controlled trials with binary or survival outcomes.

E2017: **Likelihood-based inference in control risk regression with study-specific covariates**
Presenter: **Phuc Tran**, University of Padova, Italy
Co-authors: Annamaria Guolo, Annamaria Guolo

Control risk regression is a meta-analysis approach to investigate the effectiveness of a treatment in clinical studies. It is characterized by the inclusion of a risk measure for the subjects in the control condition as a way to explain between-study heterogeneity. The measures of risk for the treatment group and the control group are summary information of each included study in the meta-analysis, and thus they are affected by an error. Properly accounting for measurement error is a necessary step for inference to be reliable. We investigate likelihood-based inference in control risk regression models in case additional error-affected covariates, aggregated at the study level, are taken into account. In such a situation, when the within-study covariances are unknown, we show how to compute them using Taylor expansions and empirical-type estimation. As an alternative, we develop a pseudo-likelihood approach, under a working conditional independence assumption, which sets unknown covariances to zero. The performance of the proposed solutions is examined in an extensive simulation study. Moreover, the methods are applied to a meta-analysis of the efficacy of convalescent plasma in the treatment of COVID-2019.

E0399: **Optimal linear filter from fading measurements with time-correlated additive noises and transmission random losses**
Presenter: **Raquel Caballero-Aguila**, Universidad de Jaen, Spain
Co-authors: Josefa Linares-Perez

The signal estimation problem in multisensor systems has developed into an important research area, due to its significant relevance in numerous applied and theoretical fields. It is common knowledge that networked systems frequently have random flaws, which, if not appropriately addressed, are likely to harm the performance of the estimators. The assessment of mathematical models and the development of estimation algorithms accounting for these random phenomena have thus received a lot of research attention. The fading or degradation of measurements (caused, e.g., by physical equipment limitations or inaccurate measurement instruments) is one of the most common uncertainties in sensor networks. Under the assumption that the measurements are affected by the fading phenomena, as well as perturbed by time-correlated additive noises and exposed to random packet dropouts during transmission, a recursive least-squares linear filtering algorithm is designed using a covariance-based methodology and a compensation strategy based on measurement prediction. For this kind of system with packet losses, the measurement differencing method, typically used to deal with the measurement noise time-correlation, is not successful, since some sensor measurements are randomly lost and, consequently, cannot be processed. Therefore, an alternative approach, based on the innovation technique and the direct estimation of the measurement noises, is used.

CO096 Room S-1.01 RISK ANALYSIS AND ASSESSMENT IN ECONOMICS AND FINANCE Chair: Alessandra Amendola

C0638: **Combining historical data sources in operational risk capital estimation**
Presenter: **Helgard Raubenheimer**, North-West University, South Africa
Co-authors: Riaan de Jongh, Mentje Gericke

The management of financial losses is crucial as banks must set aside regulatory capital to absorb unexpected losses. Banks also calculate economic capital to ensure solvency according to their risk profile. The main financial risks faced by banks are market-, credit-, and operational risks. Operational risk includes fraud, improper business practices, and so on. The annual aggregate distribution of expected losses is determined to calculate the capital needed to withstand this risk. The extreme quantiles of this distribution are of specific interest. For instance, a bank should hold capital to survive a one-in-a-thousand-year aggregate operational loss (the 99.9% VaR of the distribution). However, companies often have limited internal data available to model the distribution accurately and therefore use external sources and scenario assessments to supplement their data. We show how data sources can be integrated into the capital modelling process. We also suggest measures to challenge experts to adjust scenario assessments based on historical data.

C1034: **The impact of ESG scores on tails risk measures**
Presenter: **Alessandra Amendola**, University of Salerno, Italy
Co-authors: Luigi Aldieri, Vincenzo Candila

Over the past two decades, there has been increased attention and larger consciousness about the Environmental, Social and Governance (ESG) responsibilities of the firms. The asset allocation process has changed accordingly to consider ESG responsibilities, and it has been largely recognized that private and institutional investors are sensible to ESG factors when deciding which firms to invest in. Other key factors to which investors generally pay attention are the loss which is likely to occur at a given probability and over a specific period (Value-at-Risk - VaR); the expected loss (Expected Shortfall - ES) sustained in that portion of unfortunate events. The present contribution aim at investigates the VaR and ES of a set of listed firms with different ESG scores through different backtesting procedures.

C1327: **Choosing between persistent and stationary volatility**
Presenter: **Ilias Chronopoulos**, University of Essex, United Kingdom
Co-authors: Liudas Giraitis, George Kapetanios

A multiplicative volatility model is suggested where volatility is decomposed into a stationary and a non-stationary persistent part. We provide a testing procedure to determine which type of volatility is prevalent in the data. The persistent part of volatility is associated with a nonstationary persistent process satisfying some smoothness and moment conditions. The stationary part is related to stationary conditional heteroskedasticity. We outline theory and conditions that allow the extraction of the persistent part from the data and enable standard conditional heteroskedasticity tests to detect stationary volatility after persistent volatility is taken into account. Monte Carlo results support the testing strategy in small samples. The empirical application of the theory supports the persistent volatility paradigm, suggesting that stationary conditional heteroskedasticity is considerably less pronounced than previously thought.

C1882: **Volatility jumps and the classification of monetary policy announcements**
Presenter: **Demetrio Lacava**, Luiss University, Italy
Co-authors: Giampiero Gallo, Edoardo Otranto

In the last two decades, due to many endogenous and/or exogenous events (e.g. subprime mortgage crisis, high inflation, and the Covid-19 pandemic), even more frequent actions have been taken by many central banks, with direct implications on market volatility. Taking the example of the Federal Reserve (FED), we propose a new model-based classification of monetary policy announcements according to whether they cause a jump rather than a reduction of volatility. The proposed model (the GAS-AMEM with jumps) - which combines the distinctive features of the Generalized Autoregressive Score (GAS) model along with those of the Multiplicative Error Model (MEM) - provides an accurate classification method, while preserving the flexibility and the ability of the classical MEM of reproducing the well-known stylized facts characterizing volatility. Focusing on a short window around each FED's communication, we isolate the impact of monetary announcements by excluding any contamination carried by relevant events that may occur within a business day. By relying on both the S&P500 stock index and specific firms, we classify FED's announcements according to their effect on the stock market as a whole, on the one hand, and on specific sectors of the market, on the other hand.

CO396 Room Virtual R03 MODELLING FINANCIAL MARKETS Chair: Menelaos Karanasos

C1312: **Short- and long-run cross-country sustainability interdependences**
Presenter: **Starvoula Yfanti**, Queen Mary University of London, United Kingdom

Co-authors: Menelaos Karanasos, Jiaying Wu

The cross-country interconnectedness among sustainability equity indices is investigated. Using a bivariate corrected-DCC-MIDAS setting, we study the short- and long-run time-varying dependence dynamics among national sustainability benchmarks. Our correlation analysis identifies short- and long-run hedging characteristics and interdependences and concludes on both procyclical and countercyclical cross-country interlinkages.

C1488: **Group patterns in income inequality and economic growth**
Presenter: **Binzhi Chen**, University of Birmingham, United Kingdom
Co-authors: Yiannis Karavias, Marco Barassi

The relationship between income inequality and economic growth has been debated for a long time. The purpose is to take into account the cross-country latent group patterns by using the latest econometrics techniques, Grouped Fixed Effects, to examine the growth-inequality nexus. The estimation results from a broadly balanced panel with 70 economies from 2000-2018 indicate that non-linear effects, heterogeneity and latent group patterns exist simultaneously in the short run. Especially, we find that worldwide economies form three group structures. However, three economies, namely Ireland, Sierra Leone and Zimbabwe, do not have any group patterns with any other economies. The results are valid for both the Gini index, Gini disposable index and Gini market index, respectively. The non-linear relationship is also robust in several robustness checks. The short-run positive impact on economic growth can be treated as a complement to previous studies and Kuznets Curve.

C1348: **Can return forecasts enhance international asset allocation: Evidence from the sum of the parts approach**
Presenter: **Andrew Vivian**, Loughborough University, United Kingdom
Co-authors: Mark Wohar

The aim is to examine whether real-time return forecasts are valuable to an investor looking to allocate their portfolio across a wide selection of countries. We expand the Sum-of-Parts (SoP) method for forecasting stock returns to an international setup by adding FX returns as an additional component. We use two different methods to calculate the forecasts. The first method (Empirical Mode Decomposition) uses wavelets to frequency decompose each part into locally independent sub-signals, while the second method combines historical averages and predictive regressions. We then compare the performance of various types of portfolia under the SoP and historical average forecasts, with rebalancing taking place every period. We find that SoP forecasts deliver economic gains over the historical average, especially when the EMD method is implemented. We further demonstrate that economic gains can be generated for investors based in various different countries.

C1425: **Time-varying connectedness between the US futures markets and the macroeconomy**
Presenter: **Aris Kartsaklas**, Brunel University London, United Kingdom
Co-authors: Awon Almajali

The aim is to examine the connectedness, in return and volatility systems, among futures markets and macroeconomics variables from 1997 to 2015, covering two significant crises: the global financial crisis and the tech bubble. We utilize a time-varying parameter VAR model (TVP-VAR), which is based on the recently developed connectedness approach. The results show a changing level of connectedness with an average of 70.5% for return and 75.6% for volatility. We find that macroeconomic variables are the main contributors to the overall forecast error variance, a result that holds at both return and volatility levels. For example, non-borrowed reserves and total reserves show the highest contributions among all other macro and finance variables. Overall, our findings are robust to Bayesian prior choice and reflect the rapid influences of both crises, which is essential to formulate policies that seek to achieve financial stability.

CO520 Room Virtual R04 DATA ANALYSIS TOOLS FOR BAYESIAN INFERENCE Chair: Catherine Forbes

C1324: **Assessing the sensitivity of a Youden posterior to extremes in the classification variable**
Presenter: **Catherine Forbes**, Monash University, Australia

The two-class Youden index corresponds to an optimal threshold value for classifying subjects into one of two distinct groups. The value of the index maximises the sum of the two correct classification probabilities for each group. However, the population distributions of each group must be estimated since they are usually unknown. The empirical distribution functions for each have been used to estimate the threshold. A Bayesian bootstrap has been used to obtain a posterior distribution for the receiver operating characteristic (ROC) curve, which in turn may be used to obtain a posterior distribution for the threshold value. This Youden posterior effectively accommodates distributional uncertainty; however, it may be sensitive to extremes and asymmetry in the classification variable. To explore the impact of any potential sensitivity, we systematically re-weight individual sample observations to determine a range of posterior point and interval estimates of the Youden threshold. We illustrate our approach in the context of a derived (subjective) housing affordability indicator for an urbanised state in Malaysia.

C1248: **Bayesian case influence analysis for spatial autoregressive model**
Presenter: **Cheok Hang Lei**, Monash University, Australia

A general methodological framework is developed for Bayesian case influence analysis for a spatial autoregressive model. An algebraic representation for the importance sampling weights associated with case-deletion is derived for use with sampled draws from a full data posterior produced via Markov chain Monte Carlo. Once these weights are obtained, a principal components analysis is used on the covariance matrix of log-case deletion weights to produce low dimensional case influence summary plots. The methodology is then applied to artificial data sets, each with a single influential observation. Influential observations are detected in the plots.

C1230: **Regularized exponentially tilted empirical likelihood for Bayesian inference**
Presenter: **Eunseop Kim**, The Ohio State University, United States
Co-authors: Steven MacEachern, Mario Peruggia

Empirical likelihood extends the use of likelihood to models defined through moment conditions with minimal distributional assumptions. A likelihood closely related to empirical likelihood, called the exponentially tilted empirical likelihood, arises as a nonparametric Bayesian procedure with the prior that favors distributions close to the empirical distribution. Despite its desirable asymptotic properties, the so-called empty set problem or the convex hull constraint limits its use for Bayesian inference. We propose a hybrid method for the exponentially tilted empirical likelihood that is free from the empty set problem. The method introduces an auxiliary exponential family distribution and applies exponential tilting to the mixture of the empirical distribution and the auxiliary distribution. The auxiliary distribution prevents the empty set problem and regularizes the likelihood by forming an exponential family. We demonstrate that the method stabilizes the posterior distribution in the full parameter space, enabling more efficient posterior sampling with relatively small sample sizes.

CO084 Room BH (S) 1.01 Lecture Theatre 1 MODELLING ECONOMIC AND FINANCIAL TIME SERIES Chair: Gianluca Cubadda

C0780: **Detecting common bubbles in multivariate mixed causal-noncausal models**
Presenter: **Alain Hecq**, Maastricht University, Netherlands

Methods are proposed to investigate whether the bubble patterns observed in individual series are common to various series. We detect the non-linear dynamics using the recent mixed causal and noncausal models. Both a likelihood ratio test and information criteria are investigated, the former having better performances in our Monte Carlo simulations. Implementing our approach on three commodity prices, we do not find evidence of commonalities, although some series look very similar.

C0781: **The vector error correction index model: Representation and statistical inference**
Presenter: **Gianluca Cubadda**, University of Rome TV, Italy
Co-authors: Marco Mazzali

The multivariate index autoregressive model is extended to the case of cointegrated time series of order (1,1). In this new modelling, which we call the Vector Error-Correction Index Model (VECIM), the first differences of cointegrated time series are driven by some linear combinations of the variables that are labelled as the indexes. When the number of indexes is small compared to the sample size, the VECIM achieves a significant dimension reduction w.r.t. the classical Vector Error Correction Model (VECM), thus allowing to analyze cointegration even in medium vector autoregressive models, a setting where maximum likelihood inference for the VECM does not work well. We show that the indexes follow a VECM of smaller dimension than the number of series, that the VECIM allows decomposing the reduced form shocks into sets of common and uncommon shocks, and that the former can be further decomposed into permanent and transitory shocks. Moreover, we offer a switching algorithm to estimate the parameters of the VECIM optimally. Finally, we document the practical value of the proposed approach through both simulations and an empirical application. In particular, we search for the shocks that drive the aggregate fluctuations at different frequency bands in the US. We find that a common transitory shock generates most of the variability at the business cycle frequencies, whereas a common permanent shock drives the long run.

C0912: **Testing for the cointegration rank between periodically integrated processes**
Presenter: **Tomas del Barrio Castro**, University of the Balearic Islands, Spain

Cointegration between periodically integrated (PI) processes has been widely analyzed. However, there is currently no method that allows us to determine the cointegration rank between PI processes. A method is proposed for determining the cointegration rank between a set of PI processes based on the idea of pseudo-demodulation, as proposed recently in the context of seasonal cointegration previously. Once a pseudo-demodulated time series is obtained, a previous procedure can be applied to determine the cointegration rank. A Monte Carlo experiment shows that the proposed approach works satisfactorily for small samples.

C1330: **Hierarchical regularizers for reverse unrestricted MIDAS**
Presenter: **Marie Ternes**, Maastricht University, Netherlands
Co-authors: Alain Hecq, Ines Wilms

Reverse unrestricted MIDAS (RU-MIDAS) regressions are used to model high-frequency variables by means of low-frequency variables. However, in practice, the dimensionality of RU-MIDAS grows quickly due to the frequency mismatch between the high- and low-frequency components and the number of explanatory variables included. We propose tackling dimensionality through sparsity-inducing convex regularizers built upon the group lasso with nested groups. The regularizer encourages hierarchical sparsity patterns by prioritizing the inclusion of coefficients according to the recency of the information they contain.

CO080 Room BH (SE) 1.01 ADVANCES IN TIME SERIES METHODS Chair: Massimiliano Caporin

C0326: **Cost-at-Risk estimation on the Italian ancillary services market**
Presenter: **Luigi Grossi**, University of Padova, Italy
Co-authors: Francesco Lisi, Federico Quaglia

The aim is to explore how to evaluate the cost risk related to a market-based ancillary services procurement by an electricity Transmission System Operator (TSO). We consider the case of Terna, the Italian TSO, which operates the Italian electricity ancillary service market. We propose a two-step procedure moving from the time series of incurred costs for the years 2017-2020 and some possible costs drivers such as actual consumptions, levels of reserve requirements, network nodal and zonal constraints, as well as prices of the primary commodities affecting the Italian electricity market (natural gas and carbon dioxide). Calendar variables are also included to account for periodic and other deterministic effects. In the first step of the procedure, we model the dynamics of the cost conditional mean, which can be viewed as the predictable part of costs. A non-parametric model describing the relations among variables impacting costs, possibly non-linearly, is identified and estimated. This model relies on the basis and periodic splines. Models residuals represent the unpredictable costs component and are used to quantify the Cost-at-Risk (CaR), the equivalent of Value-at-Risk in the case of costs. For this purpose, we compute and test different approaches for 1-day and 30-days CaR. Models for conditional quantiles dynamics were identified and estimated on the residuals of the conditional mean model and were based on lagged values.

C0449: **Point and probabilistic forecast reconciliation for general linearly constrained multiple time series**
Presenter: **Tommaso Di Fonzo**, University of Padova, Italy
Co-authors: Daniele Girolimetto

Hierarchical forecast reconciliation is the post-forecasting process aimed at revising a set of incoherent base forecasts into coherent forecasts in line with cross-sectional/temporal/cross-temporal data structure. Most of the point and probabilistic hierarchical forecast reconciliation results move from the classic reconciliation formula valid for the structural representation of a hierarchical time series. When a general linearly constrained multiple time series is considered, the projection approach reconciliation formula gives a general solution. While it is well known that the classic structural reconciliation formula is equivalent to its projection approach counterpart, it is not obvious up to now if and how a structural-like reconciliation formula may be derived for a general, not genuinely hierarchical time series. Such an expression would permit extending definitions, theorems, and results found recently for probabilistic forecast reconciliation in a rather straightforward manner. We show that even for general linearly constrained multiple time series, it is possible to express the reconciliation formula according to a structural approach that keeps distinct free and basic, instead of bottom and upper (aggregated), variables. We extend the definition of probabilistic forecast reconciliation to a general linearly constrained multiple time series, and consider an empirical example.

C0308: **Conditional autoregressive G model for common factor detection in the stock market**
Presenter: **Marco Girardi**, University of Padova, Italy
Co-authors: Massimiliano Caporin

A new distribution is presented to model financial assets' realized covariances, obtained as the product of a scalar component distributed as a unit-mean inverse gamma and a matrix component following a Wishart distribution. The mean of the resulting distribution is endowed with an autoregressive moving average structure. The model captures a common factor in the assets' behaviour, which constitutes the inherent risk in the market, as well as the idiosyncratic risk component. The one-step-ahead forecasts of the covariance matrix are employed in a portfolio allocation framework with the aim of tracking the reference index performance by limiting the impact of specific risks. A possible empirical application of the model in a hybrid portfolio management strategy is also discussed.

C0323: **Penalized CAW, forecast error variance decompositions and systemic risk measurement**
Presenter: **Massimiliano Caporin**, University of Padova, Italy
Co-authors: Giuseppe Storti

Parameter estimation of the Conditional Autoregressive Wishart model under penalization is discussed. We introduce two novel Forecast Error Variance Decompositions where returns shocks impact on the entire set of realized variances and covariances, the first following a more traditional approach and the second based on simulations. From both decompositions, we derive a spillover index to monitor the systemic risk. An empirical analysis on US large-cap equities exemplifies the proposals.

CO738 Room BH (SE) 1.02 COMMODITIES: PRICING AND TRADING Chair: Ana-Maria Fuertes

C0366: **Seasonality in commodity prices: New approaches for pricing plain vanilla options**
Presenter: **Carme Frau**, University of Balearic Islands, Spain
Co-authors: Viviana Fanelli

A new term-structure model is presented for commodity futures prices, which extends a previous model by incorporating seasonal stochastic volatility represented with two different sinusoidal expressions. We obtain a quasi-analytical representation of the characteristic function of the futures log-prices and closed-form expressions for standard European options' prices using the fast Fourier transform algorithm. We price plain vanilla options on the Henry Hub natural gas futures contracts, using our model and extant models. We obtain higher accuracy levels with our model than with the extant models.

C1515: **Reasons behind words: Causes and consequences of OPEC narratives**
Presenter: **Marc Joets**, IESEG School of Management, France
Co-authors: Celso Brunetti, Valerie Mignon

Several studies have documented the effect of OPEC announcements on crude oil prices and volatility with mixed results. They all concentrate on the event study of production or price-related announcements. We propose to reverse the question by asking: what are the factors behind OPEC communications? In other words, what are the exogenous elements that influence the nature of OPEC communications? To answer this question, we propose a completely novel approach using natural language processing. We first model OPEC communications by extracting topics from OPEC press re- leases over the period March 2002-March 2021. This allows us to construct the whole spectrum of OPEC communications over time. Second, we propose a methodology to select major topics that are strictly related to crude oil prices across different maturities. Third, we analyze the causes of OPEC communications among a set of exogenous factors.

C1604: **Cross-market return predictability, commodity and capital market integration**
Presenter: **Emmanuel Eyiah-Donkor**, University College Dublin, Ireland
Co-authors: John Cotter, Valerio Poti

The cross-predictability of asset returns is studied using predictor variables specific to the commodity, stock and bond markets, and market integration thereof. We present evidence supporting the hypothesis that the commodity market is only partially integrated with the stock and bond markets. Extensive tests of predictability show that the information content of commodity-specific predictors has statistically significant predictive power for stock and bond excess returns at both short and long horizons. Portfolio analysis using mean-variance spanning tests confirms the robustness of the predictive regression tests. Specifically, we also find that the returns on a commodity predictability-based trading strategy constructed using stock and bond predictors do not improve the utility of a mean-variance investor with an existing portfolio composed of commodities and a commodity predictability-trading strategy constructed using commodity predictors. Our results have two implications. First, at the margin, different traders seem to price the strategies that exploit the predictability of each asset class. Second, there appear to be different marginal risk arbitrageurs for each asset class in line with the literature on limits to speculation and capital mobility. Traders in the stock and bond markets seem to value information from the commodity market but not vice versa.

C0361: **Commodity hedging: Traditional or Selective**
Presenter: **Ana-Maria Fuertes**, City University London, United Kingdom
Co-authors: Joelle Miffre, Adrian Fernandez-Perez

Commodity selective hedging aims at simultaneously covering the risk of the spot position and earning a speculative premium that reflects the hedgers' view of commodity futures price changes. The purpose is to confront the traditional minimum-variance hedging approach and many selective hedging strategies as deployed in the empirical literature that rely on historical averages as nave forecast, forecasts from AR or VAR models, equal-weighted forecasts combinations from univariate models or the direct integration of commodity pricing signals suggested in the recent style integration literature. Deploying the strategies out-of-sample for 24 commodities, it turns out that the hedgers expected utility gains versus no hedging are largest with traditional minimum-variance hedging, namely, when the speculative component of the selective hedge is ignored and the focus is on risk minimization. The superiority of the minimum variance hedging strategy exacerbates in bad times when commodity market volatility is high or during NBER recessions, and is unchallenged when considering transaction costs, alternative specifications of the selective hedges, rolling or expanding estimation windows or different rebalancing frequencies.

CO546 Room BH (SE) 1.05 ADVANCES IN CLIMATE AND ENERGY ECONOMICS Chair: Felix Kapfhammer

C1597: **The asymmetric effects of weather shocks on euro area prices**
Presenter: **Catalina Martinez Hernandez**, European Central Bank, Germany
Co-authors: Matteo Ciccarelli, Friderike Kuik

The impacts of weather shocks on sectoral prices in the four largest euro area economies are assessed. The effects are estimated using high-frequency weather data and monthly data on prices and output, within a set of Bayesian Vector Autoregressions, which explicitly consider the seasonal dependence of the shock. The results suggest that temperature-related shocks have significant effects on prices. The response of aggregate prices to temperature shocks occurs mainly via food, energy and service prices, and is heterogeneous across countries and over seasons. An increase in monthly mean temperatures increases prices in summer and decreases prices in other seasons of the year, with a stronger response in warmer euro area countries. An increase in temperature variability significantly impacts prices beyond the impacts of changes in means. Our results suggest that future weather shocks - increasingly frequent and intense in a changing climate - affect price and price volatility dynamics.

C2014: **The economic consequences of effective carbon taxes**
Presenter: **Felix Kapfhammer**, BI Norwegian Business School, Norway

The sectoral and macroeconomic consequences of carbon taxes are studied. In contrast to the previous literature, we develop a new monthly measure of effective carbon tax rates, which accounts for the time-varying emission coverage of taxes that are both explicitly and implicitly levied on greenhouse gas-emitting goods. Employing the new measure for four Nordic countries, we find that effective carbon taxes reduce emissions as expected, but they also decrease sectoral and macroeconomic activity - though there is some heterogeneity in the effects within and across the Nordics.

C1530: **Economic narrative processing in the case of climate change**
Presenter: **Noriyasu Kaneda**, Bank of Japan, Japan
Co-authors: Hiroki Sakaji

Important economic narratives are extracted from newspaper articles using a causal chain method and BERT, a deep learning-based language model that performs dependency parsing of economic contexts. Economic narratives are important causes of major economic events, and then they could provide clues for understanding people's beliefs and expectations. The novel framework can construct indices of economic narratives for any financial and economic issues and visualize a graphical linkage of subtopics. As a case study, we apply this method to describe a narrative for climate change. The result suggests that governments had major roles in discussing international frameworks and environmental regulations for carbon neutrality in the 2000s. Also, firms and green investors seemed to react to the progress of the climate policy debate and have tackled long-term risk management and investment in new businesses in recent years. The climate narrative could be useful for the analysis of spillover

of policy announcements, expectation formation and behavioral changes in microeconomic entities. For future work, we need to test empirically whether the novel method could be an effective approach to dissecting people's beliefs in climate policy, market bubbles or inflation and embodying narrative economics.

C1678: **Green factor: Quantifying equity returns' climate risk using green active fund allocations**
Presenter: **Jeanne Gohier**, Universite Paris 1 Pantheon Sorbonne - CNRS UMR8174, France
Co-authors: Thibault Soler

The shift to a green, low-carbon economy has generated new investment strategies and impacted asset prices. We analyse the climate risk of assets by building a brown-minus-green (BMG) factor and evaluating assets' exposure to this factor. The CARIMA project develops a similar approach on the basis of extra-financial performance (ESG) data only. While it captures significant information datapoints on the exposure to climate risk, it is very arduous to compute and to update, as it requires four different databases. Moreover, there is a significant time lag on ESG data, especially for carbon data (two years) which makes these BMG factors outdated. Here, we develop a new method to evaluate market sentiments by using green funds/ETFs database to build a BMG factor. We first use principal-component analysis to isolate the dark-green component and then compute the sensitivity of asset prices to this dark-green proxy portfolio to build the BMG factor. This methodology improves the evaluation of the assets' individual climate risk, by capturing up-to-date market sentiments which encapsulate both fundamental investor and analyst knowledge on the extra-financial performance of assets. We end by analysing the explanatory power of this BMG factor when added to the Carhart asset pricing model.

CO106 Room BH (SE) 1.06 FACTOR AND GARCH MODELS Chair: Niklas Ahlgren

C0749: **Reassessing the evidence on factor and portfolio premia**
Presenter: **Agnieszka Jach**, Hanken School of Economics, Finland
Co-authors: Jan Antell

Using a previous modelling and estimation framework, we perform a test of the mean (T_2 statistic) for a large collection of daily Fama-French factors and portfolio returns, and compare the results with those based on the standard t test. The T_2-based results provide clearly weaker evidence in favour of various premia and, in some cases, suggest their absence. On the US market, the discrepancy between the tests is particularly large for the value and momentum factors. Caution should be exercised when assessing the presence of a given premium with the t test.

C0812: **Robust Estimation and inference for time-varying unconditional volatility**
Presenter: **Genaro Sucarrat**, BI Norwegian Business School, Norway
Co-authors: Rickard Sandberg

The unconditional volatility of financial return is often time-varying. To model this, a common approach is to decompose the volatility σ_t^2 multiplicatively into a non-stochastic process g_t, and a de-volatilitised stochastic process h_t: $\sigma_t^2 = g_t h_t$. We prove the consistency and asymptotic normality of the single-step Quasi Maximum Likelihood Estimator (QMLE) of the parameters of g_t for a large class of specifications of g_t. Next, we derive a simple but robust and consistent estimator of the asymptotic coefficient covariance. The exact specification of h_t need not be estimated or known, and h_t can even be non-stationary in the distribution. This is important in empirical applications, since financial returns are frequently characterised by a non-stationary zero-process. Next, we derive a period-by-period estimator of time-varying periodic unconditional volatility. Due to the assumptions we rely upon, our results extend directly to the Multiplicative Error Model (MEM) interpretation of volatility models. So our results can also be applied to the modelling of the time-varying unconditional mean of non-negative processes (e.g. volume, duration, realised volatility, dividends and unemployment). Three applications illustrate our results.

C0970: **Testing the ATV-GARCH model**
Presenter: **Niklas Ahlgren**, Hanken School of Economics, Finland
Co-authors: Alexander Back, Timo Terasvirta

It is common for long financial time series to exhibit a gradual change in conditional and unconditional volatility. The additive time-varying (ATV-)GARCH model allows for structural change with remarkable flexibility. Instead of making all GARCH parameters time-varying, the intercept is parameterised by a logistic transition function with rescaled time as the transition variable. This specification is a parsimonious parameterisation of the very general nonparametric time-varying GARCH. It provides a simple and flexible way to capture deterministic nonlinear changes in the variance, and is particularly well suited for situations in which volatility of an asset or index is systematically increasing or decreasing (or both) over time. The model is unidentified if the intercept is constant. It is, therefore, imperative to test the constancy of the intercept before attempting to fit the model to data. We derive Lagrange multiplier (LM) tests of GARCH against ATV-GARCH. A testing-based modelling strategy is introduced and illustrated by two empirical examples.

C1337: **A structurally motivated stochastic volatility model for equity returns**
Presenter: **Alexander Back**, Hanken School of Economics, Finland

The conditional volatilities of equity returns often exhibit a long-memory property. In standard volatility models, this may lead to parameter estimates that suggest a highly persistent process, often bordering on non-stationarity. To deal with this, several researchers have proposed multiplicatively decomposing volatility into slow-moving and transient components. A recent vein in the literature has used a structural approach to identify the slow-moving component as potentially stemming from the amount of financial leverage that the firm under consideration has taken on. Consequently, the transient component is interpreted as a volatility model in assets rather than equity returns. This captures the well-known notion that leverage makes equity riskier. We use this backdrop to propose an extension of a seminal model in this literature. We argue that an autoregressive stochastic volatility model may be a better model for asset returns than its GARCH counterpart, and that the choice between the two may have non-trivial consequences in applications. We suggest a new parameterization and propose an estimation procedure based on indirect inference. A drawback of these models is that they are computationally expensive to estimate. We propose a machine learning approach that can drastically speed up estimation.

CO736 Room BH (SE) 2.09 RECENT DEVELOPMENTS IN PERSONAL CREDIT RISK MODELLING Chair: Bart Baesens

C0722: **Predicting credit rating migrations by combining financial, market, and textual data**
Presenter: **Manon Reusens**, KU Leuven, Belgium
Co-authors: Kamesh Korangi, Seppe vanden Broucke, Christophe Mues, Cristian Bravo, Bart Baesens

Estimating credit risk is critical for all entities in business. Therefore, several proxies for the creditworthiness of companies exist, for example, credit ratings issued by credit rating agencies, and much research is conducted into the predictive factors of credit risk. In literature, credit risk is often estimated using financial data. However, we predict credit rating migrations by combining financial data, market data, and textual data. Moreover, we analyze how text can facilitate the predictions of these migrations and what its added value is on top of solely using financial and market data. Furthermore, we also compare different ways to incorporate textual data. In the past, text was often summarized into so-called NLP scores, like subjectivity and polarity, before predicting credit risk. However, through recent advances in the field of Natural Language Processing, full texts can easily be incorporated with pretrained transformer models. We compare multiple models and further analyze the effects of using full text over only incorporating these summarizing NLP scores.

C0919: **A causal perspective on managing credit risk**
Presenter: **Christopher Bockel-Rickermann**, KU Leuven, Belgium

For financial intermediaries, clients not repaying debt pose a substantial risk to their businesses. Therefore, credit risk scoring is a crucial task for lenders. To date, academic research has provided many approaches and methodologies to tackling this problem. However, much of the research has so far focused on calculating the default risk for a specific borrower-loan pair and not on the outcomes of potential counterfactuals. This is likely restricting the implementability of models in credit operations. Lenders typically have significant discretion about the terms of a credit before and during origination. Accurate methods and models to aid during origination, however, are scarce. Hence, the problem of loan origination and its impact on default risks is stated as a problem of causal inference. The "debt ratio" of a credit is considered a treatment and causal machine learning is used to identify the individualized effect of the debt ratio of clients on their default probability. The resulting model may help practitioners to balance client needs and business risks and is assessed on a semi-synthetic credit risk data set. The approach aims to provide a new perspective on credit risk research and new tools to both practitioners and researchers alike.

C1826: **Credit scoring with dynamic multilayer graph neural networks**
Presenter: **Maria Oskarsdottir**, Reykjavik University, Iceland
Co-authors: Cristian Bravo, Christophe Mues, Kamesh Korangi, Sahab Zandi

Credit scoring is one of the oldest applications of data analytics, where lenders use credit scores to help adjudge the risk involved in granting a loan. For most people, access to credit is necessary to support financial wellness, and an acceptable credit score is also required for access to several financial services. While the presence of default correlation has been suspected for a long time, it has only recently been studied to some extent, with the help of network science. Borrowers, in particular, can be connected in different ways, and represented with multilayer networks to reflect various kinds of connections. We present a model for credit risk assessment leveraging a dynamic multilayer network built from a Graph Neural Network and a Recurrent Neural Network. We test our methodology in agricultural lending where sources of connections are geographic location and economic activity of borrowers. The proposed model considers different types of connections between the borrowers as well as the evolution of these connections over time. Preliminary results demonstrate that, when it comes to predicting probability of default of the borrowers, our proposed model brings both better results and interesting insights compared to traditional methods.

C2007: **Geolocation-aware credit risk modeling**
Presenter: **Margot Geerts**, KU Leuven, Belgium
Co-authors: Jochen De Weerdt, Seppe vanden Broucke

Although financial and nonfinancial data are considered key to corporate credit risk modeling, network data and supply chain data have been recently adopted as well. However, more recent research on corporate credit risk shows that credit ratings also depend on location. Corporate credit risk is approached as a geospatial problem. The geolocation of firms' headquarters allows for assessing the location effect of a company's credit risk at a finer level than in previous research. Tree-based methods consistently rank among the best-performing models for tabular data, including this application. Yet, currently available decision tree learning algorithms are suboptimal for geospatial problems. First, conventional decision tree learners are restricted to axis-parallel boundaries. For data sets including $X-Y$-coordinates, this leads to unnatural decision boundaries. Second, decision tree learners are insufficiently tailored to operate well on heterogeneous data. This creates a strong need for a tailored geospatial decision tree learning algorithm with more appropriate splits. Two multivariate decision tree splitters are proposed: diagonal splits and Gaussian splits. The former includes linear combinations of features in the set of candidate splits, and the latter approximates the decision boundary by a Gaussian and splits around it. As this introduces intractability in finding the optimal split, heuristic optimization is leveraged to achieve higher performance and scalability.

CO076 Room BH (SE) 2.10 TOPICS IN FINANCIAL ECONOMETRICS Chair: Leopold Soegner

C1545: **Index insurance and catastrophe bonds for coping with agriculture risk in a multi-region setting**
Presenter: **Christine Oetjen**, Technical University of Kaiserslautern, Germany

Systemic risk can cause large losses for index crop insurers. Motivated by recent studies on agriculture risk, we explore how regional diversification and securitization using catastrophe bonds can reduce these losses. For this, we consider an equilibrium model to optimize the expected utility from all parties. The model is applied to data from Chinese rice farmers. The statistical analysis shows that regional diversification can improve the insurer's expected utility. However, this effect is smaller and premiums increase when there is a large positive correlation between the yield of two regions. Catastrophe bonds increase the insurer's expected utility and decrease the premiums. With the help of simulations, we will look at the robustness of the model, and we will see that securitization reduces the insurer's risk of extreme losses. Combining both methods, regional diversification and securitization, delivers promising results.

C1644: **Convergence of optimal strategies in a continuous-time financial market with model uncertainty on the drift**
Presenter: **Joern Sass**, University of Kaiserslautern, Germany
Co-authors: Dorothee Westphal

In financial markets, simple portfolio strategies often outperform more sophisticated optimized ones. E.g., in a one-period setting, the equal weight or $1/N$-strategy often provides more stable results than mean-variance-optimal strategies. This is due to the estimation error for the mean and can be rigorously explained by showing that for increasing uncertainty on the means, the equal weight strategy becomes optimal, which is due to its robustness. We extend this result to continuous-time strategies in a multivariate Black-Scholes-type market. To this end, we derive optimal trading strategies for maximizing the expected utility of terminal wealth under CRRA utility when we have Knightian uncertainty on the drift, meaning that the only information is that the drift parameter lies in an uncertainty set. The investor takes this into account by considering the worst possible drift within this set. We show that a minimax theorem holds which enables us to find the worst-case drift and the optimal robust strategy quite explicitly. We derive the limits when uncertainty increases and show that a uniform strategy is asymptotically optimal. We also discuss a financial market with a stochastic drift process, combining the worst-case approach with filtering techniques. In this setting, we show how an ellipsoidal uncertainty set can be defined based on the filters, and we demonstrate that investors need to choose a robust strategy to profit from additional information.

C1655: **Bayesian reconciliation of the return predictability**
Presenter: **Borys Koval**, Vienna University of Economics and Business, Austria
Co-authors: Leopold Soegner, Sylvia Fruehwirth-Schnatter

A stable vector autoregressive (VAR) system comprising asset returns, the dividend-price ratio, and dividend growth allows one to pin down the question of return predictability to the value of one particular parameter of a restricted VAR model. This restricted VAR model is used, and return predictability is investigated in a Bayesian context. We adapt two new priors, a Jeffrey's prior and a prior based on the reduced-bias estimator, and compare our Bayesian estimation routine to other Bayesian (e.g., uniform and Reference prior) and frequentist approaches proposed in the literature by means of an extensive simulation study. In terms of root mean square error (RMSE), mean absolute error (MAE), and credible interval coverage, the approach proposed in this article leads to superior performance relative to ordinary least squares estimation, a Frequentist reduced-bias approach, and Bayesian estimation using priors proposed in the literature. We apply our methodology to S&P 500 data and find strong evidence for return predictability after properly accounting for the correlation structure and imposing theory-motivated restrictions on the dividend-price ratio.

C1850: **Nowcasting the Austrian economy with mixed-frequency VAR models**
Presenter: **Ines Fortin**, Institute for Advanced Studies, Austria
Co-authors: Jaroslava Hlouskova

Having information on GDP more timely than provided by national statistical offices, usually released with a one-quarter lag, is often desirable. This is why economists have started to build models exploiting data available at higher frequencies. These data are used to nowcast the present, in the most efficient way possible and without a loss of information. The related models explicitly rely on data at different frequencies. We nowcast Austrian GDP, investment, consumption, and exports/imports. Current research suggests that the use of large models with parameter shrinkage should provide good nowcast accuracy. However, this also requires the continuous management of big datasets which are updated at different points in time and goes along with substantial computational time. We rather build small models with few high-frequency variables, which nowcast as well as possible, applying a mixed-frequency VAR model. As benchmark models, we consider the traditional (quarterly) VAR model as well as univariate AR models and MIDAS regressions. We also provide a measure of uncertainty surrounding our nowcasts.

C2042: **About the parameterisation of hypertall transfer functions**
Presenter: **Philipp Gersing**, Vienna University of Technology, Austria

A parameterisation theory is provided for what is called hypertall rational transfer functions $k(z)$ of dimension $(n \times q)$ with $n >> q$, where the entries are rational functions. Such transfer functions appear, for example, when modeling the common component of static and dynamic factor sequences. We introduce echelon realisations for so-called noise-free representations of hypertall transfer functions. In a noise-free realisation, no error term appears in the observation equation, as is the case in the usual parameterisations of factor model models. We relate the noise-free echelon realisation to the standard realisation and show that generically factor models follow an AR(1) process. The ultimate goal in applications is to use state space realisations where small errors in the autocovariance result in small errors in the ultimate purpose of the application, e.g. forecasting. The hope is that through a better understanding of the structure of hypertall transfer functions, we can come up with representations that result in better forecasting performance.

CO694 Room BH (SE) 2.12 ANOMALY DETECTION AND FORECASTING USING MACHINE LEARNING METHODS Chair: Artem Prokhorov

C0252: **Anomaly detection with kernel density estimation on manifolds**
Presenter: **Fan Cheng**, Monash University, Australia
Co-authors: Anastasios Panagiotelis, Rob Hyndman

Manifold learning can be used to obtain a low-dimensional representation of the underlying manifold given the high-dimensional data. However, kernel density estimates of the low-dimensional embedding with a fixed bandwidth fail to account for the way manifold learning algorithms distort the geometry of the underlying Riemannian manifold. We propose a novel kernel density estimator for any manifold learning embedding by introducing the estimated Riemannian metric of the manifold as the variable bandwidth matrix for each point. The geometric information of the manifold guarantees a more accurate density estimation of the true manifold, which subsequently could be used for anomaly detection. To compare our proposed estimator with a fixed-bandwidth kernel density estimator, we run two simulations with 2-D metadata mapped into a 3-D swiss roll or twin peaks shape and a 5-D semi-hypersphere mapped in a 100-D space, and demonstrate that the proposed estimator could improve the density estimates given a good manifold learning embedding and has higher rank correlations between the true and estimated manifold density. A shiny app in R is also developed for various simulation scenarios. The proposed method is applied to density estimation in statistical manifolds of electricity usage with the Irish smart meter data. This demonstrates our estimator's capability to fix the distortion of the manifold geometry and to be further used for anomaly detection in high-dimensional data.

C1957: **A machine learning attack on illegal trading**
Presenter: **Artem Prokhorov**, University of Sydney, Australia
Co-authors: Henry Leung, Robert James

An adaptive framework is designed for the detection of illegal trading behavior. Its key component is an extension of a pattern recognition tool, originating from the field of signal processing and adapted to modern electronic systems of securities trading. The new method combines the flexibility of dynamic time warping with contemporary approaches from extreme value theory to explore large-scale transaction data and accurately identify illegal trading patterns. Importantly, our method does not need access to any confirmed illegal transactions for training. We use a high-frequency order book dataset provided by an international investment firm to show that the method achieves remarkable improvements over alternative approaches in the identification of suspected illegal insider trading cases.

C1980: **Big machine learning models that learn to estimate and forecast a class of autoregressive processes**
Presenter: **Pablo Montero Manso**, University of Sydney, Australia

Machine Learning/AI models are complex enough to learn not only how to predict a single individual autoregressive process, but the whole class of autoregressive models. For example, suppose that a given class of autoregressive models of fixed order is optimally estimated by ordinary linear least squares; that is, all potential processes within the class are linear autoregressive of a fixed order, and the 'optimal' estimation in terms of mean squared error is to fit the parameters by least squares. A machine learning model that has been trained from a large set of time series of that class can learn an estimation algorithm, potentially replicating the least squares algorithm itself. We will present: 1) An analytical result that shows how a whole class or linear autoregressive models can be optimally estimated and predicted by a nonlinear model that does not require fit to each individual time series. 2) Specific machine learning models that are trained on millions of simulated time series (AR and VAR) that learn an 'optimal' estimation procedure for the time series of that class. 3) Then compare the predictive performance of these models on new time series (without fitting them to data) against traditional estimation methods, such as the Box-Jenkins methodology.

C1870: **Stochastic frontier analysis for (co)integrated panels**
Presenter: **Anton Skrobotov**, Russian Presidential Academy of National Economy and Public Administration and SPBU, Russia
Co-authors: Artem Prokhorov

The standard stochastic frontier analysis is extended to allow integrated and cointegrated variables. We developed the statistical inference for the cointegrating vector under non-symmetric behaviour of the regression errors. Monte-Carlo evidence demonstrates the importance of taking into account cointegration.

Sunday 18.12.2022 13:35 - 15:15 Parallel Session I – CFE-CMStatistics

EO120 Room S-2.23 TOPICS IN DIMENSION REDUCTION Chair: Tatyana Krivobokova

E0732: **Uniformly valid inference based on the Lasso in linear mixed models**
Presenter: **Peter Kramlinger**, UC Davis, United States
Co-authors: Ulrike Schneider, Tatyana Krivobokova

Linear mixed models (LMMs) are suitable for clustered data and are common in, e.g., biometrics, medicine, or small area estimation. It is of interest to perform valid inference after selecting a subset of available variables. We construct confidence sets for the fixed effects in Gaussian LMMs based on the Lasso, which allows quantifying the joint uncertainty of variable selection and estimation. To this end, the properties of REML are used to separate the estimation of the regression coefficients and covariance parameters. We derive an appropriate normalizing sequence from proving the uniform Cramer consistency of the REML estimator. We then show that the resulting confidence sets for the fixed effects are uniformly valid over the parameter space of both the regression coefficients and the covariance parameters. Their superiority to naive confidence sets is validated in simulations and illustrated with a study of the acid neutralization capacity of U.S. lakes.

E0793: **Iterative regularization methods for ill-posed generalized linear models**
Presenter: **Gianluca Finocchio**, University of Vienna, Austria
Co-authors: Tatyana Krivobokova

The problem of regularized maximum-likelihood estimation in ill-posed generalized linear models is studied. Ill-posedness is assumed to be the byproduct of a low-dimensional latent factor model. We provide a class of iterative algorithms extending known penalization/projection techniques and obtain theoretical guarantees under regularity assumptions on the latent model. In particular, when the number of features and observations are both large, we propose a novel iteratively-reweighted-partial-least-squares algorithm outperforming its competitors in both computational efficiency and minimum-norm maximum-likelihood estimation. Our findings are confirmed by simulation studies on both real and generated data.

E1059: **Sufficient reductions in regression with mixed predictors**
Presenter: **Efstathia Bura**, Vienna University of Technology, Austria
Co-authors: Liliana Forzani, Rodrigo Garcia Arancibia, Pamela Llop, Diego Tomassi

Most data sets comprise measurements of continuous and categorical variables. Modelling high-dimensional mixed predictors has received limited attention in regression and classification Statistics literature. We study the general regression problem of inferring a variable of interest based on high dimensional mixed continuous and binary predictors. The aim is to find a lower dimensional function of the mixed predictor vector that contains all the modeling information in the mixed predictors for the response, which can be either continuous or categorical. The approach we propose identifies sufficient reductions by reversing the regression and modeling the mixed predictors conditional on the response. We derive the maximum likelihood estimator of the sufficient reductions, asymptotic tests for dimension, and a regularized estimator, which simultaneously achieves variable (feature) selection and dimension reduction (feature extraction). We study the performance of the proposed method and compare it with other approaches through simulations and real data examples.

E1250: **A fresh look at sparse quantile regression**
Presenter: **Paulo Serra**, VU Amsterdam, Netherlands
Co-authors: Alexandra Vegelien

In statistics, the aim is often to discover (sometimes impose) structure on observed data, and dimension plays a crucial role in this task. For instance, high-dimensional data sometimes live in a lower-dimensional space, and sparse models are a popular way to represent this. Sparse quantile regression combined with appropriate penalties produces sparse, robust estimators. We will share some results about what kinds of advantages sparse quantile regression brings over mean-based estimators, particularly in terms of robustness, support recovery, correlation between observations and in the design, asymmetric and fat-tailed distributions, and models with quantile level dependent sparseness.

EO324 Room S-1.01 TEXT ANALYSIS FOR COMPLEX DATA Chair: Annamaria Bianchi

E1085: **Using web site text to identify different types of companies**
Presenter: **Piet Daas**, Eindhoven University of Technology, Netherlands

Different types of companies are identified based on - differences in - the texts on their websites. This approach has been used to identify innovative and platform economy companies in the Netherlands and drone companies in several European countries. Usually, an initial test is performed to determine if (and how much) the website texts for the topic studied actually differ. For this, at least 2000 company website texts, including 50% positive and 50% negative cases, are routinely used. Survey data or expert findings are used to determine the actual type of company. Next, the website's texts are preprocessed, and various classification algorithms, included in the scikit-learn library of Python, are applied to determine which of them is best able to discern between the positive and negative cases; e.g. platform vs non-platform. In addition, the effect of adding WordEmbeddings-based features is routinely tested. We found that logistic regression with WordEmbeddings worked best to detect innovation (accuracy 88%), linear-SVM worked best for platform economy websites (accuracy 82%), and logistic regression worked best to detect drone companies (accuracy 82-86%). The results are usually used to obtain a small subset of companies for the type studied that are subsequently investigated further.

E1668: **Automated classification for open-ended questions with BERT**
Presenter: **Jay Gweon**, Western University, Canada

Manual coding of text data from open-ended questions into different categories is time-consuming and expensive. Automated coding uses statistical/machine learning to train on a small subset of manually coded text answers. Recently, pre-training a general language model on vast amounts of unrelated data and then adapting the model to the specific application has proven effective in natural language processing. Using two data sets, we empirically investigate whether BERT, the currently dominant pre-trained language model, is more effective at automated coding of answers to open-ended questions than other non-pre-trained statistical learning approaches. We found fine-tuning the pre-trained BERT parameters is essential as otherwise, BERT's is not competitive. Second, we found fine-tuned BERT barely beats the non-pre-trained statistical learning approaches in terms of classification accuracy when trained on 100 manually coded observations. However, BERT's relative advantage increases rapidly when more manually coded observations (e.g. 200-400) are available for training. We conclude that for automatically coding answers to open-ended questions, BERT is preferable to non-pre-trained models such as support vector machines and boosting.

E1859: **Unstructured textual data and composite indicators construction**
Presenter: **Camilla Salvatore**, University of Milano-Bicocca, Italy
Co-authors: Annamaria Bianchi, Silvia Biffignandi

This paper presents a novel approach for constructing indicators using social media data in order to augment traditional data-based indicators and discusses ways to modify the input based on social media data.Topic modelling is considered in order to identify the proportion of text related to the phenomenon under consideration. These proportions are the input for the indicator construction. However, topic modelling is computationally expensive and might be difficult to interpret for final data users. A more natural approach for the construction of composite indicators is the use of

dictionary-based methods. Based on the results of topic modelling (words that are mostly associated with a dimension), we propose to develop a context-specific dictionary that can be used to perform the same analysis more efficiently. One of the main advantages of the dictionary approach is that it can be easily implemented, interpreted, and updated regularly by experts. In this paper, we propose different methods for constructing dictionaries, considering words, stems, and a list of words augmented by word embeddings. A sensitivity analysis can be performed to determine the stability of the indicator in light of the different approaches.The specific empirical application focuses on measuring corporate social responsibility (CSR) and an original Twitter indicator is developed.

E1602: **Robust and consistent estimation of word embeddings for Italian tweets**
Presenter: **Mauro Bruno**, Istat, Italy
Co-authors: Elena Catanese

Word Embeddings (WEs) are a popular class of models for word representations used in many natural language processing tasks. They provide a low-dimensional, dense vector representation of a word. WEs are usually generated from a large corpus of structured text, e.g., Wikipedia. Tweets are short, noisy and have lexical and semantic features that are different from other types of texts. To our knowledge, extensive robustness analysis on Twitter data has not been carried out. There are several WE methods, such as, Word2Vec, Fasttext and Glove. Each method requires a fine-tuning of several hyper-parameters. Istat currently utilizes a Word2Vec approach to perform thematic analysis on Twitter data. We measure the stability of each WE method, by varying a pre-defined set of hyper-parameters. Solving word analogies is another popular benchmark for WE, based on the assumption that linear relations between word pairs are indicative of the quality of the embedding. Therefore, we compare the performance of the above-mentioned methods on the prediction of specific analogies related to the economy. Insight into the robustness of WE models and on the consistency of thematic focuses on Italian Twitter corpora will be provided, aiming at producing complementary information for Official Statistics purposes.

EO745 Room S-1.04 RECENT ADVANCES IN COPULA REGRESSION Chair: Paul Bach

E0466: **Vine copula based structural equation models**
Presenter: **Claudia Czado**, Technische Universitaet Muenchen, Germany

While there is considerable effort to construct Bayesian networks from data, there is less emphasis on understanding and quantifying conditional distributions and associated quantities of nodes given their parents from the identified Bayesian network. Often Gaussian structural equation models are utilized, which might be too restrictive. A copula-based and thus non-Gaussian non-linear structural equation model for continuous data is proposed. It utilizes a previous approach based on D-vine copulas. It allows for easy fitting and estimation of conditional quantiles. It includes a forward selection algorithm to select the most important covariates. This will be used to identify edges which can be removed from a given network. This approach will be illustrated for an experimental setting.

E0833: **Copula modelling of serially correlated multivariate data with hidden structures**
Presenter: **Radu Craiu**, University of Toronto, Canada
Co-authors: Robert Zimmerman, Vianey Leos Barajas

A copula-based extension of the hidden Markov model is considered. At each measuring time, a vector of observations is measured for each unit in the sample. The joint model produced by the copula extension allows decoding of the hidden states based on information from multiple observations. The dependence structure is integrated into the likelihood using copulas. This modification brings additional computational challenges, which are tackled using a theoretically justified variation of the EM algorithm developed within the framework of inference functions for margins. The method is illustrated using numerical experiments and a real example.

E0813: **Bivariate mixed outcome-survival additive regression**
Presenter: **Guillermo Briseno Sanchez**, TU Dortmund University, Germany
Co-authors: Andreas Groll

A bivariate regression model is proposed where the response is given by a right-censored survival time and a binary outcome. The continuous survival time is modelled using the piecewise-exponential approach, i.e. a discrete-time survival model with a suitably augmented dataset is employed. The parameters of the bivariate distribution are modelled using additive predictors given by a linear combination of suitable representations of covariates $\mathbf{x}_{k,i}$ and regression coefficients β_k, where $k = 1,\dots,K$ indexes the bivariate distribution parameters and $i = 1,\dots,n$ indexes the observations in the sample. The augmentation of the dataset required for a piecewise-exponential model results in a sequence of pseudo-observations of length $j = 1,\dots j(i)$ per individual i in the sample. Therefore, the model for the survival marginal response consists of $n' > n$ rows, which makes the construction and evaluation of a likelihood function not possible. We derive the functions necessary to tackle the aforementioned issue. A joint bivariate density is then constructed using parametric copulae, allowing for the separate specification of the dependence structure and marginal distributions. Model fitting is carried out using trust-regions implemented as a custom extension of the R package GJRM.

E1279: **Multivariate structural distributional time series**
Presenter: **Simone Maxand**, Europa-Universitat Viadrina, Germany
Co-authors: Nadja Klein

A new structural model is proposed for implicit copula time series. Such a generic distributional description of multivariate time series allows studying the interconnection of the involved series in basically all forms while simultaneously capturing the complex and highly non-linear serial dependencies through an implicit copula process. The latter is high-dimensional, but estimation is possible efficiently through Bayesian inference. We illustrate the new method on electricity demand and price data from Germany. We derive electricity price elasticities and short-term probabilistic forecasts for price and demand, both of which are crucial quantities for the efficient operation of energy markets.

EO598 Room S-1.06 MIXTURE MODELLING Chair: Sollie Millard

E0344: **A unified tool for the root selection and the hypothesis testing for mixture models**
Presenter: **Weixin Yao**, UC Riverside, United States

The aim is to introduce how to apply goodness of fit (GOF) statistics to choose a consistent root for finite mixture models. Our new method inherits both the consistency properties of distance estimators and the efficiency of the MLE. The new method is simple to use, and its computation can be easily done using existing R packages for mixture models. In addition, we will also introduce how to apply the GOF test statistics to perform hypothesis testing and model selection for finite mixture models. The limiting distribution of test statistics is simulated based on a bootstrap method. It is demonstrated through extensive empirical studies that a simple application of GOF test statistics to finite mixture models can provide comparable or even superior hypothesis testing performance compared to some existing cutting-edge testing methods.

E0586: **A one-step backfitting algorithm for estimating the semi-parametric mixture of partial linear models**
Presenter: **Sphiwe Skhosana**, University of Pretoria, South Africa
Co-authors: Sollie Millard, Frans Kanfer

The semi-parametric mixture of partial linear models (SPMPLMs) offers flexibility in modelling heterogeneous regression relationships. In addition, it reduces the curse of dimensionality problem. Given a set of covariates, the model assumes that the component regression function (CRF) is a linear combination of a parametric function of some of the covariates and a non-parametric function of the other covariates. In practice, the CRF

is usually estimated at a set of grid points using a local profile likelihood approach via the Expectation-Maximization (EM) algorithm. However, maximizing each local-likelihood function separately does not guarantee that the responsibilities obtained at the E-step of the EM algorithm align at each grid point leading to a label-switching problem. This results in non-smooth CRFs. We propose a modified EM algorithm in the form of a one-step backfitting algorithm to account for the label-switching by tracking the roughness of the CRF. Because of the computational intensity of the one-step procedure, we also propose an alternative plug-in estimation procedure. We use simulation and an application on a real-world data set to demonstrate the performance of both algorithms. In our simulation study, the proposed algorithm performs similarly to, if not better than, competitive methods for all the scenarios investigated.

E1357: **Nonparametric tilted regression estimation**
Presenter: **Seyed Mahdi Salehi**, University of Neyshabur, Iran
Co-authors: Farzaneh Boroumand, Hassan Doosti, Mohammad Taghi Shakeri
Tilting methods are employed for modifying the empirical distribution by replacing the uniform distribution of weights over data with a multinomial distribution. The tilting approach has also been utilized for minimizing the distance to an infinite-order regression estimator. We propose a tilted Nadaraya-Watson estimator and proved that it achieves a higher level of accuracy and, at the same time, preserves interesting properties of the infinite order estimator. We also showed that the tilted estimators are consistent and have desirable convergence rates. In a simulation study, we illustrated that the tilted Nadaraya-Watson estimator has a better performance than its classical version in terms of Median Integrated Squared Error.

E1424: **A fresh consideration for the mode within a mixture context**
Presenter: **JT Ferreira**, University of Pretoria, South Africa
With an astounding increase and necessitated understanding of scale and scope in modern data, flexible modelling remains a sincere and meaningful point of discussion within data analysis. Plentiful work has been studied based on mean/variance/mean-variance type of mixtures; a specific compound approach is of interest with a focus on the mode. Particular positive support distributions are considered in their known reparameterised form where the mode is explicitly available, and acts as a point of departure for studying the construction and viability of a mixture-type approach on this mode. This characterisation allows for unique leverage in terms of modelling by using the mode as an intuitive parameter that may allow for multiple bumps in data, as well as potential intuitive initialisation choices in estimation. The results are illustrated via a real data application.

EO642 Room S-1.22 NEW DEVELOPMENTS IN CENSORED AND TRUNCATED DATA Chair: Rebecca Betensky

E1531: **Regression analysis for censored and truncated event data using pseudo-observations**
Presenter: **Erik Parner**, Aarhus University, Denmark
The pseudo-observation method has become popular for performing regression analysis for censored event data. Pseudo-observations are transformations of the event data; once they are computed, they can be treated as observations for regression analysis, and often standard statistical software can be used for the analysis. Applications include regression analysis for cumulative risk, restricted means and number of life years lost due to specific causes of death. We discuss under which conditions we may expect the pseudo-observation method to provide unbiased estimates. In particular, we consider regression models for cumulative risk in a cohort with left truncation. Some variants of pseudo-observations are also discussed.

E1705: **Nonparametric and semiparametric estimation with sequentially truncated survival data**
Presenter: **Jing Qian**, University of Massachusetts, Amherst, United States
Co-authors: Rebecca Betensky, Jingyao Hou
In observational cohort studies with complex sampling schemes, truncation arises when the time to event of interest is observed only when it falls below or exceeds another random time, i.e., the truncation time. In more complex settings, observation may require a particular ordering of event times; we refer to this as sequential truncation. Estimators of the event time distribution have been developed for simple left-truncated or right-truncated data. However, these estimators may be inconsistent under sequential truncation. We propose nonparametric and semiparametric maximum likelihood estimators for the distribution of the event time of interest in the presence of sequential truncation, under two truncation models. We show the equivalence of an inverse probability-weighted estimator and a product limit estimator under one of these models. We study the large sample properties of the proposed estimators and derive their asymptotic variance estimators. We evaluate the proposed methods through simulation studies and apply the methods to an Alzheimer's disease study. We have developed an R package, seqTrun, for the implementation of our method.

E1734: **Transformation methods for smoothed estimation from interval-censored data**
Presenter: **Rebecca Betensky**, New York University, United States
Co-authors: Jing Qian
Interval censoring is common in longitudinal studies of Alzheimer's disease due to observation of monotonic processes at periodic visits, e.g., time to Clinical Dementia Rating scale score of 0.5 and amyloid at a given time. In addition, monotonic AD markers (e.g., amyloid) may not be observed at a time of interest, and thus are interval-censored at that time. Additionally, it may be of interest to estimate the time between events (latency), where the initiating event is interval-censored, such as time from CDR of 0.5 to death. The nonparametric information contained in interval-censored data lies in minimal intersections of the observed intervals. Thus, data that are highly intersecting are less informative than those that are not. We propose a novel solution through a linear transformation of the unobserved event time using a discrete Uniform random variable and a scalar parameter selected to satisfy an independence condition. We then calculate the NPMLE for the distribution of the transformed time, which yields a smoothed estimator of the original time.

E1720: **Statistical methods for doubly truncated data**
Presenter: **Carla Moreira**, University of Minho, Portugal
Truncation is a well-known phenomenon that may be present in observational studies of time-to-event data. For example, when the sample restricts to those individuals with events falling between two particular dates, they are subject to selection bias due to the simultaneous presence of left and right truncation, also known as interval sampling, leading to a double truncation. When time-to-event data is doubly truncated, the sampling information includes the variable of interest X and left-truncation and right-truncation variables U and V, but the observable population reduces to those individuals for which the variable of interest lies between left-truncation and right-truncation variables. In this case, both large and small values of X are observed in principle with a relatively small probability. The problem of estimating the distribution of X and other related curves, such as kernel density and kernel hazard functions, using nonparametric and semiparametric approaches, from a set of iid triplets with the distribution of (X,U,V) given the double truncation restriction will be presented. Several scenarios will be reported where the effect of ignoring double truncation appears in practice. Possible limitations of the nonparametric and semiparametric estimators will be discussed.

EO741 Room S-1.27 TESTING INDEPENDENCE IN HIGH-DIMENSIONAL STATISTICS Chair: Din Chen

E0351: **A likelihood ratio test for independence and a test of fit for the multivariate linear model for high-dimensional data**
Presenter: **Carlos Coelho**, NOVA University of Lisbon, NOVA-Math and NOVA.id.FCT, Portugal
A likelihood ratio test (LRT) that may be used for the test of independence of two sets of variables in the high-dimensional case is developed. The main aim is that this LRT may be used as a test of fit for the Multivariate Linear Model and the multi-way MANOVA model, in cases where the

number of response variables in the analysis is larger than the number of observations, even allowing for the cells to have just one observation, while enabling the user to test for the significance of individual factors and interactions. The aim is also that the test might be used with non-normal and even heavy-tailed and skewed distributions. A Normal asymptotic distribution is obtained for the test statistic and a test for nested models is also obtained.

E0489: **BEAUTY powered BEAST**
Presenter: **Kai Zhang**, University of North Carolina at Chapel Hill, United States
Co-authors: Zhigen Zhao, Wen Zhou

The focus is on nonparametric dependence detection with the proposed Binary Expansion Approximation of Uniformity (BEAUTY) approach, which generalizes the celebrated Euler's formula. It approximates the characteristic function of any copula with a linear combination of expectations of binary interactions from marginal binary expansions. This novel theory enables the unification of many important tests through approximations from some quadratic forms of symmetry statistics, where the deterministic weight matrix characterizes the power properties of each test. To achieve robust power, we study test statistics with data-adaptive weights, referred to as the Binary Expansion Adaptive Symmetry Test (BEAST). By utilizing the properties of the binary expansion filtration, we show that the Neyman-Pearson test of uniformity can be approximated by an oracle-weighted sum of symmetry statistics. The BEAST with this oracle provides a benchmark of feasible power against any alternative by leading all existing tests with a substantial margin. To approach this oracle power, we develop the BEAST through a regularized resampling approximation of the oracle test. The BEAST improves the empirical power of many existing tests against a wide spectrum of common alternatives and provides a clear interpretation of the form of dependency when significant.

E0510: **Model-free multiple testing using mirror statistics (MMM)**
Presenter: **Zhigen Zhao**, Temple University, United States
Co-authors: Xin Xing

The general regression analysis is considered, and the relation between a univariate response and a p-dimensional covariate is studied. We assume the general multi-index model with an unknown link function. It is assumed that the response depends on the covariate via some linear combinations, which is characterized by the central subspace. For all the covariates, we want to test the hypothesis of whether each individual predictor plays any role in the central subspace subject to the control of the false discovery rate. We combine the method of sufficient dimension reduction and the Gaussian mirror to construct the MMM method, standing for Model-free Multiple Testing using Mirror Statistics. It is shown that MMM controls the FDR at the desired level asymptotically. Numerical evidence has shown that MMM is much more powerful than all its alternatives.

E0693: **A data adaptive rank-based procedure for assessing reproducibility of high-throughput experiments**
Presenter: **Wen Zhou**, Colorado State University, United States
Co-authors: Debashis Ghosh, Austin Ellingworth

Reproducibility guarantees the consistency and validity of experimental findings, while the lack of which can lead to negative and even catastrophic effects on scientific discovery. In high-throughput studies, reproducibility has often been identified as hypotheses with coinciding test results across different experiments. The maximum rank statistic (MaRR) was introduced to identify reproducible hypotheses based on the agreement of test results across experiments. Regardless of its empirical success, the theoretical guarantees of MaRR remain largely unknown. We carefully investigate MaRR which lends it to quantifying reproducibility in high-throughput studies. We also develop a novel data adaptive rank-based statistic that balances the signal strength of a hypothesis and its variation across experiments. Based on the new statistic, we design a procedure to assess reproducibility with marginal false discovery rate (mFDR) control. By inspecting the rejection region, we show that the new procedure dominates the original MaRR statistic with superior power. We also present a revealing phase transition phenomenon of our procedure using the bivariate Gaussian canonical model. Using comprehensive simulations, we demonstrate the finite sample performance of our method, which corroborates the theoretical findings.

EO554 Room K0.16 ADVANCES IN NETWORK DATA ANALYSIS Chair: Sharmodeep Bhattacharyya

E0490: **Leave-one-out singular subspace perturbation analysis for spectral clustering**
Presenter: **Anderson Ye Zhang**, University of Pennsylvania, United States

The singular subspaces perturbation theory is of fundamental importance in probability and statistics. It has various applications across different fields. We consider two arbitrary matrices where one is a leave-one-column-out submatrix of the other one and establish a novel perturbation upper bound for the distance between two corresponding singular subspaces. It is well-suited for mixture models and results in a sharper and finer statistical analysis than classical perturbation bounds such as Wedin's Theorem. Powered by this leave-one-out perturbation theory, we provide a deterministic entrywise analysis for the performance of spectral clustering under mixture models. Our analysis leads to an explicit exponential error rate for the clustering of sub-Gaussian mixture models.

E0756: **A framework for modelling multiplex networks**
Presenter: **Swati Chandna**, Birkbeck, University of London, United Kingdom
Co-authors: Svante Janson, Sofia Olhede

Consider the setting where multiple networks are observed on the same set of nodes, also known as multiplex networks. Such data may arise in the form of networks observed over time, e.g., friendship networks, or from different subjects at a given point in time, e.g. structural brain networks from multiple individuals. Given such data, it is crucial to quantify dependence between pairs of networks and have a framework which allows for the empirical description of correlated networks. We describe an approach to modelling multiplex networks with two layers and show how it leads to simpler models that generate correlated networks.

E1350: **Change point localization in dependent dynamic nonparametric random dot product graphs**
Presenter: **Oscar Hernan Padilla**, UCLA, United States

The change point localization problem is studied in a sequence of dependent nonparametric random dot product graphs. To be specific, assume that at every time point, a network is generated from a nonparametric random dot product graph model, where the latent positions are generated from unknown underlying distributions. The underlying distributions are piecewise constant in time and change at unknown locations, called change points. Most importantly, we allow for dependence among networks generated between two consecutive change points. This setting incorporates edge dependence within networks and temporal dependence between networks, which is the most flexible setting in the published literature. To accomplish the task of consistently localizing change points, we propose a novel change point detection algorithm, consisting of two steps. First, we estimate the latent positions of the random dot product model, our theoretical result being a refined version of the state-of-the-art results, allowing the dimension of the latent positions to grow unbounded. Subsequently, we construct a nonparametric version of the CUSUM statistic that allows for temporal dependence. Consistent localization is proved theoretically and supported by extensive numerical experiments, which illustrate state-of-the-art performance.

E1431: **The manifold hypothesis for graphs**
Presenter: **Patrick Rubin-delanchy**, University of Bristol, United Kingdom

A manifold hypothesis for graphs is presented, giving several arguments. The first is empirical, showing that manifold structure appears in high-dimensional embeddings of several real-world networks, and seems to give a distorted view of a true, low-dimensional latent domain. The second

is that under regularity conditions, any latent position network model is equivalent to a manifold hypothesis in inner-product space. The last is that the manifold hypothesis explains how a sparse graph can have a high triangle density. We will also suggest several ways to exploit this hypothesis within statistical procedures.

EO649 Room K0.19 RECENT DEVELOPMENT IN MEDIATION ANALYSIS Chair: Yeying Zhu

E0211: **HIMA2: High dimensional mediation analysis and its application to epigenome-wide DNA methylation data**
Presenter: **Lei Liu**, Washington University in St. Louis, United States

Mediation analysis plays a major role in identifying significant mediators in the pathway between environmental exposures and health outcomes. With advanced data collection technology for large-scale studies, there has been growing research interest in developing a methodology for high-dimensional mediation analysis. We present HIMA2, an extension of the HIMA method. First, the proposed HIMA2 reduces the dimension of mediators to a manageable level based on the sure independence screening (SIS) method. Second, a de-biased Lasso procedure is implemented for estimating regression parameters. Third, we use a multiple-testing procedure to accurately control the false discovery rate (FDR) when testing high-dimensional mediation hypotheses. We demonstrate its practical performance using Monte Carlo simulation studies and apply our method to identify DNA methylation markers which mediate the pathway from smoking to reduced lung function in the Coronary Artery Risk Development in Young Adults (CARDIA) Study.

E0646: **Variable selection for mediation analysis with latent factors via group-wise penalization**
Presenter: **Qing Wang**, Wellesley College, United States
Co-authors: Yeying Zhu, Xizhen Cai

A mediation analysis model is considered where the outcome depends on exposure and a number of latent factors that are related to a set of observable mediators. In addition, we assume that there exist some redundant mediators that do not relate to the factors or outcome. We first apply a penalized factor analysis model with group-wise regularization to uncover the relationship between the mediators and latent factors so as to filter out irrelevant mediators. An expectation-maximization algorithm is employed to fit the penalized factor analysis model. Then, we utilize the attained latent factors to understand their relationship with both the exposure and the outcome through a two-step procedure. Simulations in both low and high-dimensional settings are considered. The results suggest that our proposed model yields a more accurate identification of the true mediators and produces a smaller bias for the mediation model compared to existing methods. Ongoing work is to incorporate simultaneous mediator and factor selection in contrast to the two-stage process.

E0736: **Comparisons of inference methods in high-dimensional mediation analysis**
Presenter: **Xizhen Cai**, Williams College, United States
Co-authors: Yeying Zhu, Yuan Huang

Mediation analysis is a framework to understand how a treatment affects the outcome through intermediate variables, namely mediators. Over the past decades, large and high-dimensional datasets have become easily stored and publicly available. This leads to many recent advances in mediation analysis in developing models to fit more complex data structures and methods for mediator selections in high-dimensional settings. The statistical inference procedure following the mediator selection also serves as an essential step in the mediation analysis. We study the effect of applying different inference procedures after the mediator selection and perform simulation studies to further compare these procedures. We will discuss our simulation settings and the findings to provide guidelines that help distinguish among various approaches, highlight the advantages and disadvantages of each, and identify ones that perform better in certain scenarios.

E1204: **Surrogate marker assessment using mediation analyses in a case-cohort design**
Presenter: **Yen-Tsung Huang**, Academia Sinica, Taiwan

The identification of surrogate markers for gold-standard outcomes in clinical trials enables future cost-effective trials that target the identified markers. Due to resource limitations, these surrogate markers may be collected only for cases and for a subset of the trial cohort, i.e., the case-cohort design. Motivated by a COVID-19 vaccine trial, we propose methods of assessing the surrogate markers for a time-to-event outcome in a case-cohort design by using mediation and instrumental variable (IV) analyses. In the mediation analysis, we decomposed the vaccine effect on COVID-19 risk into an indirect effect (the effect mediated through the surrogate marker such as neutralizing antibodies), and a direct effect (the effect not mediated by the marker), and we propose that the mediation proportions are surrogacy indices. We employed weighted estimating equations derived from nonparametric maximum likelihood estimators (NPMLEs) under semiparametric probit models for the time-to-disease outcome. We plugged in the weighted NPMLEs to construct estimators for the aforementioned causal effects and surrogacy indices, and we determined the asymptotic properties of the proposed estimators. Applying the proposed mediation and IV analyses to a mock COVID-19 vaccine trial data, we found that 84.2% of the vaccine efficacy was mediated by 50% pseudovirus-neutralizing antibody.

EO488 Room K0.20 STATISTICAL METHODS FOR MISSING DATA AND MEASUREMENT ERROR Chair: Baojiang Chen

E0591: **Variable selection and estimation for the average treatment effect with error-prone confounders**
Presenter: **Li-Pang Chen**, National Chengchi University, Taiwan
Co-authors: Grace Yi

In the framework of causal inference, the inverse-probability-weighting estimation method and its variants have been commonly employed to estimate the average treatment effect. Such methods, however, are challenged by the presence of irrelevant pre-treatment variables and measurement errors. Ignoring these features and naively applying the usual inverse probability-weighting estimation procedures may typically yield biased inference results. We develop an inference method for estimating the average treatment effect with those features taken into account. We establish theoretical properties for the resulting estimator and carry out numerical studies to assess the finite sample performance of the proposed estimator.

E0661: **Dirichlet process mixture models for the analysis of repeated attempt designs**
Presenter: **Michael Daniels**, University of Florida, United States

In longitudinal studies, it is not uncommon to make multiple attempts to collect a measurement after baseline. Recording whether these attempts are successful provides useful information for the purposes of assessing missing data assumptions. This is because measurements from subjects who provide the data after numerous failed attempts may differ from those who provide the measurement after fewer attempts. Previous models for these designs were parametric and/or did not allow sensitivity analysis. For the former, there are always concerns about model misspecification, and for the latter, sensitivity analysis is essential when conducting inference in the presence of missing data. We propose a new approach which minimizes issues with model misspecification by using Bayesian nonparametrics for the observed data distribution. We also introduce a novel approach for identification and sensitivity analysis. We re-analyze the repeated attempts data from a clinical trial involving patients with severe mental illness and conduct simulations to understand the properties of our approach better.

E0686: **A calibration method to stabilize the estimation in missing data and causal inference**
Presenter: **Baojiang Chen**, University of Texas, United States
Co-authors: Ao Yuan, Jing Qin

Missing data are commonly available in causal inference, where the marginal means of the outcome in both the treatment and control groups are estimated. The augmented inverse weighting (AIW) estimator was commonly used to estimate the marginal mean of the outcome due to its

doubly robust property. However, the AIW estimator can be severely biased if both the propensity score (PS) and the outcome regression (OR) models are misspecified. One possible reason is that the misspecification of the PS or/and OR model yields some extreme values in these models, which can have a great influence on the marginal mean estimate. We propose a calibrated augmented inverse weighting estimator for the marginal mean, which can control for these extreme values' influence, hence providing a stable marginal mean estimator. The proposed estimator also enjoys the doubly robust property. We also introduce the Box-Cox transformation in the outcome regression model to reduce the possibility of model misspecification. A smearing estimate is used to estimate the conditional mean of the outcome. Finally, we extend this method to handle high dimensional covariates in the PS and OR models. Asymptotic results are also developed. Extensive simulation studies demonstrate that the proposed method performs better than peers by providing a more stable estimate. We apply this method to an AIDS clinical trial study.

E0954: **A versatile estimation procedure without estimating the nonignorable missingness mechanism**
Presenter: **Jiwei Zhao**, University of Wisconsin-Madison, United States
The focus is on an estimation problem in a regression setting where the outcome variable is subject to nonignorable missingness, and identifiability is ensured by the shadow variable approach. We propose a versatile estimation procedure where modeling of missingness mechanism is completely bypassed. We show that our estimator is easy to implement and we derive the asymptotic theory of the proposed estimator. We also investigate some alternative estimators under different scenarios. Comprehensive simulation studies are conducted to demonstrate the finite sample performance of the method. We apply the estimator to a children's mental health study to illustrate its usefulness.

EO422 Room K0.50 RECENT ADVANCES IN NONPARAMETRIC METHODS Chair: Bojana Milosevic

E1132: **Computational and statistical limits in high dimensional independent component analysis**
Presenter: **Arnab Auddy**, Columbia University, United States
Co-authors: Ming Yuan
The independent components analysis (ICA) model is a popular semiparametric model where one observes a d-dimensional vector $X = AS$ for an unknown invertible mixing matrix A and a random vector S consisting of independent components. Despite its usage in a variety of applications, existing statistical results in such models are restricted to the case of fixed dimension d. We will address the issues of computability and statistical inference in the ICA model when d is allowed to grow. We will first see that there exists a computational limit, in terms of the sample size n and the dimension d, below which it is computationally hard to recover any column of A. On the other hand, if we are above this limit, it is, in fact, possible to estimate the columns of A at a parametric rate, without estimating the unknown marginal distributions of S. Additionally, we show that our estimators are asymptotically normal (for sufficiently large d and n) whenever we are above the computational limit.

E1199: **Least-squares estimation of a quasiconvex regression function**
Presenter: **Rohit Patra**, LinkedIn, United States
Co-authors: Somabha Mukherjee
A new approach is developed for the estimation of a multivariate function based on the economic axioms of quasiconvexity (and monotonicity). On the computational side, we prove the existence of the quasiconvex-constrained least squares estimator (LSE) and provide a characterization of the function space to compute the LSE via a mixed integer quadratic programme. On the theoretical side, we provide finite sample risk bounds for the LSE via a sharp oracle inequality. Our results allow for errors to depend on the covariates and to have only two finite moments. We illustrate the superior performance of the LSE against some competing estimators via simulation. Finally, we use the LSE to estimate the production function for the Japanese plywood industry and the cost function for hospitals across the US.

E1773: **Kernel PCA for multivariate extremes**
Presenter: **Marco Avella-Medina**, Columbia University, United States
Co-authors: Richard Davis, Gennady Samorodnitsky
Kernel PCA is proposed as a method for analyzing the dependence structure of multivariate extremes, and it is demonstrated that it can be a powerful tool for clustering and dimension reduction. We provide some theoretical insights into the preimages obtained by kernel PCA, demonstrating that under certain conditions, they can effectively identify clusters in the data. We build on these new insights to characterize rigorously the performance of kernel PCA based on an extremal sample, i.e., the angular part of random vectors for which the radius exceeds a large threshold. More specifically, we focus on the asymptotic dependence of multivariate extremes characterized by the angular or spectral measure in extreme value theory and provide a careful analysis in the case where the extremes a generated from a linear factor model. We give theoretical guarantees on the performance of kernel PCA preimages of such extremes by leveraging their asymptotic distribution and Davis-Kahan perturbation bounds. Our theoretical findings are complemented by numerical experiments illustrating the finite sample performance of our methods.

E0183: **Pointwise error bounds for fused lasso**
Presenter: **Sabyasachi Chatterjee**, University of Illinois at Urbana Champaign, United States
An element-wise error bound is obtained for the Fused Lasso estimator for any general convex loss function ρ. We then focus on the special cases when either ρ is the square loss function (for mean regression) or is the quantile loss function (for quantile regression) for which we derive new point-wise error bounds. Even though error bounds for the usual Fused Lasso estimator and its quantile version have been studied before, our bound appears to be new. This is because all previous works bound a global loss function like the sum of squared error or a sum of Huber losses in the case of quantile regression. Clearly, element-wise bounds are stronger than global loss error bounds as it reveals how the loss behaves locally at each point. Our element-wise error bound also has a clean and explicit dependence on the tuning parameter, which informs the user of a good choice of ρ. In addition, our bound is non-asymptotic with explicit constants and is able to recover almost all the known results for Fused Lasso (both mean and quantile regression) with additional improvements in some cases.

EO584 Room S0.03 SPATIAL TRANSCRIPTOMICS DATA MODELING AND ANALYSIS Chair: Qiwei Li

E1830: **Spectral clustering using gene expression and histology identifies disease-relevant spatial domains in SRT**
Presenter: **Kyle Coleman**, University of Pennsylvania, United States
Co-authors: Jian Hu, Daiwei Zhang, Mingyao Li
Spatially resolved transcriptomics (SRT) provides an unprecedented opportunity to integrate gene expression with histology and spatial location information when studying the disease. A prominent goal in SRT data analysis is the identification of spatial domains through the clustering of SRT spots. We present SpeCTrE (Spectral Clustering using Transcriptomics and H&E Histology), a deep learning-based spectral clustering algorithm for the grouping of SRT spots into spatial domains that are distinct with respect to gene expression and histology. SpeCTrE first employs HIPT to extract spot-level histology features while capturing the long-range histological dependencies among spots. The algorithm then constructs two adjacency matrices representing the transcriptional and histological similarities of each pair of spots. The columns of the transcriptomics adjacency matrix are projected onto the top eigenvectors of the normalized Laplacian of the histology adjacency matrix to obtain a modified transcriptomics adjacency matrix containing histological information useful for clustering. Using this updated adjacency matrix and a multilayer perceptron, SpeCTrE obtains spot-level feature vectors that are used as input for the k-means clustering algorithm. Through analyses of SRT datasets from cancerous tissue sections and extensive benchmark evaluations, we show that SpeCTrE outperforms state-of-the-art spatial clustering methods in separating spots into disease-relevant spatial domains.

E1922: **SPROD: De-noising spatial transcriptomics data based on position and image information**
Presenter: **Bing Song**, University of Texas Southwestern Medical Center, United States
Co-authors: Yunguan Wang, Shidan Wang, mingyi Chen, Yang Xie, Guanghua Xiao, Li Wang, Tao Wang
Spatially resolved transcriptomics (SRT) provides gene expression close to, or even superior to, single-cell resolution while retaining the physical locations of sequencing and often also providing matched pathology images. However, SRT expression data suffer from high noise levels, due to the shallow coverage in each sequencing unit and the extra experimental steps required to preserve the locations of sequencing. Fortunately, such noise can be removed by leveraging information from the physical locations of sequencing, and the tissue organization reflected in corresponding pathology images. We develop Sprod, based on latent graph learning of matched location and imaging data, to impute accurate SRT gene expression. We validate Sprod comprehensively and demonstrate its advantages over previous methods for removing drop-outs in single-cell RNA-sequencing data. We show that, after imputation by Sprod, differential expression analyses, pathway enrichment, and cell-to-cell interaction inferences are more accurate. Overall, we envision de-noising by Sprod to become a key first step towards empowering SRT technologies for biomedical discoveries.

E1923: **iIMPACT: Integrating image and molecular-based profiles to analyze and cluster spatial transcriptomics data**
Presenter: **Qiwei Li**, The University of Texas at Dallas, United States
Co-authors: Xi Jiang, Guanghua Xiao, Lin Xu
The breakthrough in spatial transcriptomics (ST) has enabled comprehensive molecular characterization at the cellular level while preserving spatial information. Meanwhile, pathology imaging powered by artificial intelligence enables the histology characterization of single cells. Understanding the spatial organization of cells and their heterogeneous gene expression profiles will provide deeper biological insights. To address these two problems, we develop iIMPACT, a multi-stage method to cluster and analyze ST data. The first stage is an interpretable Bayesian mixture model, which combines a Gaussian component to model the molecular profile and a multinomial component for cell abundances, and incorporates the spatial information by a Markov random field prior. After region segmentation, we develop a zero-inflated generalized linear regression model under the Bayesian framework to study the association between the cellular pattern and gene expression. Applying our method to a publicly available breast cancer dataset, we found that iIMPACT outperforms existing clustering methods in terms of segmentation accuracy and generates the most biologically meaningful cancer-related genes.

E1784: **RUV statistical methods based on generalised linear models for omics data**
Presenter: **Alysha De Livera**, La Trobe University, Australia
Co-authors: Terry Speed, Agus Salim
Unwanted variations in omics data, not only inevitably arise from various technical sources, such as the use of multiple analytical platforms, batches, laboratories, long-run of samples and temperature changes within instruments, but also from unwanted biological variations, such as different cell sizes which are often unmeasurable. Failure to carry out a suitable approach to removing unwanted variation (RUV) in the statistical analysis of omics data, leads to increases in Type I and Type II errors, spurious correlation, as well as artificial clustering and poor classification of the biological samples. Over the last decade, RUV statistical methods have been established as popular, widely-used methods in multiple omics fields for removing unwanted variation. We will go through the latest developments in the RUV methods based on generalised linear models, demonstrate their applications to RNA-sequencing and single-cell data, and compare their performance with the existing methods as well as RUV counterparts which are based on linear models.

EO594 Room S0.13 MODAL INFERENCE Chair: Rosa Crujeiras

E0827: **Nonparametric test for density modes**
Presenter: **Federico Ferraccioli**, Joint Research Centre - European Commission, Italy
Co-authors: Giovanna Menardi
A nonparametric resampling procedure is proposed to test the significance of a mode, with the aim of evaluating whether a region of relatively high observed density reflects the actual presence of a mode in the true distribution underlying a set of data. The method leverages on Morse theory and stochastic gradient methods to characterize the local properties of the modes. This allows the definition of an asymptotic test, based on the concept of gradient ascent paths and relying on resampling methods, to approximate the distribution of the test statistic under the null hypothesis.

E1049: **On robustness issues in modal clustering**
Presenter: **Giovanna Menardi**, University of Padova, Italy
Co-authors: Marco Rudelli, Luca Greco
Deviations from model assumptions, along with the presence of a certain amount of outlying observations, are common in many practical statistical applications. Clustering techniques make no exception, yet some caution is required in this context. First, small clusters could be mistaken for outlying observations, or vice versa. Second, the concept of outlier itself shall be defined with respect to a cluster, rather than the entire data set, and depends on the considered notion of cluster. While robust methods have been proposed in both distance- and model-based clustering, the issue has been largely neglected in the modal framework. Clusters are associated with the domains of attraction of the modes of the density underlying data. Nonparametric methods, usually employed for density (and hence modes) estimation, are known to be vulnerable to the presence of outliers, and prone to the sparsity of data in high dimensions, as much of the probability mass is led to flow to the tails of the density, possibly giving rise to the birth of spurious modes. Robustness issues are discussed in this framework, and suitable measures to flag outliers are explored, especially with a view to trimming methods for modal clustering.

E1625: **Mode and ridge detection on a sphere**
Presenter: **Yen-Chi Chen**, University of Washington, United States
Directional data consist of observations distributed on a (hyper)sphere, and appear in many applied fields, such as astronomy, ecology, and environmental science. Both statistical and computational problems of kernel smoothing for directional data are studied. We generalize the classical mean shift algorithm to directional data, which allows us to identify local modes of the directional kernel density estimator (KDE). The statistical convergence rates of the directional KDE and its derivatives are derived, and the problem of mode estimation is examined. We also prove the ascending property of the directional mean shift algorithm and investigate a general problem of gradient ascent on the unit hypersphere. To demonstrate the applicability of the algorithm, we evaluate it as a mode clustering method on both simulated and real-world data sets.

E1791: **Estimating the number of directional clusters from density-based methods**
Presenter: **Paula Saavedra-Nieves**, Universidade de Santiago de Compostela, Spain
Co-authors: Rosa Crujeiras
Set estimation is focused on the reconstruction of a set (or the estimation of any of its features) from a random sample of points. Target sets to be estimated appear in different contexts, but from a distribution-based perspective, level set estimation is a problem of interest. Actually, this theory is also linked to clustering methods: the number of population clusters is defined as the number of connected components of density level sets. This topic has received some attention in the literature, especially for densities supported in a Euclidean space. However, this clustering approach can be easily extended to more general settings such as the circle or the sphere. We derive some methodology for estimating the number of directional clusters as the number of connected components of directional level sets. An extensive simulation study shows the performance of the proposed estimator for densities supported on the unit circle and the sphere.

EO557 Room Safra Lecture Theatre STATISTICAL INFERENCE ON FUNCTIONAL TIME SERIES Chair: Siegfried Hoermann

E0263: **Dynamic factor model for functional time series: Identification, estimation, and prediction**
Presenter: **Nazarii Salish**, University Carlos III de Madrid, Spain
Co-authors: Sven Otto

A functional dynamic factor model for time-dependent functional data is proposed. We decompose a functional time series into a predictive low-dimensional common component consisting of a finite number of factors and an infinite-dimensional idiosyncratic component that has no predictive power. The conditions under which all model parameters, including the number of factors, become identifiable are discussed. Our identification results lead to a simple-to-use two-stage estimation procedure based on functional principal components. As part of our estimation procedure, we solve the separation problem between the common and idiosyncratic functional components. In particular, we obtain a consistent information criterion that provides joint estimates of the number of factors and dynamic lags of the common component. Finally, we illustrate the applicability of our method in a simulation study and to the problem of modeling and predicting yield curves. In an out-of-sample experiment, we demonstrate that our model performs well compared to the widely used term structure Nelson-Siegel model for yield curves.

E0950: **White noise testing for functional time series**
Presenter: **Mihyun Kim**, West Virginia University, United States
Co-authors: Gregory Rice, Piotr Kokoszka

White noise tests are reviewed in the context of functional time series, and compare many of them using the custom-developed R package wwntests. The tests are categorized based on whether they are conducted in the time domain or spectral domain, and whether they are valid for i.i.d. or general uncorrelated noise. We also review and extend several residual-based goodness-of-fit tests of popular models used in functional data analysis. Through numerous simulation experiments and a data application, we demonstrate the use of these tests, and are able to provide practical guidance on their implementation, benefits, and drawbacks.

E1315: **Estimation of functional ARMA models**
Presenter: **Thomas Kuenzer**, Graz University of Technology, Austria

Functional auto-regressive moving average (FARMA or ARMAH) models allow for flexible and natural modelling of functional time series. While there are many results on pure autoregressive (FAR) models in Hilbert spaces, results on estimation and prediction of FARMA models are considerably more scarce. We devise a simple two-step method to estimate ARMA models in separable Hilbert spaces. Estimation is based on dimension-reduction using principal components analysis of the functional time series. We establish consistency of the proposed estimators under simple assumptions by employing a data-driven criterion to select the dimensionality of the principal component subspaces used in the estimation procedure. The empirical performance of the estimation algorithm is evaluated in a simulation study, where it performs better than competing methods.

E1684: **The maximum of the periodogram of a sequence of functional data**
Presenter: **Vaidotas Characiejus**, University of Southern Denmark, Denmark
Co-authors: Siegfried Hoermann, Clement Cerovecki

The detection of periodic signals in functional time series is investigated when the length of the period is not assumed to be known. A natural test statistic for the detection of periodicities is the maximum overall fundamental frequencies of the Hilbert-Schmidt norm of the periodogram operator. Using recent advances in Gaussian approximation theory, we show that under certain assumptions, the appropriately standardised test statistic belongs to the domain of attraction of the Gumbel distribution. The asymptotic results allow us to construct tests for hidden periodicities. We demonstrate the performance of our methodology in a simulation study, and we also illustrate the usefulness of our approach by examining periodicities in the air quality data from Graz, Austria and showing that our approach is not only able to detect the presence of periodic signals, but it is also able to reveal the structure of periodicities in the data.

EO631 Room Virtual R01 STATISTICAL ADVANCES IN BIOMEDICAL RESEARCH Chair: Bingxin Zhao

E0609: **Orthogonal common-source and distinctive-source decomposition between high-dimensional data views**
Presenter: **Hai Shu**, New York University, United States

Modern biomedical studies often collect multi-view data, that is, multiple types of data measured on the same set of objects. A typical approach to the joint analysis of two high-dimensional data views/sets is to decompose each data matrix into three parts: a low-rank common-source matrix that captures the shared information across data views, a low-rank distinctive-source matrix that characterizes the individual information within each single data view, and an additive noise matrix. Existing decomposition methods often focus on the orthogonality between the common-source and distinctive-source matrices, but inadequately consider the more necessary orthogonal relationship between the two distinctive-source matrices. The latter guarantees that no more shared information is extractable from the distinctive-source matrices. We propose a novel decomposition method that defines the common-source and distinctive-source matrices from the L2 space of random variables rather than the conventionally used Euclidean space, with careful construction of the orthogonal relationship between distinctive-source matrices. The proposed estimators of common-source and distinctive-source matrices are shown to be asymptotically consistent and have reasonably better performance than some state-of-the-art methods in both simulated data and real data analysis.

E0744: **An empirical Bayes regression for multi-tissue EQTL data analysis**
Presenter: **Fei Xue**, Purdue University, United States
Co-authors: Hongzhe Li

The Genotype-Tissue Expression (GTEx) project collects samples from multiple tissues to study the relationship between single nucleotide polymorphisms (SNPs) and gene expression in each tissue. However, most existing eQTL analyses only focus on single tissue information. We develop a multi-tissue eQTL analysis that improves the single-tissue cis-SNP gene expression association analysis by borrowing information across tissues. Specifically, we propose an empirical Bayes regression model for SNP-expression association analysis using data across multiple tissues. To allow the effects of SNPs to vary greatly among tissues, we use a mixture distribution as the prior, which is a mixture of a multivariate Gaussian distribution and a Dirac mass at zero. The model allows us to assess the cis-SNP gene expression association in each tissue by calculating the Bayes factors. We show that the proposed estimator of the cis-SNP effects on gene expression achieves the minimum Bayes risk among all estimators. Analyses of the GTEx data show that our proposed method is superior to traditional regression methods in terms of predicting accuracy for gene expression levels.

E1341: **SUMMIT: An integrative approach for better transcriptomic data imputation improves causal gene identification**
Presenter: **Chong Wu**, The University of Texas MD Anderson Cancer Center, United States

Genes with moderate to low expression heritability may explain a large proportion of complex trait etiology, but such genes cannot be sufficiently captured in conventional transcriptome-wide association studies (TWASs), partly due to the relatively small available reference datasets for developing expression genetic prediction models to capture the moderate to low genetically regulated components of gene expression. We introduce a new method, the Summary-level Unified Method for Modeling Integrated Transcriptome (SUMMIT), to improve the expression prediction model accuracy and the power of TWAS by using a large expression quantitative trait loci (eQTL) summary-level dataset. We apply SUMMIT to the eQTL summary-level data provided by the eQTLGen consortium. Through simulation studies and analyses of genome-wide association study (GWAS)

summary statistics for 24 complex traits, we show that SUMMIT improves the accuracy of expression prediction in blood, successfully builds expression prediction models for genes with low expression heritability, and achieves higher statistical power than several benchmark methods. Finally, we conduct a case study of COVID-19 severity with SUMMIT and identify 11 likely causal genes associated with COVID-19 severity.

E1465: **Robust statistical inference for cell type deconvolution**
Presenter: **Jingshu Wang**, University of Chicago, United States

Cell type deconvolution is a computational approach to infer individual cell types' proportions from bulk transcriptomics data. Most existing methods for cell type deconvolution only provide point estimation of the cell type proportions without any uncertainty quantification, though the estimates can be very noisy due to various sources of biases and randomness. We will discuss a new statistical framework MEAD for cell type deconvolution to get more efficient estimators and construct asymptotically valid confidence intervals both for each individual's cell type proportion and for quantifying how cell type proportions change across multiple bulk individuals in downstream regression analyses. Our statistical inference takes into account the biological randomness of gene expressions across cells and individuals, gene-gene dependence, and cross-platform biases and sequencing errors, without any parametric assumptions. We also provide identification conditions of the cell type proportions when there are arbitrary platforms-specific biases across sequencing technologies.

EO526 Room Virtual R02 RECENT ADVANCES IN STATISTICAL INFERENCE Chair: Monika Bhattacharjee

E0562: **One-step weighting for the generalization of causal inference**
Presenter: **Ambarish Chattopadhyay**, Stanford University, United States
Co-authors: Eric Cohn, Jose Zubizarreta

Weighting methods are often used to generalize estimates of causal effects from a study sample to a target population. Traditional methods construct the weights by separately modeling the treatment assignment and the study selection probabilities and then multiplying functions (e.g., inverses) of the estimated probabilities. These estimated multiplicative weights may not produce adequate covariate balance and can be highly variable, resulting in biased and/or unstable estimators, particularly when there is limited covariate overlap across populations or treatment groups. To address these limitations, we propose a weighting approach for both randomized and observational studies that weights each treatment group directly in 'one go' towards the target population. We present a general framework for generalization problems by characterizing the study and target populations in terms of probability distributions. Under this framework, we justify this one-step weighting approach. By construction, this approach directly balances covariates relative to the target population and produces stable weights. Moreover, this approach does not require individual-level data from the target population. We connect this approach to inverse probability and inverse odds weighting. We show that the one-step weighting estimator for the target average treatment effect is consistent, asymptotically Normal, doubly robust, and semiparametrically efficient.

E0580: **Asymptotics of large autocovariance matrices**
Presenter: **Monika Bhattacharjee**, IIT Bombay, India

The high dimensional moving average process is considered and the asymptotics for eigenvalues of its sample autocovariance matrices are explored. Under quite weak conditions, we prove, in a unified way, that the limiting spectral distribution (LSD) of any symmetric polynomial in the sample autocovariance matrices, after suitable centering and scaling, exists and is nondegenerate. We use methods from free probability in conjunction with the method of moments to establish our results. In addition, we are able to provide a general description of the limits in terms of some freely independent variables. We also establish asymptotic normality results for the traces of these matrices. We suggest statistical uses of these results in problems such as order determination of high dimensional MA and AR processes and testing of hypotheses for coefficient matrices of such processes.

E1368: **Detection of intervention effects on time series**
Presenter: **Anish Ganguli**, Walmart, India

Intervention effects is a very popular term in the study of Time Series Analysis. There are a lot of factors which intervene with sales, and as a result, the pattern of the sales might get impacted a lot which eventually affects the forecasted sales numbers and, finally, the business decisions. The major problem with these intervention effects is to identify whether the effect is significant or not and also, if it is significant, how can we adjust for the impact or include its impact in the modelling/forecasting process? The objective is to review the approaches by which we can understand if the effects are significant or not in terms of impacting our Time Series sales data. To understand the significance of these effects, two approaches have been considered, viz. Autoregressive Distributed Lags (ARDL) and Google's Causal Impact (CI). These two approaches are helpful in determining if the intervention effect in consideration is significantly impacting our Time Series or not.

E1233: **Weighted l_1-penalized corrected quantile regression for high-dimensional temporally dependent measurement errors**
Presenter: **Nilanjan Chakraborty**, Washington University in Saint Louis, United States
Co-authors: Monika Bhattacharjee, Hira Lal Koul

The focus is on a high dimensional quantile regression model in presence of measurement errors. We consider as parse high-dimensional errors-in-variables linear regression model, where measurement errors in the covariates are assumed to have linear stationary temporal dependence and known Laplace marginal distributions. The regression errors are assumed to be independent-identically distributed random variables and have non-sub-Gaussian tails. Convergence results of the weighted l_1-penalized corrected quantile estimator of the regression parameter vector are established. An appropriate data-adaptive algorithm is given for obtaining a suitable choice of weights. Model consistency is established for the adaptive estimator. A simulation study has also been conducted to assess the finite sample performance of the proposed estimator.

EO633 Room Virtual R03 COUNTERFACTUAL ANALYSIS AND OPTIMAL POLICY Chair: Jiaying Gu

E1639: **Policy learning for optimal dynamic treatment regimes with observational data**
Presenter: **Shosei Sakaguchi**, University of Tokyo, Japan

Statistical decisions for dynamic treatment assignment problems are studied. Many policies involve dynamics in their treatment assignments where treatments are sequentially assigned to individuals across multiple stages, and the effect of treatment at each stage is usually heterogeneous with respect to the prior treatments, past outcomes, and observed covariates. We consider learning an optimal dynamic treatment regime that guides the optimal treatment assignment for each individual at each stage based on the individual's history. We propose two doubly-robust learning approaches using observational data under the assumption of sequential ignorability. The first approach solves the treatment assignment problem at each stage through backward induction, and the second approach solves the whole dynamic treatment assignment problem simultaneously across all stages. Using doubly-robust estimators of treatment effect scores and cross-fitting, each of the approaches can achieve the minimax optimal convergence rate $O_p(n^{-1/2})$ of welfare regret even when nuisance components are non-parametrically estimated.

E1658: **Optimal decision rules under partial identification**
Presenter: **Kohei Yata**, University of Wisconsin-Madison, United States

A class of statistical decision problems is considered in which the policymaker must decide between two alternative policies to maximize social welfare (e.g., the population mean of an outcome) based on a finite sample. The central assumption is that the underlying, possibly infinite-dimensional parameter, lies in a known convex set, potentially leading to partial identification of the welfare effect. An example of such restrictions

is the smoothness of counterfactual outcome functions. As the main theoretical result, we obtain a finite-sample decision rule (i.e., a function that maps data to a decision) that is optimal under the minimax regret criterion. This rule is easy to compute, yet achieves optimality among all decision rules; no ad hoc restrictions are imposed on the class of decision rules. We apply the results to the problem of whether to change a policy eligibility cutoff in a regression discontinuity setup. We illustrate the approach in an empirical application to the BRIGHT school construction program in Burkina Faso, where villages were selected to receive schools based on scores computed from their characteristics. Under reasonable restrictions on the smoothness of the counterfactual outcome function, the optimal decision rule implies that it is not cost-effective to expand the program.

E1686: **Counterfactual identification and latent space enumeration in discrete outcome models**
Presenter: **Thomas Russell**, Carleton University, Canada
Co-authors: Jiaying Gu, Thomas Stringham

A unified framework is provided for partial identification of counterfactual parameters in a general class of discrete outcome models allowing for endogenous regressors and multidimensional latent variables, all without parametric distributional assumptions. The main theoretical result is that, when the covariates are discrete, the infinite-dimensional latent variable distribution can be replaced with a finite-dimensional version that is equivalent from an identification perspective. The finite-dimensional latent variable distribution is constructed in practice by enumerating regions of the latent variable space with a new and efficient cell enumeration algorithm for hyperplane arrangements. We then show that bounds on a certain class of counterfactual parameters can be computed by solving a sequence of linear programming problems, and show how the researcher can introduce additional assumptions as constraints in the linear programs. Finally, we apply the method to a mobile phone choice example with heterogeneous choice sets, as well as an airline entry game example.

E1780: **Empirical welfare maximization with constraints**
Presenter: **Liyang Sun**, CEMFI, Spain

When designing eligibility criteria for welfare programs, policymakers naturally want to target the individuals who will benefit the most. Two new econometric approaches are proposed to selecting an optimal eligibility criterion when individuals' costs to the program are unknown and need to be estimated. One is designed to achieve the highest benefit possible while satisfying a budget constraint with high probability. The other is designed to optimally trade off the benefit and the cost from violating the budget constraint. The setting we consider extends the previous literature on Empirical Welfare Maximization by allowing for uncertainty in estimating the budget needed to implement the criterion, in addition to its benefit. Consequently, my approaches improve the existing approach as they can be applied to settings with imperfect takeup or varying program needs. We illustrate my approaches empirically by deriving an optimal budget-constrained Medicaid expansion in the US.

EO552 Room Virtual R04 NOVEL METHODS IN MICROBIOME DATA ANALYSIS Chair: Anna Plantinga

E0249: **SMRmix for integrating multiple microbiome datasets at community level**
Presenter: **Ni Zhao**, Johns Hopkins University, United States

Recent studies have highlighted the importance of human microbiota in our health and diseases. However, in many areas of research, individual microbiome studies often offer inconsistent results due to the limited sample sizes and the heterogeneity in study populations and experimental procedures. Integrative analysis of multiple microbiome datasets is necessary. However, statistical methods that incorporate multiple microbiome datasets and account for the study heterogeneity are not available in the literature. We develop a mixed-effect similarity matrix regression (SMRmix) approach for identifying community-level microbiome shifts between outcomes. SMRmix has a close connection with the microbiome kernel association test, one of the most popular approaches for such a task but is only applicable when we have a single study. Via extensive simulations, we show that SMRmix has well-controlled type I error and higher power than some potential competitors. We also applied SMRmix to data from the HIV-reanalysis consortium, a collective effort that obtained all publicly available data on the gut microbiome and HIV in December 2017, and obtained a coherent association of gut microbiome with HIV infection, and with MSM status (i.e. men who have sex with men).

E0561: **Covariate-adjusted Bayesian kernel regression for learning effect sizes of selected microbiome**
Presenter: **Liangliang Zhang**, Case Western Reserve University, United States
Co-authors: Christine Peterson

Current microbiome profiling methods allow for very fine resolution of the strains present in each sample. When associated with patient-level outcomes, the abundant features tend to be more influential than the rare features. Therefore, we propose a Bayesian kernel method to convolute all the microbial features together and study their joint impact in nonlinear kernel regression. The method can balance the impact of abundant features and rare features by taking into account their internally linked kinship structures and provide a similarity function which further helps in categorizing patient groups. Unlike the linear regression setting, there is no clear form of effect sizes in kernel regression. We transform the Kernel space back to the original space to obtain estimated effect sizes for individual microbiome features, which is usually not an easy task to accomplish. In addition, the model will provide us with improved uncertainty assessment both at the joint level and the individual level.

E0753: **Mixture margin random-effects copula models for inferring microbial co-variation networks**
Presenter: **Hongzhe Li**, University of Pennsylvania, United States

Longitudinal microbiome studies, in which data on a single subject are collected repeatedly over time, are becoming increasingly common in biomedical research. Such studies provide an opportunity to study the inherently dynamic nature of a microbiome in a way that cannot be done using cross-sectional studies. We develop random-effects copula models with mixed zero-beta margins to identify biologically meaningful temporally conserved co-variation between two bacterial taxa, while accounting for the excessive zeros seen in 16S rRNA and metagenomic sequencing data. The model assumes a random-effects model for the dependence parameter in the copulas, which captures the conserved microbial co-variation while allowing for a time-specific dependence parameters. Our analysis of the longitudinal pediatric DIABIMMUNE cohort identifies changes in both local and global patterns of microbial co-variation networks in infants treated with antibiotics. Our results show that the no-antibiotics network is less dependent on individual taxon, thus making it more stable than the antibiotics network and more robust to both targeted and random attacks.

E0933: **A new approach to testing mediation of the microbiome at both the community and individual taxon levels**
Presenter: **Yijuan Hu**, Emory University, United States
Co-authors: Ye Yue

Understanding whether and which microbes played a mediating role between exposure and a disease outcome is essential to develop clinical interventions to treat the disease by modulating the microbes. Existing methods for mediation analysis of the microbiome are often limited to a global test of community-level mediation or selection of mediating microbes without control of the false discovery rate (FDR). We propose a new approach based on inverse regression that regresses the microbiome data at each taxon on the exposure and the exposure-adjusted outcome. This approach fits nicely into our Linear Decomposition Model (LDM) framework, so our new method LDM-med, implemented in the LDM framework, enjoys all the features of the LDM, e.g., allowing an arbitrary number of taxa to be tested simultaneously, supporting continuous, discrete, or multivariate exposures and outcomes (including survival outcomes) as well as adjustment of confounders, and offering analysis of the taxon data at the relative abundance or presence-absence scale. Using extensive simulations, we showed that LDM-med always preserved the FDR of testing individual taxa and had adequate sensitivity. LDM-med always controlled the type I error of the global test and had compelling power over existing methods. The flexibility of LDM-med for a variety of mediation analyses is illustrated by an application to a murine microbiome dataset, which identified several plausible mediating taxa.

EO628 Room Virtual R05 RECENT DEVELOPMENTS IN SURVIVAL ANALYSIS Chair: Chi Hyun Lee

E0952: **Evaluation of the natural history of disease by combining incident and prevalent cohorts: Application to the Nun Study**
Presenter: **Daewoo Pak**, Yonsei University, Korea, South
Co-authors: Jing Ning, Richard Kryscio, Yu Shen

The Nun study is a well-known longitudinal epidemiology study of aging and dementia that recruited elderly nuns who were not yet diagnosed with dementia (i.e., incident cohort) and who had dementia prior to entry (i.e., prevalent cohort). In such a natural history of disease study, multistate modeling of the combined data from both incident and prevalent cohorts is desirable to improve the efficiency of inference. While important, the multistate modeling approaches for the combined data have been scarcely used in practice because prevalent samples do not provide the exact date of disease onset and do not represent the target population due to left-truncation. We demonstrate how to adequately combines both incident and prevalent cohorts to examine risk factors for every possible transition in studying the natural history of dementia. We adapt a four-state nonhomogeneous Markov model to characterize all transitions between different clinical stages, including plausible reversible transitions. The estimating procedure using the combined data leads to efficiency gains for every transition compared to those from the incident cohort data only.

E0960: **Regression analysis of multivariate recurrent event data allowing time-varying dependence**
Presenter: **Wen Li**, The University of Texas, United States
Co-authors: Mohammad Rahbar, Sean Savitz, Jing Zhang, Sori Lundin, Amirali Tahanan, Jing Ning

In multivariate recurrent event data, each patient may repeatedly experience more than one type of event. Analysis of such data gets further complicated by the time-varying dependence structure among different types of recurrent events. The available literature regarding the joint modeling of multivariate recurrent events assumes a constant dependency over time, which is strict and often violated in practice. To close the knowledge gap, we propose a class of flexible shared random effects models for multivariate recurrent event data, that allow for time-varying dependence to adequately capture complicated correlations among different types of recurrent events. We developed an expectation-maximization (EM) algorithm for stable and efficient model fitting. Extensive simulation studies demonstrated that the estimators of the proposed approach have satisfactory finite sample performance. We applied the proposed model and the estimating method to a cohort of stroke patients identified from the University of Texas Houston Stroke Registry and evaluated the effects of risk factors and the dependence structure of different types of post-stroke readmission events.

E1340: **A mixture model for estimating the risk of prostate cancer progression in active surveillance**
Presenter: **Yibai Zhao**, Fred Hutch Cancer Center, United States

Active surveillance has become a widely accepted management strategy to reduce overtreatment for low-risk prostate cancer patients. Prostate cancer is monitored through biopsies at scheduled visits to detect cancer progression to high risk, and the time to progression is left-censored. Because of biopsy misclassification, there are additional challenges to address. Individuals with high-grade cancer at the time of diagnosis (i.e., prevalent cases) may undergo active surveillance due to the imperfect sensitivity of biopsy, and some low-risk cancers may remain indolent indefinitely. In addition to the heterogeneity of cancers, observed data are subject to misclassification at each visit. We assume a mixture model for progressive and indolent cancers as well as the prevalent cases where the proportional hazards model incorporates the effect of either time-independent or time-varying covariates on cancer progression. We propose a semiparametric likelihood-based approach to handle interval-censored observations while accounting for the misclassification rates of biopsy. We conduct simulation studies to investigate the performance of the proposed approach under various settings. We apply the proposed approach to the Canary Prostate Active Surveillance Study to evaluate potential risk factors for cancer progression and to estimate the indolent fraction under a range of biopsy sensitivity rates.

E1400: **Estimating time-varying treatment effects on restricted mean survival time in large patient databases**
Presenter: **Chi Hyun Lee**, University of Massachusetts Amherst, United States
Co-authors: Wen Li, Yu Shen, Jing Ning

The restricted mean survival time (RMST), which is defined as the life expectancy up to a specific time point, has recently attracted substantial attention as an alternative to the hazard ratio for quantifying the treatment effect in clinical studies. We propose a flexible model to estimate the effect of treatment based on RMST. The effect of treatment is expressed as a function of restriction time to better characterize the dynamic trend of its effect on survival. To account for possible heterogeneity across patients in large databases, we incorporate the propensity scores for receiving treatment into the model. We further introduce an ensemble approach to aggregate estimators constructed based on subsamples of the observed failure times. We evaluate the finite sample performance of the proposed single model and ensemble-based approaches through simulations, and apply the proposed methods to the study of primary inflammatory breast cancer for assessing the effect of trimodality therapy on survival.

EO671 Room Virtual R06 ADVANCES IN GENERATIVE MODELLING Chair: Susan Wei

E1079: **Monotonicity and double descent in uncertainty estimation with Gaussian processes**
Presenter: **Liam Hodgkinson**, University of Melbourne, Australia
Co-authors: Chris van der Heide, Fred Roosta, Michael Mahoney

Contrary to what classical learning theory suggests, it is known that the quality of many modern machine learning models improves as the number of parameters increases. For predictive performance, these effects have recently been quantified with the double descent learning curve, which shows that larger models exhibit smaller test errors under appropriate regularization. We will present an analogous theory for models which estimate uncertainty, namely Gaussian processes (GP). In particular, contrary to popular belief, we will prove under a few assumptions that model quality of GPs under marginal likelihood improves monotonically in the number of covariates (even synthetic ones), provided an appropriate degree of regularization is imposed. To support our theory, we will show a variety of experiments, where we find this phenomenon holds beyond our considered assumptions and depends on several key factors, including kernel regularity and data conditioning.

E1389: **Training energy-based models with diffusion contrastive divergence**
Presenter: **Tianyang Hu**, Huawei Noah's Ark Lab, China

Diffusion probabilistic models have found great success in generative modeling. We show that the diffusion process can benefit other generative models as well, specifically energy-based models (EBMs). In training EBMs, Contrastive Divergence (CD) is a popular approach, which relies on drawing samples from a few MCMC steps starting from the data distribution to the EBM distribution. CD suffers when data is high-dimensional, and the mixing of MCMC can be problematic. To this end, we propose a novel training scheme termed Diffusion Contrastive Divergence (DCD). Different from CD, DCD gradually transfers both the data distribution and the EBM distribution with suitable diffusion processes, which can benefit from the close form conditional transition and save the computational cost on simulating stochastic differential equations. A novel method to characterize the evolving energy function along the transition process is developed. The proposed DCD is both analyzed theoretically and evaluated empirically, with comprehensive comparisons to existing training methods of EBMs.

E1464: **On the convergence of coordinate ascent variational inference**
Presenter: **Anirban Bhattacharya**, Texas AM University, United States
Co-authors: Debdeep Pati, Yun Yang

As a computational alternative to Markov chain Monte Carlo approaches, variational inference (VI) is becoming more and more popular for approximating intractable posterior distributions in large-scale Bayesian models due to its comparable efficacy and superior efficiency. Some

recent works provide theoretical justifications for VI by proving its statistical optimality for parameter estimation under various settings; meanwhile, formal analysis of the algorithmic convergence aspects of VI is still largely lacking. We consider the common coordinate ascent variational inference (CAVI) algorithm for implementing the mean-field (MF) VI of optimizing a Kullback-Leibler divergence objective functional over the space of all factorized distributions. Focusing on the two-block case, we analyze the convergence of CAVI by leveraging the extensive toolbox from functional analysis and optimization. We provide general conditions for certifying global or local exponential convergence of CAVI. As illustrations, we apply the developed theory to a number of examples, and derive explicit problem-dependent upper bounds on the algorithmic contraction rate.

E0318: **Black box variational inference with a deterministic objective: Faster, more accurate, and even more black box**
Presenter: **Tamara Broderick**, MIT, United States

Automatic differentiation variational inference (ADVI) offers fast and easy-to-use posterior approximation in multiple modern probabilistic programming languages. However, its stochastic optimizer lacks clear convergence criteria and requires tuning parameters. Moreover, ADVI inherits the poor uncertainty estimates of mean-field variational Bayes (MFVB). We introduce "deterministic ADVI" (D-ADVI) to solve these issues. In particular, we replace the intractable MFVB objective with a Monte-Carlo approximation; subsequently, fixing the Monte Carlo draws allows the use of off-the-shelf deterministic optimization tools. We show that D-ADVI reliably finds good solutions with default settings (unlike ADVI) and is faster and more accurate than ADVI. Moreover, unlike ADVI, D-ADVI is amenable to linear response corrections, yielding more accurate posterior covariance estimates. We demonstrate the benefits of D-ADVI on a variety of real-world problems.

EO538 Room Virtual R07 STATISTICAL ANALYSIS OF COMPLEX DEPENDENT DATA Chair: Weichi Wu

E0802: **Test for independence of infinite dimensional random elements**
Presenter: **Subhra Sankar Dhar**, IIT Kanpur, India
Co-authors: Suprio Bhar

A test for independence is studied for two random elements X and Y lying in an infinite dimensional space $\mathcal{H}$ (specifically, a real separable Hilbert space equipped with the inner product $\langle .,. \rangle_{\mathcal{H}}$). A measure of association is proposed based on the appropriate difference between the joint probability density function of the bivariate random vector $(\langle l_1, X\rangle_{\mathcal{H}}, \langle l_2, Y\rangle_{\mathcal{H}})$ and the product of marginal probability density functions of the random variables $\langle l_1, X\rangle_{\mathcal{H}}$ and $\langle l_2, Y\rangle_{\mathcal{H}}$, where $l_1 \in \mathcal{H}$ and $l_2 \in \mathcal{H}$ are two arbitrary elements. It is established that the proposed measure of association equals zero if and only if the random elements are independent. In order to carry out the test of whether X and Y are independent or not, the sample version of the proposed measure of association is considered as the test statistic, and the asymptotic distributions of the test statistic under the null and the local alternatives are derived. The performance of the new test is investigated for simulated data sets, and the practicability of the test is shown for one real data set related to climatology.

E0820: **Adaptive frequency band analysis for functional time series**
Presenter: **Pramita Bagchi**, George Mason University, United States

The frequency-domain properties of nonstationary functional time series often contain valuable information. These properties are characterized by their time-varying power spectrum. Practitioners seeking low-dimensional summary measures of the power spectrum often partition frequencies into bands and create collapsed measures of power within bands. However, standard frequency bands have primarily been developed through manual inspection of time series data and may not adequately summarize power spectra. We propose a framework for adaptive frequency band estimation of nonstationary functional time series that optimally summarizes the time-varying dynamics of the series. We develop a scan statistic and search algorithm to detect changes in the frequency domain. We establish the theoretical properties of this framework and develop a computationally-efficient implementation. The validity of our method is also justified through numerous simulation studies and an application to analyzing electroencephalogram data in participants alternating between eyes open and eyes closed conditions.

E0913: **Simultaneous inference for time series functional linear regression**
Presenter: **Zhou Zhou**, The Chinese University of Hong Kong, China
Co-authors: Yan Cui

The focus is on the problem of joint simultaneous confidence band (JSCB) construction for regression coefficient functions of time series scalar-on-function linear regression when the regression model is estimated by a roughness penalization approach with flexible choices of orthonormal basis functions. A simple and unified multiplier bootstrap methodology is proposed for the JSCB construction, which is shown to achieve the correct coverage probability asymptotically. Furthermore, the JSCB is asymptotically robust to inconsistently estimated standard deviations of the model. The proposed methodology is applied to a time series data set of the electricity market to visually investigate and formally test the overall regression relationship as well as perform model validation. A uniform Gaussian approximation and comparison result overall Euclidean convex sets for normalized sums of a class of moderately high-dimensional stationary time series is established.

E0941: **Time-varying correlation network analysis of non-stationary multivariate time series with complex trends**
Presenter: **Lujia Bai**, Tsinghua University, China
Co-authors: Weichi Wu

A unified approach is proposed for the inference of time-varying cross and autocorrelation curves of multivariate time series, which are observed once at a time and are non-stationary with piece-wise smooth trends. The dimension of the time series and the number of lags of cross and autocorrelation considered are allowed to diverge. The framework enables us to visualize the evidence of connections in the network based on various asymptotically correct multiple hypothesis testing of correlation functions, which is implemented via a difference-based and nonparametric estimator as well as bootstrap-assisted procedures for generating critical values. Therefore, the inferred network is able to capture functional relationships among second-order structures of time series. We prove the asymptotic validity of the correlation network inference procedure, and demonstrate its effectiveness in finite samples by simulation studies and empirical applications in finance.

EO494 Room Virtual R08 NOVEL STATISTICAL DEVELOPMENTS FOR ECONOMICS AND FINANCE Chair: Ramses Mena

E0670: **Stock market clustering methods**
Presenter: **Fidel Selva**, Universidad Nacional Autonoma de Mexico, Mexico

Financial markets have benefited greatly from classification systems like SIC and NAICS as a basis for dissecting the asset universe into comparable groups for analysis, index tracking and forecasting purposes. In the case of equity, the industry classifications in place are based on each company's portrayed business activity, earnings analysis and market perception. Those systems require human intervention, particularly from fundamental analysts, which derives in a resource-consuming process if one considers the data volume of the whole US Stock market. To avoid this manual classification, several distances, similarity and model-based clustering methods are explored for the CRSP database from 2000 to 2020 for the daily returns of the top 3000 stocks by liquidity. The goodness of fit of resulting groups is assessed internally with dispersion and likelihood measures, and externally by backtesting a dollar-neutral mean reversion strategy. For this last measure, the Sharp ratio was greater for some of the proposed methods.

E1066: **Kurtosis-based risk parity: Methodology, portfolio effects and properties**
Presenter: **Consuelo Nava**, University of Turin, Italy
Co-authors: Maria Grazia Zoia, Maria Debora Braga

A risk parity strategy is introduced based on portfolio kurtosis as a reference measure. This strategy allocates the asset weights in a portfolio in a manner that allows a homogeneous distribution of responsibility for portfolio returns' huge dispersion, since portfolio kurtosis puts more weight on extreme outcomes than standard deviation does. Therefore, the goal is not the minimization of kurtosis, but rather its fair diversification among assets. An original closed-form expression for portfolio kurtosis is devised to set up the optimization problem for this type of risk parity strategy. The latter is then compared with the one based on standard deviation by using data from a global equity investment universe and implementing an out-of-sample analysis. The kurtosis-based risk parity strategy has interesting portfolio effects, with lights and shadows. It outperforms the traditional risk parity according to main risk-adjusted performance measures. In terms of asset allocation solutions, it provides extremely unbalanced and more erratic portfolio weights (albeit without excluding any component) in comparison to those pertaining to the traditional risk parity strategy.

E0839: **Copula particle filters**
Presenter: **Carlos Erwin Rodriguez**, IIMAS-UNAM, Mexico
Co-authors: Stephen Walker

A novel analysis of the state space model is presented. It is shown that by modifying the standard recursive update, it is possible to apply a copula model to eliminate a particular integral, which is typically performed using importance sampling. With Bayesian models, copulas have recently been shown to provide predictive densities directly, avoiding integrals altogether. As in every particle filter algorithm, particles are generated; hence the proposed algorithm is named the Copula Particle Filter (CPF). As a by-product, the likelihood function of the model is obtained and used for parameter inference.

E1833: **On the mean of log-returns random distributions**
Presenter: **Ilia Naumkin**, IIMAS, Mexico
Co-authors: Ramses Mena

The study of the distribution of means of random probability measures has received considerable attention from the theoretical viewpoint during the last 30 years. However, the mathematical complexity inherent to such random objects has prevented them from reaching real applications successfully. We present a novel mechanism to efficiently compute the moments of such random means; specifically, we focus on Gibbs-type nonparametric priors. The motivation behind our proposal lies in the study of log-returns distributions of assets in financial markets, which we model via random distributions.

EO124 Room BH (SE) 1.02 APPLICATIONS OF DATA SCIENCE IN CAUSAL INFERENCE, IMAGING, AND FINANCE Chair: Stan Finkelstein

E0859: **Sensitivity analysis for violations of proximal identification assumptions**
Presenter: **Raluca-Ioana Cobzaru**, Massachusetts Institute of Technology, United States
Co-authors: Roy Welsch, Stan Finkelstein, Zach Shahn, Kenney Ng

Causal inference from observational data often rests on the unverifiable assumption of no unmeasured confounding. Recently, proximal inference has been introduced to leverage negative control outcomes and exposures as proxies to adjust for bias from unmeasured confounding. However, some of the key assumptions that proximal inference relies on are themselves empirically untestable. Additionally, the impact of violations of proximal inference assumptions on the bias of effect estimates is not well understood. We derive bias formulas for proximal inference estimators under a linear structural equation model data-generating process. These results are the first step toward sensitivity analysis and quantitative bias analysis of proximal inference estimators. While limited to a particular family of data-generating processes, our results offer some more general insight into the behavior of proximal inference estimators.

E1151: **What are the projections of donations and wealthier people: New forecasts for the tax incentives in the canton of Geneva**
Presenter: **Marta Pittavino**, University of Geneva, Switzerland
Co-authors: Giedre Lideikyte Huber

This is the second part of the first large-scale empirical legal analysis of tax incentives for charitable giving in Switzerland, and one of the few studies globally. Using unique longitudinal data, including household income and wealth of the entire taxpayers' population of the Canton of Geneva, Switzerland, we study patterns of charitable deductions and characteristics of donors making such deductions. Our study period extends over a decade (2001-2011). This period also encompasses a legal reform that raised ceilings for charitable deductions. We observe that an overwhelming majority of donors make deductions that never reach the legal ceiling, especially after the reform. Analyzing the deduction patterns in the entire donor population, we observe that deducting charitable donations has become increasingly popular during the study period. In addition, we find that donors' relative generosity tends to decrease when their income and wealth increase. Therefore, we investigated how wealthier people are affected by the tax incentives and how their related donations are conducted. Projections for the upcoming five years, from the end of our study period, are provided. An estimation of future donations and donors is given, based on the previously studied 11 years of data. Those results have important tax policy implications and relevance in modeling tax incentives for charitable giving, in both Switzerland and elsewhere.

E1420: **Portfolio construction using robust NLP incorporating noisy social media text**
Presenter: **Jennifer Zou**, Harvard University, United States
Co-authors: Roy Welsch, Frank Xing

Social media data provides valuable insight into retail investors' market perceptions in close to real-time; however, the signals can be noisy due to misspellings, abbreviations, and other representational differences. Furthermore, natural language processing (NLP) models for handling such texts have been shown to suffer from a number of robustness issues. We present a method for obtaining more robust semantic vector embeddings from social media (Twitter) data by training on a combination of clean and artificially generated noisy texts. We then demonstrate the improved performance of portfolios constructed using these robust estimates in simulation.

E1219: **Machine learning for Alzheimer's patients stratification and target identification**
Presenter: **Aamna AlShehhi**, Khalifa University, United Arab Emirates

Alzheimer's disease (A.D.) is an insidious, progressive, and degenerative neurodegenerative disease that destroys normal brain functionality. According to the World Health Organization (WHO), Alzheimer's disease is the most common form of dementia and contributes to approximately 70% of all dementia cases. In 2018 Alzheimer's Association reported that an estimated 5.7 million Americans were diagnosed with A.D. The number of patients is expected to double by 2050. A.D. is known to be caused by the presence and aggregates of tau neurofibrillary tangles and amyloid-beta (A) plaques in the brain. The heterogeneity of A.D. patients and the complexity of the disease genomic mechanisms create challenges for disease diagnosis, addressing patients' needs, and understanding treatment response. That is why machine learning and deep learning models can play a vital role in addressing those challenges. Our study aims to cluster A.D. patients into clinically homogeneous groups by linking Electronic Health Records (HER) patients' data with genomic information using different machine learning models and discovering a pivotal biomarker related to each specific stratum. A different disease pathway detected for different cohorts is confirmed and discovered.

EO050 Room BH (SE) 1.05 ROBUST CAUSAL INFERENCE IN BIOLOGY AND ECONOMICS Chair: Zijian Guo

E0536: **Causal inference with corrupted data: Measurement error, missing values, discretization, and differential privacy**
Presenter: **Rahul Singh**, MIT, United States

Co-authors: Anish Agarwal

The 2020 US Census will be published with differential privacy, implemented by injecting synthetic noise into the data. Controversy has ensued, with debates that center on the painful trade-off between the privacy of respondents and the precision of economic analysis. Is this trade-off inevitable? To answer this question, we formulate a semiparametric model of causal inference with high dimensional data that may be noisy, missing, discretized, or privatized. We propose a new end-to-end procedure for data cleaning, estimation, and inference with data cleaning-adjusted confidence intervals. We prove consistency, Gaussian approximation, and semiparametric efficiency by finite sample arguments. The rate of Gaussian approximation is $n^{-1/2}$ for semiparametric estimands such as average treatment effect, and it degrades gracefully for nonparametric estimands such as heterogeneous treatment effect. Our key assumption is that the true covariates are approximately low rank, which we interpret as approximate repeated measurements and validate in the Census. In our analysis, we provide nonasymptotic theoretical contributions to matrix completion, statistical learning, and semiparametric statistics. We verify the coverage of the data cleaning-adjusted confidence intervals in simulations. Finally, we conduct a semi-synthetic exercise calibrated to privacy levels mandated for the 2020 US Census.

E1027: **Paradoxes and resolutions for semiparametric data fusion of individual data and summary statistics**

Presenter: **Wang Miao**, Peking University, China

External summary statistics have been used as constraints on the internal data distribution, which promised to improve the statistical inference in the internal data; however, paradoxical results arise in such data integration: efficiency loss may occur if the uncertainty of the summary statistics is not negligible and estimation bias can emerge if they are obtained from a different population from the internal study. We investigate these paradoxical results in a semiparametric framework. We establish the semiparametric efficiency bound for estimating a general functional of the internal data distribution, which is shown to be no larger than that using only internal data. We propose a data-fused efficient estimator that achieves this bound so that the efficiency paradox is resolved. This initial data-fused estimator is further regularized with adaptive lasso penalty so that the resultant estimator can achieve the same asymptotic distribution as the oracle one that uses only unbiased summary statistics, which resolves the bias paradox. Simulations and applications to a Helicobacter pylori infection dataset are used to illustrate the proposed methods.

E1193: **Testing overidentifying restrictions with high-dimensional data and heteroskedasticity**

Presenter: **Ziwei Mei**, The Chinese University of Hong Kong, Hong Kong

Co-authors: Qingliang Fan, Zijian Guo

A new test is proposed for overidentifying restrictions (called the Q test) with high-dimensional data. This test is based on estimation and inference for quadratic forms of high-dimensional parameters. It is shown to have the desired asymptotic size and power properties under heteroskedasticity, even if the number of instruments and covariates is larger than the sample size. Simulation results show that the new test performs favorably compared to existing alternative tests under the scenarios when those tests are feasible or not. An empirical example of the trade and economic growth nexus manifests the usefulness of the proposed test.

E1201: **A novel penalized inverse-variance weighted estimator for Mendelian randomization with applications to COVID-19 outcomes**

Presenter: **Zhonghua Liu**, Columbia University, United States

Mendelian randomization utilizes genetic variants as instrumental variables (IVs) to estimate the causal effect of an exposure variable on an outcome of interest, even in the presence of unmeasured confounders. However, the popular inverse-variance weighted (IVW) estimator could be biased in the presence of weak IVs, a common challenge in MR studies. We develop a novel penalized inverse-variance weighted (pIVW) estimator, which adjusts the original IVW estimator to account for the weak IV issue by using a penalization approach to prevent the denominator of the pIVW estimator from being close to zero. Moreover, we adjust the variance estimation of the pIVW estimator to account for the presence of balanced horizontal pleiotropy. We show that the recently proposed debiased IVW (dIVW) estimator is a special case of our proposed pIVW estimator. We further prove that the pIVW estimator has smaller bias and variance than the dIVW estimator under some regularity conditions. We also conduct extensive simulation studies to demonstrate the performance of the proposed pIVW estimator. Furthermore, we apply the pIVW estimator to estimate the causal effects of five obesity-related exposures on three coronavirus disease 2019 (COVID-19) outcomes.

EO668 Room BH (S) 2.02 STRUCTURED PRIOR DISTRIBUTIONS FOR COMPLEX MODELS Chair: Sarah Elizabeth Heaps

E1782: **Challenges and successes with structured prior modelling in survey adjustment**

Presenter: **Lauren Kennedy**, Monash University, Australia

Multilevel regression and poststratification have grown in popularity as a method to adjust for non-response and non-probability samples in surveys. One of the hallmarks is the use of regularization or partial pooling to ensure efficient predictions for rarer groups in the sample (either through sampling or through underlying population demographics). We will discuss how structured priors help to achieve this regularization, when they are useful to consider, and some of the complexities of identifying a "good" prior to use.

E1074: **Bayesian cumulative shrinkage for infinite factorizations**

Presenter: **Sirio Legramanti**, University of Bergamo, Italy

Co-authors: Daniele Durante, David Dunson

The dimension of the parameter space is typically unknown in a variety of models that rely on factorizations. For example, in factor analysis, the number of latent factors is not known and has to be inferred from the data. Although classical shrinkage priors are useful in such contexts, increasing shrinkage priors can provide a more effective approach that progressively penalizes expansions with growing complexity. We propose a novel increasing shrinkage prior, called the cumulative shrinkage process, for the parameters that control the dimension in overcomplete formulations. Our construction has broad applicability and is based on an interpretable sequence of spike-and-slab distributions which assign increasing mass to the spike as the model complexity grows. Using factor analysis as an illustrative example, we show that this formulation has theoretical and practical advantages relative to current competitors, including an improved ability to recover the model dimension. An adaptive Markov chain Monte Carlo algorithm is proposed, and the performance gains are outlined in simulations and in an application to personality data.

E0713: **Bayesian estimation of correlation matrices of longitudinal data**

Presenter: **Riddhi Pratim Ghosh**, Bowling Green State University, United States

Co-authors: Bani Mallick, Mohsen Pourahmadi

Estimation of correlation matrices is a challenging problem due to the notorious positive-definiteness constraint and high dimensionality. Reparameterising Cholesky factors of correlation matrices in terms of angles or hyperspherical coordinates where the angles vary freely in the range $[0,\pi)$ has become popular in the last two decades. However, it has not been used in Bayesian estimation of correlation matrices, perhaps due to a lack of clear statistical relevance and suitable priors for the angles. We show for the first time that for longitudinal data, these angles are the inverse cosine of the semi-partial correlations (SPCs). This simple connection makes it possible to introduce physically meaningful selection and shrinkage priors on the angles or correlation matrices with emphasis on selection (sparsity) and shrinking towards the longitudinal structure. Our method deals effectively with the positive-definiteness constraint in posterior computation. We compare the performance of our bayesian estimation based on angles with some recent methods based on partial autocorrelations through simulation and apply the method to data related to a clinical trial on smoking.

E0384: **Structured prior distributions for the covariance matrix in latent factor models**

Presenter: **Sarah Elizabeth Heaps**, Durham University, United Kingdom

Factor models are widely used for dimension reduction in the analysis of multivariate data. This is achieved through a (sparse) factorisation of a $p \times p$ covariance matrix; a latent factor representation allows this to be interpreted as the sum of a diagonal matrix of idiosyncratic variances and a shared variation matrix equal to the product of a $p \times k$ factor loadings matrix ($k << p$) and its transpose. Historically, little attention has been paid to incorporating prior information in Bayesian analyses using factor models where, at best, the prior for the factor loadings matrix is invariant with respect to the order of the variables. A class of structured priors is developed that can encode ideas of dependence structure about the shared variation matrix. The construction allows data-informed shrinkage towards sensible parametric structures within a framework that facilitates inference on the number of factors. Using an unconstrained reparameterisation of stationary vector autoregressions, the methodology is also extended to stationary dynamic factor models. For computational inference, parameter-expanded Markov chain Monte Carlo samplers are proposed, including an efficient adaptive Gibbs sampler. Finally, substantive applications showcase the scope of the methodology and its inferential benefits.

EO743 Room BH (S) 2.03 BAYESIAN AND ROBUST INSIGHTS IN DATA ANALYSIS AND CLASSIFICATION Chair: Miguel de Carvalho

E1253: **Bayesian nonparametric inference for the overlap coefficient: Application to disease diagnosis**
Presenter: **Vanda Inacio**, FCiencias.ID - Associacao para a Investigacao e Desenvolvimento de Ciencias, Portugal

Diagnostic tests play an important role in medical research and clinical practice. The ultimate goal of a diagnostic test is to distinguish between diseased and nondiseased individuals and before a test is routinely used in practice, it is a pivotal requirement that its ability to discriminate between these two states is thoroughly assessed. The overlap coefficient, which is defined as the proportion of overlap area between two probability density functions, has gained popularity as a summary measure of diagnostic accuracy. We propose two Bayesian nonparametric estimators, based on Dirichlet process mixtures, for estimating the overlap coefficient. We further introduce the covariate-specific overlap coefficient and develop a Bayesian nonparametric approach based on Dirichlet process mixtures of additive normal models for estimating it. A simulation study is conducted to assess the empirical performance of our proposed estimators. Two illustrations are provided: one concerned with the search for biomarkers of ovarian cancer and another one aimed to assess the age-specific accuracy of glucose as a biomarker of diabetes.

E1309: **Divide and conquer: Cluster analysis with a different number of clusters**
Presenter: **Andrej Svetlosak**, University of Edinburgh, United Kingdom
Co-authors: Raffaella Calabrese

Joint clustering modelling can be challenging. Results of similarity-based methods (via k-means and k-medoids) often do not reflect data structures, while model-based clustering (via mixture models), as we show, nearly always leads to the same number of components on each margin. We address these drawbacks by proposing a novel approach – which we refer to as divide and conquer clustering – that lies on the interface between model-based clustering and similarity-based clustering. Our approach consists of three steps, provides interpretable cluster solutions, and allows for a differing number of components on the margins. We achieve this by first modelling each margin separately by recently introduced non-local prior mixtures, which treat the number of components as a model parameter. Second, we learn about the set of possible joint clusters (proto-clusters) obtained via Voronoi tessellation on the product space of the marginal component means. Lastly, the final joint clusters are the Voronoi faces centred at the local density maxima of the joint distribution. These are obtained by dividing up proto-clusters with a density below a threshold between the remaining Voronoi faces. In this sense, the high-density areas divide and conquer low-density regions. We analyse and compare the performance of the method with selected state-of-the-art clustering methods. The results on both simulated and real datasets suggest an on-par or better performance than competing methods.

E1142: **Modeling residuals with mixtures of Gaussian with non-Gaussian components for robustness and outlier detection**
Presenter: **Alexandra Posekany**, University of Technology Vienna, Austria

Outliers and systematically skewed or heavy-tailed data frequently occur in data analytical problems of many fields ranging from economics to bioinformatics. A specific notion of Bayesian robustness is robustifying the likelihood, as the backbone of the model. Constructing normally distributed likelihood models is often due to computational convenience, in the same way as classical inference with approximate normality is. Independent of sample size, data in many applied fields nowadays do not fulfil this assumption, and linear and non-linear models for regression or classification suffer from that. We aim to provide a robust estimation of parameters of the "main part of the data" through a normal or skewed distribution as likelihood, while simultaneously identifying the "outlying part of the data" represented by one or more skewed or heavy-tailed mixture components. Through the component labels and posterior weights we can identify the noisy or outlying parts of the data for filtering or inspecting the data quality.

E1619: **Robust approximate Bayesian inference**
Presenter: **Nicola Sartori**, University of Padova, Italy
Co-authors: Erlis Ruli, Laura Ventura

A method is illustrated that allows the construction of pseudo posterior distributions based on unbiased estimating functions. In particular, such estimating functions are used to construct suitable summary statistics in Approximate Bayesian Computation algorithms. The composite score function is a prominent example of estimating function that can be used in complex models when the likelihood computation is too demanding or when a full model specification could be too strong. The latter case implies weaker model assumptions than a full Bayesian analysis and therefore can lead to a more robust inference. In order to directly address the robustness of the posterior distribution, we propose the use of M-estimating functions instead. The theoretical properties of the corresponding robust posterior distributions are discussed.

EO625 Room BH (S) 2.05 COMPUTATIONS AND METHODS IN BAYESIAN NONPARAMETRICS Chair: Mario Beraha

E0665: **Mixture representations for likelihood ratio ordered distributions**
Presenter: **Andres Barrientos**, Florida State University, United States
Co-authors: Michael Jauch, Victor Pena, David Matteson

Mixture representations for likelihood ratio ordered distributions are introduced. Essentially, the ratio of two probability densities, or mass functions, is monotone if and only if one can be expressed as a mixture of one-sided truncations of the other. To illustrate the practical value of the mixture representations, we address the problem of density estimation for likelihood ratio ordered distributions. In particular, we propose a nonparametric Bayesian solution which takes advantage of the mixture representations. The prior distribution is constructed from Dirichlet process mixtures and has large support on the space of pairs of densities satisfying the monotone ratio constraint. With a simple modification to the prior distribution, we can test the equality of two distributions against the alternative of likelihood ratio ordering. We develop a Markov chain Monte Carlo algorithm for posterior inference and demonstrate the method in a biomedical application.

E0677: **Inferring taxonomic placement from DNA barcoding aiding in discovery of new taxa**
Presenter: **Alessandro Zito**, Duke University, United States
Co-authors: Tommaso Rigon, David Dunson

Predicting the taxonomic affiliation of DNA sequences collected from biological samples is a fundamental step in biodiversity assessment. This task is performed by leveraging on existing databases containing reference DNA sequences whose taxa are known. However, environmental sequences can be from organisms that are either unknown to science or for which there are no reference sequences available. Thus, the taxonomic novelty

of a sequence needs to be accounted for when doing classification. We propose Bayesian nonparametric taxonomic classifiers, BayesANT, which use species sampling model priors to allow new taxa to be discovered at each taxonomic rank. Using a simple product multinomial likelihood with conjugate Dirichlet priors at the lowest rank, a highly flexible algorithm is developed to provide a probabilistic prediction of the taxa placement of each sequence at each rank. As an illustration, we run our algorithm on a carefully annotated library of Finnish arthropods. To assess the ability of BayesANT to recognize novelty and to correctly predict known taxonomic affiliations, we test it on two training-test splitting Scenarios, each with a different proportion of taxa unobserved in training. We show how our algorithm attains excellent prediction performances and reliably quantifies classification uncertainty, especially when many sequences in the test set are affiliated with taxa unknown in training.

E0889: **Bayesian clustering of high-dimensional data via latent repulsive mixtures**
Presenter: **Lorenzo Ghilotti**, University of Milano-Bicocca, Italy
Co-authors: Mario Beraha, Alessandra Guglielmi

In modern applications, it is common to collect high-dimensional data and be interested in clustering subjects based on them. It has been shown that mixture models produce inconsistent inference in that setting, proposing a general class of models overcoming such an issue, called Lamb. Their approach consists in linking observations to a set of low-dimensional latent factors through a matrix of loadings, and performing model-based clustering via nonparametric mixture models on the latent space. Lamb models are likely to be misspecified, thus leading to inconsistent clustering. Repulsive mixture models have recently provided empirical evidence about robustness to misspecification, limitedly to low-dimensional data. We propose, within the class of Lamb models, to employ a repulsive mixture model to cluster the latent factors. To this end, we propose a general construction for anisotropic determinantal point processes (DPPs), which guarantees the analytical availability of their spectral densities. We employ such a DPP as prior in a repulsive mixture model on the latent factors, and let the matrix of factor loadings drive the anisotropic behavior, so that separation is indeed induced between the high-dimensional centers of different clusters. An efficient MCMC algorithm is proposed, and the methodology is compared to existing methods.

E1203: **Bayesian nonparametric multilayer clustering of longitudinal data**
Presenter: **Beatrice Franzolini**, Agency for Science, Technology and Research, Singapore, Singapore
Co-authors: Maria De Iorio

A new class of Bayesian nonparametric models is introduced to make inferences on an ordered collection of partitions of the same objects, namely, a multilayer partition. The class is suited to analyze panel/longitudinal data, where repeated observations are collected over time for the same observational units. The core is a conditional partial exchangeable structure, which we argue is a natural and general modeling strategy for this context. The resulting class of Bayesian models guarantees analytical and computation tractability both in terms of the clustering structure and the underlying random probabilities measures. It allows predictions for any new number of observations and may, ultimately, constitute a powerful reference framework -currently missing in the literature- for the development of tailored Bayesian nonparametric models for panel data. We further explore in detail two specific models within this class: one based on a novel prior and another employing the well-known hierarchical Dirichlet process as a building block.

EO068 Room K2.31 (Nash Lec. Theatre) INNOVATIVE AND PRACTICAL STRATEGIES IN CAUSAL INFERENCE Chair: Andrew Spieker

E0235: **Combining experimental and non-experimental data to examine treatment effect heterogeneity**
Presenter: **Carly Lupton-Smith**, Johns Hopkins Bloomberg School of Public Health, United States
Co-authors: Elizabeth Stuart

Determining what works for whom is a key goal in prevention and treatment across a variety of areas, including mental health. Identifying effect moderator factors that relate to the size of treatment effects is crucial for the delivery of treatment and prevention interventions, but doing so is incredibly difficult using standard study designs. Randomized trials, the gold standard for estimating average effects, are typically underpowered to detect moderation. Large-scale nonexperimental studies may provide another way to examine the effect of moderation, but can suffer from confounding. Recent machine learning and Bayesian methods advances are described to combine randomized trials and electronic health record (EHR) data to examine effect heterogeneity. We present results from simulation studies comparing a set of recently proposed methods for combining data sources, with the goal of estimating conditional average treatment effects. We also provide an initial application of the methods to data from randomized trials and electronic health record data of individuals receiving medication treatment for major depressive disorder.

E0698: **thinkCausal: A tool to help researchers learn while they do**
Presenter: **Jennifer Hill**, New York University, United States
Co-authors: George Perrett

Causal inference is a necessary tool in education research for answering pressing and ever-evolving questions around policy and practice. Increasingly researchers are using more complicated machine learning algorithms to estimate causal effects. These methods take some of the guesswork out of analyses, decrease the opportunity for p-hacking, and are often better suited for more fine-tuned causal inference tasks such as identifying varying treatment effects and generalizing results from one population to another. However, these more sophisticated methods are more difficult to understand and are often only accessible in more technical, less user-friendly software packages. The thinkCausal project is working to address these challenges by developing a highly-scaffolded, multi-purpose causal inference software package in R Shiny with the BART predictive algorithm as a foundation. The software will scaffold the researcher through the data analytic process and provide options to access technology-based teaching tools to understand foundational concepts in causal inference and machine learning. What we have accomplished will be outlined, and the challenges and opportunities in building this type of tool will be discussed.

E0981: **An influence function based instrumental variable estimator of censored medical costs**
Presenter: **Nicholas Illenberger**, NYU Langone Health, United States
Co-authors: Nandita Mitra

Studies aimed at estimating medical costs accrued under different treatments are critical to making informed healthcare policy decisions. Because cost analyses often use data from observational sources, their results may be biased due to unmeasured confounding or informative cost censoring. We introduce a partitioned, instrumental variable estimator of the complier average treatment effect on costs. Given a valid instrument, our estimator provides unbiased estimates of the target treatment effect in the presence of unmeasured confounding. Additionally, the use of a partitioned cost estimator allows us to address informative cost censoring and improve efficiency by utilizing data from patients with partially observed medical costs. Our proposed estimator is based on influence functions, allowing for multiple robust, efficient, and flexible semiparametric estimation. We present results from simulation studies to assess the performance of our proposed estimator under varying degrees of censoring and strength of IV. We apply our approach to a study assessing the costs of surgical and non-surgical interventions for gallstones and hemorrhaging using observational data.

E2000: **Robust covariate-assisted bounds in instrumental variable designs**
Presenter: **Alexander Levis**, Carnegie Mellon University, United States
Co-authors: Matteo Bonvini, Zhenghao Zeng, Luke Keele, Edward Kennedy

When exposure of interest is confounded by unmeasured factors, an instrumental variable (IV) can be used to identify and estimate certain causal contrasts. Identification of the marginal average treatment effect (ATE) from IVs typically relies on strong untestable structural assumptions.

When one is unwilling to assert such structural assumptions, IVs can nonetheless be used to construct bounds on the ATE. Famously, linear programming techniques were employed to prove tight bounds on the ATE for a binary outcome, in a randomized trial with noncompliance and no covariate information. We demonstrate how these bounds remain useful in observational settings with baseline confounders of the IV, as well as randomized trials with measured baseline covariates. The resulting lower and upper bounds on the ATE are non-smooth functionals, and thus standard nonparametric efficiency theory is not immediately applicable. To remedy this, we introduce a novel margin condition, and propose an influence function-based estimator of the ATE bounds for a binary outcome that can attain parametric convergence rates when nuisance functions are modeled flexibly. We demonstrate the properties of this estimator in simulation studies and real data. Finally, we discuss various relevant design issues, and propose extensions to continuous outcomes.

EO703 Room K2.40 STATISTICAL METHODS IN BRAIN IMAGING Chair: Hernando Ombao

E0964: **Bayesian multi-object data integration in the study of primary progressive aphasia**
Presenter: **Aaron Scheffler**, University of California, San Francisco, United States
Co-authors: Rajarshi Guhaniyogi, Rene Gutierrez

Clinical researchers collect multiple images from separate modalities (sources) to investigate questions of human health that are inadequately explained by considering one image source at a time. Viewing the collection of images as multi-objects, the successful integration of multi-object data produces a sum of information greater than the individual parts. Still, this integration can be hindered by data complexity. Each image contains structural information, indexing spatial information, or network information, indexing connectivity among the image, which reinforce each other but is challenging to merge. We propose a Bayesian regression framework that provides inference and prediction for a multi-object outcome as a function of a scalar predictor. Our framework will accommodate multiple image outcomes having different structures and identify image regions associated with the scalar predictor jointly via efficient hierarchical prior structures that scale to high-resolution image data volume. A working example is provided for the association of language comprehension scores with multi-object image data to explore the neural underpinnings of language loss in primary progressive aphasia patients.

E1415: **Supervised modeling of multiple networks for multimodal neuroimaging data**
Presenter: **Sharmistha Guha**, Texas A&M University, United States

Novel Bayesian methodologies are developed to combine information across multiple data sources to better characterize complex physical and biological systems that are inadequately explained by considering one data source at a time. The motivation comes from ever-growing brain-imaging data from multiple modalities or sources which can together offer an in-depth understanding of the human brain in health and disease.

E1440: **A functional model for studying common trends across trial time in eye-tracking experiments**
Presenter: **Damla Senturk**, University of California Los Angeles, United States
Co-authors: Donatello Telesca, Catherine Sugar, Mingfei Dong

Eye-tracking (ET) experiments commonly record the continuous trajectory of a subject's gaze on a two-dimensional screen throughout repeated presentations of stimuli (referred to as trials). Even though the continuous path of gaze is recorded during each trial, commonly derived outcomes for analysis collapse the data into simple summaries, such as looking times in regions of interest, latency to looking at stimuli, number of stimuli viewed, number of fixations or fixation length. In order to retain information in trial time, we utilize functional data analysis (FDA) for the first time in literature in the analysis of ET data. More specifically, novel functional outcomes for ET data, referred to as viewing profiles, are introduced that capture the common gazing trends across trial time which are lost in traditional data summaries. The mean and variation of the proposed functional outcomes across subjects are then modeled using functional principal components analysis. Applications to data from a visual exploration paradigm conducted by the Autism Biomarkers Consortium for Clinical Trials showcase the novel insights gained from the proposed FDA approach, including significant group differences between children diagnosed with autism and their typically developing peers in their consistency of looking at faces early on in trial time.

E1885: **Hidden Markov and semi-Markov Models for dynamic connectivity analysis in resting-state fMRI**
Presenter: **Mark Fiecas**, University of Minnesota, United States

Motivated by a study on adolescent mental health, a dynamic connectivity analysis is conducted using resting-state functional magnetic resonance imaging (fMRI) data. A dynamic connectivity analysis investigates how the interactions between different regions of the brain, represented by the different dimensions of a multivariate time series, change over time. Hidden Markov models (HMMs) and hidden semi-Markov models (HSMMs) are common analytic approaches for conducting dynamic connectivity analyses. We will give an overview of HMMs and HSMMs and their utility of dynamic connectivity analysis. We will describe how we can assess model fit using pseudo-residuals. We use these models to conduct a dynamic connectivity analysis on fMRI data obtained from female adolescents, where we show how dwell-time distributions vary across the severity of non-suicidal self-injury (NSSI). We will provide empirical evidence of the limitations of HMMs and HSMMs with respect to model fit, and discuss potential steps for further development of these models for dynamic connectivity analysis.

EO176 Room K2.41 RANDOM MATRIX THEORY AND ITS APPLICATIONS Chair: Yanrong Yang

E0762: **Optimal and adaptive shrinkage estimators for general large covariance and precision matrices**
Presenter: **Xiucai Ding**, UC Davis, United States

Some recent results are shown on the estimation of high dimensional covariance and precision matrices using Stein's invariant (shrinkage) estimators under various loss functions. We provide the first general analytical formulas for Stein's estimators for various loss functions for both the spiked and non-spiked models. Based on our formulas, we also propose optimal and adaptive estimators for these shrinkers. An algorithm and R package are provided to conduct the calculations. In order to study the asymptotics of our estimators, we establish the asymptotic normality for all the non-outlier eigenvectors and their associated quantum unique ergodicity (QUE) for a potentially spiked model.

E0877: **On eigenvalues of a high-dimensional Kendalls rank correlation matrix with dependence**
Presenter: **Zeng Li**, Southern University of Science and Technology, China
Co-authors: Cheng Wang, Qinwen Wang

Limiting spectral distribution of a high dimensional Kendall's rank correlation matrix is investigated. The underlying population is allowed to have a general dependence structure. The result no longer follows the generalized Marchenko-Pastur law, which is brand new. It is the first result on rank correlation matrices with dependence. As applications, we study Kendall's rank correlation matrix for multivariate normal distributions with a general covariance matrix. From these results, we further gain insights into Kendall's rank correlation matrix and its connections with the sample covariance/correlation matrix.

E1055: **Spiked eigenvalues of high-dimensional sample autocovariance matrices: CLT and applications**
Presenter: **Yanrong Yang**, The Australian National University, Australia
Co-authors: Han Xiao Han

High-dimensional autocovariance matrices play an important role in dimension reduction for high-dimensional time series. We establish the central limit theorem (CLT) for spiked eigenvalues of high-dimensional sample autocovariance matrices, which is developed under general conditions.

The spiked eigenvalues are allowed to go to infinity in a flexible way without restrictions in divergence order. Moreover, the number of spiked eigenvalues and the time lag of the autocovariance matrix under study could be either fixed or tending to infinity when the dimension p and the time length T go to infinity together. As a further statistical application, a novel autocovariance test is proposed to detect the equivalence of spiked eigenvalues for two high-dimensional time series. Various simulation studies are illustrated to justify the theoretical findings. Furthermore, a hierarchical clustering approach based on the autocovariance test is constructed and applied to clustering mortality data from multiple countries.

E1124: **Optimal network community inference under severe degree heterogeneity**
Presenter: **Jingming Wang**, Harvard University, United States
Co-authors: Tracy Ke

The estimation of mixed memberships is a problem of great interest in the analysis of large social network data. A symmetric network with n nodes and K communities is considered, where each node i has a mixed membership vector π_i (a K-dimensional probability mass function). It is of great interest to reveal the optimal error rates for estimating the membership vectors. The degree corrected mixed membership (DCMM) model is adopted, and two loss metrics are considered, an unweighted l^1-loss and a degree-weighted l^1-loss. The minimax rates for both loss metrics are obtained under a very broad setting that allows for a wide range of network sparsity, severe degree heterogeneity, weak signals, and diverging K. A spectral algorithm is also proposed, which achieves the minimax rates for both metrics, up to some logarithmic factors. Technically, a lot of effort is required in the study of a sharp entry-wise large deviation bound for the leading eigenvectors of the regularized graph Laplacian. This result plays a key role in motivating and understanding the rate-optimal spectral algorithm.

EC824 Room K0.18 STATISTICS FOR IMAGES AND BRAIN SIGNALS Chair: Brenda Betancourt

E1580: **Cross-scale dependence based on multi-resolution analysis of multivariate time series with application to EEG data**
Presenter: **Haibo Wu**, King Abdullah University of Science and Technology, Saudi Arabia
Co-authors: Marina Knight, Hernando Ombao

The goal is to develop a novel statistical approach to characterizing functional interactions between channels in a brain network. Wavelets are effective for capturing transient properties of non-stationary signals. Wavelets give a multi-scale decomposition of signals and, thus, can be few for studying potential cross-scale interactions between signals. We develop scale-specific sub-processes of a multivariate locally stationary wavelet stochastic process. Under this proposed framework, a novel cross-scale dependence measure and its estimation are developed, and it provides a measurement for the dependence structure of components at different scales of multivariate time series. Extensive simulation studies are conducted to demonstrate the theoretical properties of the model hold true in practice. The proposed cross-scale analysis is applied to the electroencephalogram (EEG) data to study alterations in the functional connectivity structure in children diagnosed with attention deficit hyperactivity disorder (ADHD). Our approach identified some interesting cross-scale interactions between channels in the brain network. The proposed framework can be applied to other signals.

E1871: **Tail transfer entropy: A new extremal dependence measure for studying connectivity in a brain network**
Presenter: **Paolo Victor Redondo**, King Abdullah University of Science and Technology, Saudi Arabia
Co-authors: Raphael Huser, Hernando Ombao

Brain signals, such as electroencephalograms (EEG), record neuronal activity in the cortex. During the execution of a cognitive function, large amplitude signals are associated with high activity, i.e., indication of functioning brain regions. The interest now is to infer on the impact of these amplitudes from a brain region to other regions in the context of extremes. We develop a new measure called tail transfer entropy (TTE) to quantify the amount of information transferred from the tail distribution of one signal to another signal's tails. As a result, an extremal brain connectivity network may be constructed. To estimate TTE, we propose a copula-based approach through the vine copula structure embedded with extreme value theory. Lastly, we illustrate our proposed measure based on some numerical experiments and provide interesting and novel findings on the analysis of EEG recordings linked to a visual task.

E1944: **Unbiased and robust analysis of co-localization in super-resolution images**
Presenter: **Hui Zhang**, Northwestern University, United States

Spatial data from high-resolution images abound in many scientific disciplines. For example, single-molecule localization microscopy, such as stochastic optical reconstruction microscopy, provides super-resolution images to help scientists investigate the co-localization of proteins and hence their interactions inside cells, which are key events in living cells. However, there are few accurate methods for analyzing co-localization in super-resolution images. The current methods and software are prone to produce false-positive errors and are restricted to only 2-dimensional images. We will propose a novel statistical method to effectively address the problems of unbiased and robust quantification and comparison of protein co-localization for multiple 2- and 3-dimensional image datasets. This method significantly improves the analysis of protein co-localization using super-resolution image data, as shown by its excellent performance in simulation studies and an analysis of light chain 3-lysosomal-associated membrane protein 1 protein co-localization in cell autophagy. Moreover, this method is directly applicable to co-localization analyses in other disciplines, such as diagnostic imaging, epidemiology, environmental science, and ecology.

E1485: **Decomposition of multivariate brain signals in multi-subject replicated setting**
Presenter: **Guillermo Cuauhtemoctzin Granados Garcia**, King Abdullah Universe of Science and Technology, Saudi Arabia
Co-authors: Raquel Prado, Hernando Ombao

A Bayesian nonparametric model is proposed to understand (change to investigate) the dynamic dependence between multivariate brain signals in experiments involving several trials and subjects. The proposed method models a multiple-channel signal as a mixture of latent second-order autoregressive processes (AR(2)). Each latent AR(2) represents a unique quasi-periodic wave shared across channels. The channel dependence structure is inferred via a Hierarchical Dirichlet process model allowing us to model the strength of each latent wave by borrowing information across trials and subjects. A Metropolis within Gibbs algorithm is implemented for posterior distribution computation, mixed with optimization strategies to improve computational performance. The model's effectiveness is demonstrated in a simulation study of two groups with smooth and abrupt changes in the channel dependency structure and signals oscillatory behavior. Lastly, the novel method is used to compare the EEG recordings of a group of alcoholics and a control group during a visual recognition experiment.

EC819 Room S0.11 STATISTICS AND ECONOMETRICS MODELLING AND APPLICATIONS Chair: Philipp Otto

E1795: **Bayesian spatio-temporal methods: an application to forecasting short-term defaults of firms in a commercial network**
Presenter: **Silvia Montagna**, University of Turin, Italy
Co-authors: Raffaele Argiento, Claudia Berloco

To protect financial institutions from unexpected credit losses, it is of primary importance to foresee any evidence of contagion of liquidity distress across a network of firms. This term indicates a situation of lack of solvency of a firm (e.g., a customer) that propagates to other firms (e.g., its suppliers). We look for evidence of contagion of liquidity distress on an Intesa Sanpaolo proprietary dataset by means of Bayesian spatial and spatio-temporal models. The results indicate that such models can detect cases of distress not yet apparent from covariate information collected on the firms by instead borrowing information from the network, leading to improved forecasting performance on the prediction of short-term default with respect to state-of-the-art methods.

E0915: **Amount-dependent fraud detection considering aggregated losses**
Presenter: **Jorge C-Rella**, Abanca servicios financieros, Spain
Co-authors: Ricardo Cao, Juan Vilar Fernandez
Fraud detection is a significantly difficult problem due to an intrinsic extreme unbalance and class overlap. In classical approaches, the likelihood of belonging to the positive class is estimated and a threshold is selected in order to classify an observation. As this approach does not consider the amount, it produces suboptimal decision rules in cost-sensitive classification problems. A new approach is proposed constructing a two-dimensional decision space considering the two variables on which losses depend, namely the estimated fraud probability and the loan amount. This expansion allows more freedom to the decision region, from which a new proposed algorithm takes advantage in order to obtain the optimal decision rule in terms of aggregated losses on the sample. Classical classification rules are contained in the algorithm search, so an improvement is consistently expected, which is shown with a practical application and a series of simulations.

E1799: **The smoots Package in R for semiparametric modeling of trend stationary time series**
Presenter: **Dominik Schulz**, Paderborn University, Germany
Co-authors: Yuanhua Feng, Thomas Gries, Sebastian Letmathe
An introduction is given to the new package in R called "smoots" (smoothing time series), developed for data-driven local polynomial smoothing of trend-stationary time series. Functions for data-driven estimation of the first and second derivatives of the trend are also built-in. It is first applied to monthly changes in the global temperature. The quarterly US-GDP series shows that this package can also be well applied to a semiparametric multiplicative component model for non-negative time series via the log-transformation. Furthermore, we introduced a semiparametric Log-GARCH and a semiparametric Log-ACD model, which can be easily estimated by the "smoots" package. Of course, this package applies to suitable time series from any other research area. The smoots package also provides a useful tool for teaching time series analysis, because many practical time series follow an additive or a multiplicative component model.

E1671: **Statistical inference for multivariate linear regression models with stochastic volatility**
Presenter: **WenJing Cai**, McGill University, Canada
Co-authors: Jean-Marie Dufour
The Multivariate Linear Regression (MLR) model is extended by assuming the correlated high-order stochastic volatility (SV(P)) structure for the disturbance term. The complication of estimation comes from the high dimension variance-covariance matrix of the disturbance term and the estimation of deep parameters in the SV(p) models. There are no existing references to estimate such a model, and the related estimation is sampling algorithms, which are computationally expensive and initial values dependent. We propose the Regression-based Simple Moment based estimation and ARMA-based estimation. All of them are computationally inexpensive and initial values independent. The analysis provides the asymptotic properties of our estimators and imposes two regularizations to improve efficiency. The simulation study shows that the proposed estimators perform well in terms of lower bias and root mean square error (RMSE) compared with the generalized method of moment (GMM) estimators and Bayesian MCMC estimators. We apply the proposed estimators to construct the prediction intervals from 1988 to 2021 for monthly returns of CRSP. We find that the prediction intervals constructed by the proposed estimators are reliable and have shorter interval widths compared to those constructed by Bayesian MCMC estimators or by regardless of correlated SV(p) structure in the error term.

CI017 Room BH (S) 1.01 Lecture Theatre 1 RECENT ADVANCES IN QUANTILE REGRESSION Chair: Carlos Lamarche

C0176: **Hotelling tubes, confidence bands and conformal inference**
Presenter: **Roger Koenker**, University of Illinois, United States
Stochastic frontier models and methods constitute a rare departure from the usual econometric obsession with models for conditional means. They also provided an early stimulus for the development of quantile regression methods. After a brief tutorial on Hotelling tube methods for constructing confidence bands for nonparametric quantile regression, strengthened performance guarantees for such bands are described based on recent developments in conformal inference. These methods may be considered to be a rather idiosyncratic new approach to nonparametric inference for stochastic frontier models.

C0177: **Inference on quantile processes in partially linear models with applications to the impact of unemployment benefits**
Presenter: **Zhongjun Qu**, Boston University, United States
Co-authors: Jungmo Yoon, Pierre Perron
Methods are proposed to estimate and conduct inference on conditional quantile processes for models with both nonparametric and (locally or globally) linear components. We derive their asymptotic properties, optimal bandwidths, and uniform confidence bands over quantiles allowing for robust bias correction. Our framework covers the sharp regression discontinuity design, which is used to study the effects of unemployment insurance benefits extensions, focusing on heterogeneity over quantiles and covariates. We show economically strong effects in the tails of the outcome distribution. They reduce the within-group inequality, but can be viewed as enhancing between-group inequality, although helping to bridge the gender gap.

C0178: **Unconditional quantile partial effects via conditional quantile regression**
Presenter: **Antonio Galvao**, Michigan State University, United States
Semi-parametric procedures are developed for the estimation and inference of unconditional quantile partial effects using quantile regression coefficients. The main result is based on the fact that, for continuous covariates, unconditional quantile effects are a weighted average of conditional ones. We propose a two-step estimator for the unconditional effects. In the first step, one estimates a structural quantile regression model, and in the second stage, a non-parametric regression is applied. We establish the asymptotic properties of the estimator. Inference is based on estimation of the asymptotic variance of the estimator. Monte Carlo simulations show evidence that the estimator has very good finite sample performance and is robust to the selection of bandwidth and kernel. To illustrate the proposed methods, we study the canonical study of returns to education to estimate unconditional effects.

CO605 Room S0.12 ROBUST ESTIMATION IN STOCHASTIC FRONTIER MODELS Chair: Ian Wright

C0220: **Production analysis with asymmetric noise**
Presenter: **Oleg Badunenko**, Brunel University London, United Kingdom
Co-authors: Daniel Henderson
Symmetric noise is the prevailing assumption in production analysis, but it is often violated in practice. Not only does asymmetric noise cause least-squares models to be inefficient, but it can also hide important features of the data which may be useful to the firm/policymaker. We outline how to introduce asymmetric noise into a production or cost framework as well as develop a model to introduce inefficiency into said models. We derive closed-form solutions for the convolution of the noise and inefficiency distributions, the log-likelihood function, and inefficiency, as well as show how to introduce determinants of heteroskedasticity, efficiency, and skewness to allow for heterogeneous results. We perform a Monte Carlo study and profile analysis to examine the finite sample performance of the proposed estimators. We outline R and Stata packages that we have developed and apply to three empirical applications to show how our methods lead to improved fit, explain features of the data hidden by assuming

symmetry, and how our approach is still able to estimate efficiency scores when the least-squares model exhibits the well-known wrong skewness problem in production analysis.

C0689: **On asymmetry and quantile estimation of the stochastic frontier model**
Presenter: **Ian Wright**, University of Miami, United States
Co-authors: William Horrace, Christopher Parmeter

Quantile regression has become common in applied economic research. Recently, these methods have been adapted for use with the stochastic frontier model. However, the composed nature of the error term is ignored, drawing into question if a stochastic quantile frontier is actually estimated. Here we demonstrate that a particular distributional pair is consistent with the intent of these earlier proposals but is not, in fact, a quantile estimator. A unique feature of this distributional pairing is that both distributions can be asymmetric. We further discuss the identification and practical issues associated with this model.

C0892: **Income stochastic frontiers: Methodological advances for income inequalities investigations**
Presenter: **Graziella Bonanno**, University of Salerno, Italy

The existence of inequalities in income distribution is closely related to the standard conceptualization of efficiency. The idea is to measure the differences between a potential income, obtained for an individual with particular socio-economic characteristics given his investment in human capital, and the income actually received. In particular, we estimate a Mincer equation incorporating human capital variables such as experience, education and occupation. In studies related to earnings frontiers, some scholars use the Stochastic Frontier Approach to get wage efficiency and refer to traditionally approach employing Normally distributed errors. The implicit hypothesis of this specification is that wages follow a log-Normal distribution. However, it has been shown that the latter distribution, particularly in the case of incomes, is not suitable due to the poor ability to describe both the upper and lower tails of the observed distribution of incomes. To overcome this problem, the starting point of our specification is to use the Dagum distribution for the random variable income, for which it has been shown that it fits very well with the entire income distribution. To perform a first empirical analysis in order to test the new SF specification, individuals' data are used derived from IT-SILC (Eurostat), which aims to collect timely and comparable cross-sectional and longitudinal data on income, poverty, social exclusion and living conditions.

C0977: **Distributional assumptions and the sensitivity of stochastic frontier efficiency predictions**
Presenter: **Alex Stead**, University of Leeds, United Kingdom
Co-authors: Phill Wheat, William Greene

Efficiency scores may be used to inform important regulatory, managerial, or policy decisions. In these and other applied settings, it is desirable that they be robust to small changes in the sample. However, it is well known that under standard distributional assumptions, even a single observation can have a dramatic effect on efficiency predictions, e.g. if it leads to 'wrong skewness'. Recent findings show that alternative distributional assumptions can improve the robustness of parameter estimates. We derive influence functions for the conditional mean efficiency predictor, and show that efficiency predictions from robust specifications are less sensitive to contaminating observations than those from non-robust specifications. We also discuss important differences in the way the models handle outliers with respect to prediction uncertainty.

CO208 Room BH (SE) 1.01 ADVANCES IN SEMIPARAMETRIC MODELS FOR PANEL DATA Chair: Daniel Henderson

C1007: **Recasting investment efficiency in China: A semiparametric stochastic efficiency investment panel model**
Presenter: **Taining Wang**, Capital University of Economics and Business, China
Co-authors: Taining Wang, Feng Yao

A semiparametric model is proposed for investment efficiency analysis with three advantages over conventional models. First, our model distinguishes the cause of investment inefficiency due to underinvestment (by financial constraints) from that due to overinvestment (by agency costs). Furthermore, we assume firms make one single decision by choosing under, over, or efficient investment with probabilities influenced by financial friction variables. These are absent features in existing models, which either explicitly assumes the absence of agency costs or implicitly assume an infeasible scenario where underinvestment and overinvestment appear simultaneously. Second, our model is not restricted by parametric specifications widely imposed in existing models, thus alleviating the risk of rendering misleading inferences on investment inefficiency due to model misspecification. Third, our model disentangles fixed effects from investment inefficiency, and circumvents incidental parameter problems through model transformation. We apply our model to investigate investment inefficiency using a panel of Chinese-listed firms during 2006-2020. Both inefficient investments in firms with different ownerships are, on average, persistent and even intensified over the past 15 years.

C1001: **Functional-coefficient regression models for panel data**
Presenter: **Xiao Huang**, Kennesaw State University, United States

The local linear regression method is used to estimate functional coefficients in panel data with interactive fixed effects. The use of interactive fixed effects not only captures cross-section dependence but also mitigates the endogeneity problem while maintaining flexible context-based covariate effects. We establish the uniform strong consistency result of the functional-coefficient estimator. A cross-validation procedure combined with an iterative algorithm is used in Monte Carlo simulation to demonstrate the good finite sample properties of the proposed estimator.

C0217: **Efficient estimation of a semiparametric panel data model with common factors and spatial dependence**
Presenter: **Alexandra Soberon**, Universidad de Cantabria, Spain
Co-authors: Antonio Musolesi, Juan Manuel Rodriguez-Poo

International carbon markets are an appealing and increasingly popular tool for countries to regulate carbon emissions. By putting a price on carbon, carbon markets make pollution less attractive for regulated firms. However, many observers remain skeptical of initiatives such as the European Union Emissions Trading System (EUETS), whose price remained low (compared to the social cost of carbon). The aim is to shed light on this dilemma by analyzing the effect of the EU ETS on CO2 emissions with a semiparametric panel data model where several types of cross-sectional dependence (CSD) and heteroscedasticity are allowed. A new estimator that extends the common correlated effect (CCE) approach to this framework is proposed. However, the initial estimator ignores the CSD and heteroscedasticity, which will lead to a loss of efficiency. Thus, Generalized Least Squares (GLS)-type estimators are proposed. Under rather standard conditions, the parametric estimators are shown to be root-NT consistent, and the asymptotic normality of the nonparametric estimators is also established. Further, the GLS-type estimators are shown to dominate the other. Small sample properties of the estimators are investigated by Monte Carlo experiments and an empirical application on the effect of the EU ETS is conducted.

C0926: **Estimation and inference for varying-coefficient multidimensional fixed-effects panel data models**
Presenter: **Daniel Henderson**, University of Alabama, United States
Co-authors: Christopher Parmeter, Alexandra Soberon

A general estimation method is presented for a varying coefficient multidimensional panel data regression model and offers an array of hypothesis testing avenues. We derive the asymptotic distribution of our estimator, and to construct valid tests, we develop the necessary central limit theory to conduct inference. The presence of multiple effects over differing dimensions requires nontrivial changes to the central limit theory for U-statistics. The types of inference we can conduct offer a diverse array of hypotheses for applied work, and we explicitly present test statistics for some of the most important hypothesis tests. A detailed set of simulations supports our estimators asymptotic developments and reveals that our testing

infrastructure possesses correct asymptotic size and high power.

CO042 Room BH (SE) 2.05 REGIME CHANGE MODELING II (VIRTUAL) Chair: Giovanni Di Bartolomeo

C1339: **Modelling resilience to environmental shocks**
Presenter: **Michael Kuhn**, International Institute for Applied Systems Analysis, Austria
Co-authors: Stefan Wrzaczek
A simple economic-ecological framework of resource extraction is studied that explicitly incorporates (i) a regime-changing shock and (ii) the scope for a Skiba-point to lead to divergent behaviors towards either systemic recovery or system collapse following the shock. This structure allows us to model resilience (in the sense of regime-changing shocks either being avoided or systemic recovery being possible and optimal) and its behavioral implications in a meaningful way. Specifically, we propose a model-based measure of resilience that, upon proper calibration of the model can be employed in numerical assessments of the implications for the resilience of different extraction policies. Applying a framework of marginal valuations of state variables and model parameters allows improving analytical insights into the model and distinguishing different channels driving the optimal behavior. We also study a simple example and show that for that particular case, the anticipation of shocks is leading to less precautionary behavior and would, therefore, compromise systemic resilience.

C1398: **Green transition, investment horizon, and dynamic portfolio decisions**
Presenter: **Willi Semmler**, New School for Social Research, United States
Co-authors: Ibrahim Tahri, Joao Braga
The purpose is to analyze the implications of investors' short-term oriented asset holding and portfolio decisions (or short-termism), and its consequences on green investments. We adopt a dynamic portfolio model, which, contrary to conventional static mean-variance models, allows us to study optimal portfolios for different decision horizons. Our baseline model contains two assets, one asset with fluctuating returns and another asset with a constant risk-free return. The asset with fluctuating returns can arise from fossil-fuel-based sectors or from clean energy-related sectors. We consider different drivers of short-termism: the discount rate, the nature of discounting (exponential vs hyperbolic), and the decision horizon of investors themselves. We study first the implications of these determinants of short-termism on the portfolio wealth dynamics of the baseline model. We find that portfolio wealth declines faster with a higher discount rate, with hyperbolic discounting, and with a shorter decision horizon. We extend our model to include a portfolio of two assets with fluctuating returns. For both model variants, we explore the cases where innovation efforts are spent on fossil fuel or clean energy sources. Detailing dynamic portfolio decisions in such a way may allow us for better pathways to empirical tests and may provide guidance to some online financial decision-making.

C1407: **Sustainable investment under inflation and interest rate risks**
Presenter: **Ibrahim Tahri**, PIK (Potsdam Institute for Climate Impact Research), Germany
The aim is to study an optimal investment problem, in the presence of stochastic inflation and interest rates. This is of particular interest, given the current context of rising inflation. Over an extended period of time, there could be significant changes in the interest rate and inflation rate; hence, uncertainty about these rates may have significant impacts on investment decisions, in particular, towards environmental or climate-friendly investments (due to their long-term nature). The agents' investment opportunity set consists of four instruments; a saving account, two stocks (one general and one green), and a green bond. The investment decisions are made so as to maximize an expected power utility on terminal wealth. We consider two specific cases; one with a nominal green bond and the other case when the green bond is inflation-indexed (IIB). The purpose is to portray the importance of IIBs in hedging inflation risk. The theoretical model is estimated using current market data, in order to construct the optimal investment strategy for an actual real market situation.

C0869: **The U.S. economic dynamics and inflation persistence: A regime-switching perspective**
Presenter: **Giovanni Di Bartolomeo**, Sapienza University of Rome, Italy
Co-authors: Elton Beqiraj, Giuseppe Ciccarone
The purpose is to contribute to the debate on the causes of the Great Moderation by revisiting the U.S. business cycle accounting for switches in the inflation-intrinsic persistence, formalized as hazard-function changes. We trace the existing contrasting evidence back to an identification problem that biases regime estimates and leads to misleading interpretations. Once we account for persistence switches, the empirical outcomes provide clear support to the good luck interpretation and suggest reinterpreting monetary regimes more in line with the central bankers' view. Structural changes in price and wage adjustments are also shown to play important and opposite roles in the Great Inflation.

CO424 Room BH (SE) 2.09 CONTEMPORARY ISSUES IN MODELLING AND FORECASTING INFLATION Chair: Svetlana Makarova

C0208: **Inflation at risk**
Presenter: **Francesca Loria**, Federal Reserve Board, United States
The purpose is to investigate how macroeconomic drivers influence the predictive inflation distribution and establish two key findings. First, the recent muted response of the conditional mean of inflation to economic conditions does not convey a complete picture of inflation dynamics. Indeed, we find ample variability in the tails of the inflation outlook that remains even when focusing on the most recent period of stable and low mean inflation. Second, we document that tight financial conditions carry substantial downside inflation risks in the United States and in the Euro Area, a feature overlooked by much of the literature but consistent with financial amplification mechanisms. Finally, we show that evidence from financial market quotes, from survey data and from a regime-switching model of inflation is consistent with our findings and use our model to track inflation risks during the Covid-19 crisis.

C0381: **Big data forecasting of South African inflation**
Presenter: **Kevin Kotze**, University of Cape Town, South Africa
The use of statistical learning techniques and big data to enhance the accuracy of inflation forecasts is investigated. We make use of a large dataset for the disaggregated prices of consumption goods and services, which we partially reconstruct, and a large suite of different statistical learning and traditional time series models. We find that the statistical learning models are able to compete with most benchmarks over medium to longer horizons, despite the fact that we only have a relatively small sample of available data, but are usually inferior over shorter horizons. Our findings suggest that this result may be attributed to the ability of these models to make use of relevant information, when it is available, and may be particularly useful during periods of crisis, when deviations from the steady state are more persistent. We find that the accuracy of the central bank's near-term inflation forecasts compares favourably with those of other models, while the inclusion of off-model information, such as electricity tariff adjustments and other sources of within-month data, provides these models with a competitive advantage. Lastly, we generate Shapley values for selected statistical learning models to identify the most important contributors to future inflationary pressure and also investigate the relative performance of the different models as we experienced the effects of the pandemic.

C0437: **Global inflation forecasting: Benefits from machine learning methods**
Presenter: **Tobias Soussi**, Aarhus University, Denmark
Co-authors: Marcelo Medeiros, Erik Christian Montes Schutte
Inflation forecasting for a vast panel of countries is considered. We combine the information from common factors driving global inflation and country-specific inflation to build a set of different models. We also rely on new advances in the Machine Learning literature. We show that random

forests and neural networks are very competitive models, and their superiority, although stable across most of the time period considered, increases during recessions. We also show that it is easier to forecast countries with more developed economies. The forecasting gains seem to be partially explained by the degree of trade openness and the volatility of inflation within a year.

C0452: **Inflation expectations and consumption with machine learning**
Presenter: **Lenno Uuskula**, University of Tartu, Estonia
Co-authors: Diana Gabrielyan

Measures of inflation expectations are extracted from online news to build real interest rates that capture underlying consumer expectations. The new measure is infused into various Euler consumption models. While benchmark models based on traditional risk-free returns rates fail, models built with novel news-driven inflation expectations indices improve upon benchmark models and result in strong instruments. Our positive findings highlight the role played by the media for consumer expectation formation and allow for the use of such novel data sources for other key macroeconomic relationships.

CO146 Room BH (SE) 2.10 PARAMETER UNCERTAINTY IN PORTFOLIO OPTIMIZATION AND ASSET PRICING Chair: Nathan Lassance

C0215: **On the optimal combination of naive and mean-variance portfolio strategies**
Presenter: **Rodolphe Vanderveken**, UCLouvain, Belgium
Co-authors: Nathan Lassance, Frederic Vrins

A disheartening fact in portfolio choice is that the naive equally weighted portfolio often outperforms the estimated optimal mean-variance portfolio out of sample. The value of portfolio optimization is reaffirmed by combining the two portfolios to optimize out-of-sample performance. We show that the seemingly natural constraint that the two combination weights sum to one is unnecessary and has several undesirable consequences. In particular, the resulting portfolio combination overinvests in the mean-variance portfolio and underperforms the risk-free asset for sufficiently risk-averse investors. We derive the combination of the equally weighted and mean-variance portfolios that relaxes the constraint and prove that it avoids the undesirable properties of the constrained combination. Moreover, we demonstrate that even though the optimal combination coefficients suffer from more estimation error than the constrained ones, the optimal portfolio combination delivers better out-of-sample performance for most risk-aversion levels. The empirical analysis confirms the superiority of our approach relative to the previous rule and other benchmarks. In general, our novel portfolio rules deliver out-of-sample gains relative to the equally weighted portfolio and the risk-free asset for any degree of risk aversion, and hence render portfolio theory beneficial to all investors with mean-variance preferences.

C0547: **In-sample and out-of-sample Sharpe ratios of multi-factor asset pricing models**
Presenter: **Xiaolu Wang**, Iowa State University, United States
Co-authors: Raymond Kan, Xinghua Zheng

For many multi-factor asset pricing models proposed in the literature, their implied tangency portfolios have substantially higher sample Sharpe ratios than that of the value-weighted market portfolio. In contrast, such a high Sharpe ratio is rarely delivered by professional fund managers. One reason that real-world investor cannot attain the high sample Sharpe ratios of the multi-factor models is estimation risk. We study the effect of estimation risk on the out-of-sample Sharpe ratio of a multi-factor asset pricing model by obtaining the finite sample distribution of the out-of-sample Sharpe ratio conditional on the observed in-sample Sharpe ratio. For an investor who does not know the mean and covariance matrix of the factors in a model, the out-of-sample Sharpe ratio of an asset pricing model is substantially worse than its in-sample Sharpe ratio. After taking into account estimation risk, many of the multi-factor asset pricing models no longer outperform the value-weighted market portfolio.

C1699: **Shrinking against sentiment: Exploiting behavioral biases in portfolio optimization**
Presenter: **Nathan Lassance**, UCLouvain, Belgium
Co-authors: Alberto Martin-Utrera

The performance of mean-variance portfolios is shown to be the sum of a market and an arbitrage component and the exposure of a mean-variance portfolio to each component is shown to depend on their in-sample performance. Consequently, mean-variance portfolios are highly affected by the arbitrage component and suffer from large estimation errors. However, shrinking the sample covariance matrix of returns toward the identity allows mean-variance portfolios to give more relevance to the market and alleviate the impact of parameter uncertainty. We time the exposure to each component by shrinking more when investor sentiment is low, which provides sizable economic gains with lower turnover than competing benchmarks.

C0555: **Dynamic shrinkage estimation of the high-dimensional minimum-variance portfolio**
Presenter: **Taras Bodnar**, Stockholm University, Sweden
Co-authors: Nestor Parolya, Erik Thorsen

New results in random matrix theory are derived, which allow the construction of a shrinkage estimator of the global minimum variance (GMV) portfolio when the shrinkage target is a random object. More specifically, the shrinkage target is determined as the holding portfolio estimated from previous data. The theoretical findings are applied to develop theory for dynamic estimation of the GMV portfolio, where the new estimator of its weights is shrunk to the holding portfolio at each time of reconstruction. Both cases with and without overlapping samples are considered. The non-overlapping samples correspond to the case when different data of the asset returns are used to construct the traditional estimator of the GMV portfolio weights and to determine the target portfolio, while the overlapping case allows intersections between the samples. The theoretical results are derived under weak assumptions imposed on the data-generating process. No specific distribution is assumed for the asset returns except from the assumption of finite $4+\varepsilon$, $\varepsilon > 0$, moments. Also, the population covariance matrix with an unbounded spectrum can be considered. The performance of new trading strategies is investigated via an extensive simulation. Finally, the theoretical findings are implemented in an empirical illustration based on the returns on stocks included in the S&P 500 index.

CO400 Room BH (SE) 2.12 STATISTICAL ANALYSIS OF CLIMATE DATA Chair: Liudas Giraitis

C1044: **The predictability of sea surface temperatures in El Nino regions**
Presenter: **Tommaso Proietti**, University of Roma Tor Vergata, Italy
Co-authors: Alessandro Giovannelli

The El Nino Southern Oscillation induces the alternation of persistent warming and cooling of sea surface temperatures along the equator in the east-central Pacific. We investigate the predictability and the persistence of the El Nino cycle by considering sea surface temperature anomalies averaged over four rectangular areas that subdivide the equatorial Pacific (Nino 1, 2, 3-4, and 4 regions). Predictability is defined both in terms of the prediction error variance and by the mutual information between past and future. By means of the decomposition of the mutual information, we can assess the information gains arising from modelling the time series jointly. Persistence measures the strength of the auto- and cross-covariances by comparing the long-run variance to the unconditional variance of the series. Our methodology is based on regularized univariate and multivariate Levinson-type algorithms for estimating the autocovariance and the cross-covariance matrix of the series.

C1073: **Climate change heterogeneity: A new quantitative approach**
Presenter: **Jesus Gonzalo**, Universidad Carlos III de Madrid, Spain
Co-authors: Lola Gadea

Climate change is a non-uniform phenomenon. A new quantitative methodology is proposed to characterize, measure, and test the existence of climate change heterogeneity. First, we introduce a new testable warming typology based on the evolution of the trend of the whole temperature distribution and not only on the average. Second, we define the concepts of warming acceleration and warming amplification in a testable format. And third, we introduce the new testable concept of warming dominance to determine whether region A is suffering a worse warming process than that region. Applying this three-step methodology, we find that Spain and the Globe experience clear distributional warming processes (beyond the standard average)but of different types. In both cases, this process is accelerating over time and asymmetrically amplified. Overall, warming in Spain dominates the Globe in all the quantiles except the lower tail of the global temperature distribution that corresponds to the Arctic region. Our climate change heterogeneity results open the door to the need for a non-uniform causal-effect climate analysis that is beyond the standard causality in mean as well as for a more efficient design of the mitigation-adaptation policies. In particular, the heterogeneity we find suggests that these policies should contain a common global component and a clear local-regional element. Future climate agreements should take the whole temperature distribution into account.

C1088: **Nonparametric tests based on records to detect trends in the upper tail of climate series**

Presenter: **Ana C Cebrian**, University of Zaragoza, Spain

Co-authors: Jorge Castillo-Mateo, Jesus Asin

There is clear evidence of global warming in mean temperature. However, not only changes in the mean are important. Changes in the tails of the distributions are also relevant, and the occurrence of hot extremes has serious consequences on human health, agriculture, etc. Some nonparametric tests are presented, based on the theory of records, which are useful to quantify and assess the existence of trends (global warming) in the upper tail of a temperature series. The tests are based on the number of records up to time t and the asymptotic normal distribution of these variables, regardless of the distribution of the original series. Some modifications of the basic statistic, splitting the series and using different types of records (lower and upper and forward and backward records) are proposed to increase the power of the tests. Since global warming is a spatial phenomenon, another test is developed to be applied in a spatial framework. These tests are applied to analyze the effects of global warming in the upper tails of different types of temperature series in a set of locations in the North-East of Spain.

C1154: **Estimation of cyclical time series with application to climate data**

Presenter: **Liudas Giraitis**, Queen Mary University of London, United Kingdom

Co-authors: Fulvia Marotta

A new approach is introduced for the estimation of time series with cyclically varying parameters. An application of this methodology to modelling Central England temperatures allows for uncovering unexpected features of the data-generating model and the change of its patterns and dynamics, which produce the change of daily temperatures over time. The model also provides a useful tool for the analysis and forecasting of changing patterns of climate data using simulations.

Sunday 18.12.2022 15:45 - 17:00 Parallel Session J – CFE-CMStatistics

EV772 Room Virtual R06 COMPUTATIONAL STATISTICS Chair: Aaron Scheffler

E0346: **Simulation-based inference for high dimensional implicit models and application to partially observed processes**
Presenter: **Joonha Park**, University of Kansas, United States

In many applications, a probabilistic model for a given system is defined implicitly by a simulation algorithm. Such implicit models can be simulated at any parameter value, but often the probability density function cannot be evaluated. We consider the case where the system described by an implicit model is partially observed. Parameter estimation for such partially observed, implicit models can, in principle, be carried out by repeated Monte Carlo simulations at various parameter values and comparing the log measurement densities of the observed data. However, the average of log measurement densities has a downward Jensen bias, which increases with increasing model dimensions. Under certain asymptotic assumptions, we develop methods for constructing Monte Carlo confidence intervals for the log-likelihood of data and the maximum likelihood estimate given the data. Furthermore, for models that satisfy local asymptotic normality (LAN), we develop a method for constructing a confidence interval for the unknown parameter value. We show that our methods can enable likelihood-based inference for partially observed, high-dimensional, mechanistic models for stochastic processes using numerical experiments.

E1907: **Component contribution maximization: An estimation approach for stochastic expectation-maximization**
Presenter: **Alexander Sharp**, University of Waterloo, Canada
Co-authors: Ryan Browne

The Stochastic EM algorithm replaces the E-step with a Monte Carlo approximation, trading monotonicity for the potential to escape local maxima. A consequence is that the final parameter value returned by the algorithm is no longer guaranteed to be the best estimate. Common solutions include averaging the tail of the chain, or choosing the value associated with the largest likelihood value. We prove that when the model parameter is a scalar, this second estimator is asymptotically Laplace distributed and consistent for the maximum likelihood estimate (mle), with a convergence rate that is square the convergence rate of the average. We further show, however, that as the parameter dimension increases, this estimator becomes bounded arbitrarily far from the mle. In light of this shortcoming, a new estimator for the high dimensional parameter case is proposed, which we show is consistent and successfully achieves the faster convergence rate previously observed only in the single-dimensional case. We demonstrate through multiple simulation studies the increased performance this estimator provides over topical approaches.

E1633: **Assessment and calculation of model complexity of deep neural networks**
Presenter: **Rene-Marcel Kruse**, University of Goettingen, Germany
Co-authors: Benjamin Saefken, Thomas Kneib

Model selection is an area of research in the field of statistics that receives a lot of attention, with the concept of model averaging being revisited frequently. The idea of qualified and quantifiable model selection, however, so far has received little attention in the field of learning-driven methods such as machine and deep learning. We focus our attention on bringing concepts such as model complexity and degrees of freedom to the field of deep learning to derive a means to leverage the insights to perform data-driven model assessment and model averaging of deep learning models. We illustrate the theoretical and practical problems of translating these statistical techniques to the domain of deep learning. Further, we apply the proposed methods to examples of simulation studies as well as applications based on real-world data to illustrate the validity of our approach.

EO518 Room S-2.23 ADVANCES IN STATISTICAL METHODS FOR BOUNDED DATA Chair: Agnese Maria Di Brisco

E0271: **Non-parametric regression models for compositional data**
Presenter: **Michail Tsagris**, University of Crete, Greece

Compositional data arise in many real-life applications, and versatile methods for properly analyzing this type of data in the regression context are needed. To this end, we consider an extension to the classical *k-NN* regression, termed α-*k-NN* regression, that yields a highly flexible non-parametric regression model for compositional data through the use of the α-transformation. Our model is further extended to the α-kernel regression by adopting the Nadaraya-Watson estimator. Unlike many of the recommended regression models for compositional data, zeros values (which commonly occur in practice) are not problematic, and they can be incorporated into the proposed models without modification. Extensive simulation studies and real-life data analyses highlight the advantage of using these non-parametric regressions for complex relationships between the compositional response data and Euclidean predictor variables. Both suggest that α-*k-NN* and α-kernel regressions can lead to more accurate predictions compared to current regression models, which assume a, sometimes restrictive, parametric relationship with the predictor variables. In addition, the α-*k-NN* regression, in contrast to α-kernel regression, enjoys a high computational efficiency rendering it highly attractive for use with large-scale, massive, or big data.

E0740: **Power logit regression for modeling bounded data**
Presenter: **Silvia Ferrari**, University of Sao Paulo, Brazil
Co-authors: Francisco F Queiroz

The main purpose is to introduce a new class of regression models for bounded continuous data, commonly encountered in applied research. The models, named the power logit regression models, assume that the response variable follows a distribution in a wide, flexible class of distributions with three parameters, namely the median, a dispersion parameter and a skewness parameter. A comprehensive set of tools is offered for likelihood inference and diagnostic analysis, and the new R package PLreg is introduced. Applications with real and simulated data show the merits of the proposed models, the statistical tools, and the computational package.

E0973: **A generalization of the latent Dirichlet allocation**
Presenter: **Roberto Ascari**, University of Milano-Bicocca, Italy
Co-authors: Alice Giampino

Over recent years, text modeling techniques have been employed in several applications, including the detection of latent topics in text documents. A widespread statistical tool for topic modeling is the Latent Dirichlet Allocation (LDA), which allows for a document representation in terms of topic composition. A well-known limitation of the LDA is related to the stiffness of the Dirichlet prior imposed on the topic distributions. The aim is to perform a preliminary study of the flexible Dirichlet (FD) as an alternative prior. The latter is a generalization of the Dirichlet distribution allowing for a finite mixture structure. The introduction of additional parameters ensures more flexibility, still maintaining the model interpretability, as well as conjugacy to the multinomial model. The latter property allows for a Collapsed Gibbs Sampling-based estimation procedure. The generalization of the LDA based on the FD distribution is illustrated via an application to a real dataset.

EO716 Room S-2.25 RECENT ADVANCES IN LEARNING UNDER DISTRIBUTION SHIFTS Chair: Yao Li

E0223: **Adaptive and robust multi-task learning**
Presenter: **Kaizheng Wang**, Columbia University, United States
Co-authors: Yaqi Duan

The purpose is to study the multi-task learning problem that aims to simultaneously analyze multiple datasets collected from different sources and learn one model for each of them. We propose a family of adaptive methods that automatically utilize possible similarities among those tasks

while carefully handling their differences. We derive sharp statistical guarantees for the methods and prove their robustness against outlier tasks. Numerical experiments on synthetic and real datasets demonstrate the efficacy of our new methods.

E0935: **Approximate selective inference via maximum likelihood**
Presenter: **Snigdha Panigrahi**, University of Michigan, United States
Co-authors: Jonathan Taylor

Several strategies have been developed recently to ensure valid inferences after model selection; some of these are easy to compute, while others fare better in terms of inferential power. We will address post-selection inference through approximate maximum likelihood estimation. The goal is to: (i) efficiently utilize hold-out information from selection with the aid of randomization, (ii) bypass expensive MCMC sampling from exact conditional distributions that are hard to evaluate in closed forms. At the core of our new method is the solution to a fairly simple, convex optimization problem in a few dimensions.

E1313: **Defending against backdoor attack**
Presenter: **Yao Li**, University of North Carolina at Chapel Hill, United States

Backdoor attacks are getting increasing attention as studies have shown that federated learning systems can be easily fooled by them. However, defenses against such attacks are not investigated sufficiently in federated learning. Federated learning is an emerging machine learning technique as it addresses the problem of data privacy by updating models on local clients and aggregating the global model without accessing the local data. However, such distributed nature makes it vulnerable to backdoor attacks, as attackers can send malicious model updates to insert a backdoor in the global model. We study the differences between malicious updates and benign updates and propose a detection method to filter out malicious updates from attackers to protect the federated training process.

EO620 Room S-1.01 CLUSTERING OF COMPLEX DATA STRUCTURES Chair: Maria Brigida Ferraro

E0306: **Time series clustering based on prediction accuracy of global forecasting models**
Presenter: **Angel Lopez Oriona**, Universidade da Coruña, Spain
Co-authors: Jose Vilar, Pablo Montero-Manso

A novel method to perform clustering of time series is proposed. The procedure is based on the traditional K-means clustering algorithm and relies on two iterative steps: (i) K global forecasting models are fitted via pooling by considering the series pertaining to each cluster and (ii) each series is assigned to the group associated with the model producing the best forecasts according to a particular criterion. The resulting clustering partition contains groups which are optimal in terms of overall forecasting error and thus, the technique is able to detect the different prediction patterns existing in a given database. A simulation study shows that our method outperforms several alternative procedures concerning both clustering effectiveness and forecasting accuracy. The approach is also applied to perform clustering in three real-time series datasets.

E0779: **Impact of missing data on mixtures and clustering**
Presenter: **Christophe Biernacki**, Inria, France

The frequency of missing data increases with the growing size of modern datasets, making this topic important in the research agenda of statisticians. First, we introduce the MCAR mechanism for mixed data (quantitative and categorical) mixture models and illustrate it on a biological data set. Second, as a more theoretical but important step, we discuss the impact of missing values on the EM algorithm for Gaussian mixtures in the MAR situation. We exhibit the fact that the quite familiar degeneracy problem is aggravated during the EM runs, leading to dangerously slow and also more frequent events than with complete data. Finally, we discuss the impact of missing not-at-random values (MNAR mechanism) on the partition estimation provided from mixtures (Gaussian or not). In particular, we defend the advantage of embedding the missingness mechanism directly within the clustering modeling step. A new MNAR model is introduced, discussed and experimented on a medical data set.

E1159: **Deep clustering: A new clustering method in the sequential approach**
Presenter: **Claudia Rampichini**, University of Rome La Sapienza, Italy

Deep clustering is a recent technique that exploits the potential of neural networks to overcome the problems of conventional clustering methods. Specifically, an autoencoder neural network is used to obtain a dimensional reduction in which a clustering algorithm is involved; the approach can be sequential or simultaneous. The focus is on a new clustering method that uses and enhances the information provided by the membership degrees, derived from a fuzzy algorithm, in order to improve clustering performance. First, a fuzzy algorithm is applied, and then the units that have an unclear assignment are reclassified using a crisp algorithm. A unit has an unclear assignment when membership degrees are close to each other. A summary measure of membership degrees is chosen and, based on a threshold value, a subset of the units are reclassified. The adequacy of the proposal is checked by means of several benchmark data sets and the results show margins for improvement in performance compared to both fuzzy and crisp algorithms.

EO186 Room S-1.04 MARGINAL AND CONDITIONAL INFERENCE FOR DEPENDENT DATA Chair: Glen McGee

E0441: **Marginal additive models: Simultaneous marginal and conditional non-linear regression for cluster-correlated data**
Presenter: **Alex Stringer**, University of Waterloo, Canada
Co-authors: Glen McGee

Regression models for cluster-correlated data model either the population-averaged or the cluster-conditional mean response. We introduce a Marginal Additive Model (MAM), which produces simultaneous estimates of cluster-conditional and population-averaged effects in regression models with non-linear covariate effects, including longitudinal and spatial models. The method is applied to a longitudinal study of beaver foraging habits, in which population-averaged inferences are desired, but only cluster-conditional inferences had previously been made; and a well-known spatial analysis of loa loa parasite infection rates, in which we argue that the usual assumption of independence of responses at the same spatial location presumably made due to the lack of methods for fitting marginal models to spatial data with additional within cluster-correlation is inappropriate, a challenge to which the MAM offers a solution. On the technical side, standard errors are obtained using efficient numerical linear algebra that avoids storing square matrices whose dimension depends on the sample size, a situation that occurs when attempting to apply the delta method in this context.

E0444: **A transformation perspective on marginal and conditional models**
Presenter: **Torsten Hothorn**, University of Zurich, Switzerland
Co-authors: Luisa Barbanti

Clustered observations are ubiquitous in controlled and observational studies and arise naturally in multicenter trials or longitudinal surveys. We present a novel model for the analysis of clustered observations where the marginal distributions are described by a linear transformation model and the correlations by a joint multivariate normal distribution. The joint model provides an analytic formula for the marginal distribution. Owing to the richness of transformation models, the techniques are applicable to any type of response variable, including bounded, skewed, binary, ordinal, or survival responses. We discuss the analysis of two clinical trials aiming at the estimation of marginal treatment effects. In the first trial, the pain was repeatedly assessed on a bounded visual analog scale and marginal proportional-odds models are presented. The second trial reported disease-free survival in rectal cancer patients, where the marginal hazard ratio from Weibull and Cox models is of special interest. An implementation is available in the "tram" add-on package to the R system and was benchmarked against established models in the literature.

E1078: **A class of generalized linear mixed models adjusted for marginal interpretability**
Presenter: **Peter Craigmile**, The Ohio State University, United States

Two popular approaches for relating correlated measurements of a non-Gaussian response variable to a set of predictors are to fit a marginal model using generalized estimating equations and to fit a generalized linear mixed model (GLMM) by introducing latent random variables. The first approach is effective for parameter estimation, but leaves one without a formal model for the data with which to assess the quality of fit or make individual-level predictions for future observations. The second approach overcomes these deficiencies, but leads to parameter estimates that must be interpreted conditional on the latent variables. To obtain marginal summaries, one needs to evaluate an analytically intractable integral or use attenuation factors as an approximation. Further, we note an unpalatable implication of the standard GLMM. To resolve these issues, we turn to a class of marginally interpretable GLMMs that lead to parameter estimates with a marginal interpretation while maintaining the desirable statistical properties of a conditionally specified model and avoiding problematic implications. We establish the form of these models under the most commonly used link functions and address computational issues. For logistic mixed effects models, we introduce an accurate and efficient method for evaluating the logistic-normal integral.

EO662 Room S-1.06 ADVANCES IN NONPARAMETRIC CONTROL CHARTS Chair: Manuela Cazzaro

E0310: **Transparent sequential learning for nonparametric sequential process monitoring**
Presenter: **Peihua Qiu**, University of Florida, United States

Machine learning methods have been widely used in process control and monitoring. For handling statistical process control (SPC) problems, conventional supervised machine learning methods would have difficulties because a required training dataset containing both in-control and out-of-control (OC) process observations is rarely available in SPC applications. In addition, many machine learning methods work like black boxes, and it is difficult to interpret their learning mechanisms. In the SPC literature, there have been some existing discussions on how to handle the lack of OC observations in the training data, using the one-class classification, artificial contrast, and some other ideas. However, these approaches have their own limitations. We present a recent method that extends the self-starting process monitoring idea to a general learning framework for monitoring processes with serially correlated data. Under the new framework, process characteristics to learn are well specified in advance, and process learning is sequential in the sense that the learned process characteristics keep being updated during process monitoring. The learned process characteristics are then incorporated into a control chart for detecting process distributional shifts based on all available data by the current observation time. Numerical studies show that process monitoring based on the new learning framework is reliable and effective for SPC.

E0818: **Multivariate control charts based on the L^p depth**
Presenter: **Giuseppe Pandolfo**, University of Naples Federico II, Italy
Co-authors: Carmela Iorio, Michele Staiano, Massimo Aria, Roberta Siciliano

When monitoring key quality features of a process via multivariate control charts, previous knowledge may not be enough to adopt a unique model for all the variables. In the case no specific parametric model turns out to be appropriate, alternative solutions have to be considered and adopting nonparametric methods to build control charts appears a reasonable choice. Among the existing non-parametric techniques, data depth functions are gaining a growing interest in multivariate quality control. Within the literature, several notions of depth are effective for this purpose, even in the case of deviation from the normality assumption. However, the use of the L^p depth has been surprisingly neglected so far. Hence, the goal is to investigate the behaviour of the L^p depth in the statistical process control and to compare its performances to those of the Mahalanobis depth, which is often adopted to build depth-based control charts.

E0721: **A novel approach to the change-point methodology in nonparametric control charts**
Presenter: **Claudio Giovanni Borroni**, University of Milano - Bicocca, Italy
Co-authors: Manuela Cazzaro, Paola Maddalena Chiodini

When a stream of single measurements is available, the change-point paradigm is often applied to monitor for possible shifts in location, scale, or other characteristics of the underlying distribution. Any test statistics useful for the two-sample problem can be adapted to this task, by computing it on every split of the stream into two parts and by looking at its maximum value on the set of all possible splits. Especially when no prior assumptions on the distribution are made, that is when a nonparametric chart is built; however, some problems arise due to the sample sizes, which are occasionally imposed on the two-sample test statistic. More specifically, some splits necessarily produce two largely unbalanced samples, of which one has a size too limited to generate a sensible comparison. We propose to revise the change-point paradigm to avoid such inconveniences and, specifically, to work just with tests based on two balanced samples. Our methodology can be applied both to a phase-I control chart, both to a self-starting chart, which we believe to be its more useful application. We show that the determination of critical values to guarantee the desired ARL in the in-control state is made easier, and we explore conditions under which the chart can lead to competitive values of the ARL in the out-of-control state.

EO735 Room S-1.22 ENTITY RESOLUTION, BIOMEDICAL AND NUCLEAR FORENSICS DATA MODELING Chair: Sharmistha Guha

E1347: **Pathway analysis over brain structural network with a survival outcome**
Presenter: **Yize Zhao**, Yale University, United States
Co-authors: Xinyuan Tian, Fan Li, Denise Esserman, Li Shen, Xiwen Zhao

Technological advancements in noninvasive imaging techniques provide an unprecedented opportunity to understand how the human nervous system is supported by molecular profiles and how it affects behaviors through the construction of whole brain interconnections on white matter fiber tracts, known as brain structural connectivity. Existing approaches to analyze structural connectivity frequently disaggregate the entire network into a vector of unique edges or summary measures, leading to a substantial loss of information. We propose an integrative Bayesian framework to model the effect pathway between each component and the potential mediating role of brain structural connectivity between genetic exposure and survival outcome. To accommodate the neurobiological architectures of connectivity, including symmetry, hollow and dense interconnections among hub nodes, we develop a structural modeling framework including a symmetric matrix-variate AFT model, and a symmetric matrix response regression to characterize the effect paths. We further impose within-graph sparsity and between-graph shrinkage to identify informative network configurations and eliminate the interference of noisy components. Extensive simulations confirm the superiority of our method compared with existing alternatives. We apply the proposed method to analyze the landmark ADNI study, and obtain neurobiologically plausible insights.

E1458: **Overcoming censored predictors with imputation to model the progression of Huntingtons Disease**
Presenter: **Sarah Lotspeich**, Wake Forest University, United States

Clinical trials to test experimental treatments for Huntington's disease are expensive, so it is prudent to enrol subjects whose symptoms may be most impacted by the treatment during follow-up. However, modeling how symptoms progress to identify such subjects is problematic since time to diagnosis, a key predictor, can be censored. Imputation is an appealing strategy where censored predictors are replaced with their conditional means, the calculation of which requires estimating and integrating over its conditional survival function from the censored value to infinity. However, despite efforts to make conditional mean imputation as flexible as possible, it still makes restrictive assumptions about the censored predictor (such as proportional hazards) that may not hold in practice. We develop a suite of extensions to conditional mean imputation to encourage its applicability to a wide range of clinical settings. We adopt new estimators for the conditional survival function to offer more efficient and robust inference and propose an improved conditional mean calculation. We discuss in simulations when each version of conditional mean imputation is

most appropriate and evaluate our methods as we model symptom progression from Huntington's disease data. Our imputation suite is implemented in the open-source R package, imputeCensRd.

E1430: **Inverse prediction using functional data in a Bayesian framework**
Presenter: **Audrey McCombs**, Sandia National Laboratories, United States
Co-authors: Katherine Goode, Kurtis Shuler, Derek Tucker, Adah Zhang, Daniel Ries

Inverse prediction models have commonly been developed to handle scalar data from physical experiments. However, it is not uncommon for data to be collected in functional form, after which it must be aggregated to fit the structure of traditional methods. The resulting loss of information can be costly in expensive experiments. The functional inverse prediction (FIP) framework is a general approach which uses the full information in functional response data to provide inverse predictions. We build upon the FIP framework by applying Bayesian methods, creating more seamless uncertainty quantification, and adding flexibility through seemingly unrelated regression (SUR). Basis functions represent the functional response data in a matrix of response variables, which are regressed against a matrix of predictor variables for which inverse prediction is desired. Each variable in the response matrix defines a valid regression with its own set of predictor variables and mean function form, with error terms across the regression equations assumed to be correlated. The Bayesian implementation of the FIP is demonstrated with an application to nuclear forensics. Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC, a wholly-owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

EO150 Room S-1.27 MODEL-FREE INFERENCE (VIRTUAL) Chair: Asaf Weinstein

E1992: **Risk control for online learning models**
Presenter: **Yaniv Romano**, Technion—Israel Institute of Technology, Israel

Modern machine learning algorithms have achieved remarkable performance in a myriad of applications, and are increasingly used to make impactful decisions in the hiring process, criminal sentencing, and healthcare diagnostics. The use of data-driven algorithms in high-stakes applications is exciting yet alarming: these methods are extremely complex, often brittle, and notoriously hard to analyze or interpret. Naturally, concerns have been raised about the reliability of the output of such machines. The focus is on making reliable predictions in an online setting, in which the underlying data distribution can drastically—and even adversarially—shift over time. We will introduce statistical tools that can be wrapped around any online "black-box" machine learning model to provide valid and informative uncertainty estimates.

E0997: **BONuS: Multiple multivariate testing with a data-adaptive test statistic**
Presenter: **Chiao-Yu Yang**, UC Berkeley, United States
Co-authors: Lihua Lei, Nhat Ho, William Fithian

A new adaptive empirical Bayes framework is proposed, the Bag-Of-Null-Statistics (BONuS) procedure, for multiple testing where each hypothesis testing problem is itself multivariate or nonparametric. BONuS is an adaptive and interactive knockoff-type method that helps improve the testing power while controlling the false discovery rate (FDR), and is closely connected to the "counting knockoffs" procedure. Contrary to procedures that start with a -value for each hypothesis, our method analyzes the entire data set to adaptively estimate an optimal -value transform based on an empirical Bayes model. Despite the extra adaptivity, our method controls FDR in finite samples even if the empirical Bayes model is incorrect or the estimation is poor. An extension, the Double BONuS procedure, validates the empirical Bayes model to guard against power loss due to model misspecification.

E2016: **Learn then test: Calibrating predictive algorithms to achieve risk control**
Presenter: **Stephen Bates**, UC Berkeley, United States

Learn then Test is introduced, a framework for calibrating machine learning models so that their predictions satisfy explicit, finite-sample statistical guarantees regardless of the underlying model and (unknown) data-generating distribution. The framework addresses, among other examples, false discovery rate control in multi-label classification, intersection-over-union control in instance segmentation, and the simultaneous control of the type-1 error of outlier detection and confidence set coverage in classification or regression. To accomplish this, we solve a key technical challenge: the control of arbitrary risks that are not necessarily monotonic. Our main insight is to reframe the risk-control problem as multiple hypothesis testing, enabling techniques and mathematical arguments different from those in the previous literature. We use our framework to provide new calibration methods for several core machine learning tasks with detailed worked examples in computer vision

EO712 Room K0.16 ADVANCES IN MULTIPLE NETWORK DATA ANALYSIS Chair: Jesus Arroyo

E0607: **Communication network dynamics in a large organizational hierarchy**
Presenter: **Nathaniel Josephs**, Yale University, United States

Most businesses impose a supervisory hierarchy on employees to facilitate the management, decision-making, and collaboration. In contrast, routine inter-employee communication patterns within workplaces tend to emerge more naturally, as a consequence of both supervisory relationships and the needs of the organization. Scholars of organizational management have proposed theories relating organizational trees to communication dynamics and measures of business performance. Separately, network scientists have studied the topological structure of communication patterns in different types of organizations. However, the nature of the relationship between a formal organizational structure and emergent communications between employees remains unclear. To address this, we study associations between organizational hierarchy and communication dynamics among approximately 200,000 employees of a large software company. We propose new measures of communication reciprocity and new shortest-path distances for trees to characterize the frequency of messages passed up, down, and across the organizational hierarchy. We discuss the relationship of routine employee communication patterns to supervisory hierarchies in this company, and empirically evaluate several theories of organizational management and performance.

E0611: **Dependent structures in network data**
Presenter: **Sharmodeep Bhattacharyya**, Oregon State University, United States
Co-authors: Shirshendu Chatterjee, Soumendu Sundar Mukherjee

Statistical analysis of networks generated from exchangeable network models has been extensively studied in the literature. One primary property of exchangeable network models is the conditional independence of edge formation. We extend the framework of network formation to include dependent edges with an emphasis on generating networks with all five properties of sparsity, small-world, community structure, power-law degree distribution, and transitivity or high triangle count. We propose a class of models, called Transitive Inhomogeneous Erdos-Renyi (TIER) models, which we show have all five properties. We also perform inferential tasks, such as parameter estimation, community detection, and change-point detection for sequences of dependent networks from Inhomogeneous Erdos-Renyi (IER) and TIER models. We validate our results using simulation studies too. If time permits, we will talk about some recent developments in the estimation of the number of communities using Bethe Hessian matrices.

E0703: **Lost in the shuffle: Testing power in the presence of errorful network vertex labels**
Presenter: **Vince Lyzinski**, University of Maryland, College Park, United States
Co-authors: Ayushi Saxena

Many two-sample network hypothesis testing methodologies operate under the implicit assumption that the vertex correspondence across networks is a priori known. We consider the degradation of power in two-sample graph hypothesis testing when there are misaligned/label-shuffled vertices across networks. In the context of stochastic block model networks, we theoretically explore the power loss due to shuffling for a pair of hypothesis tests based on Frobenius norm differences between estimated edge probability matrices or between adjacency matrices. The loss in testing power is further reinforced by numerous simulations and experiments, both in the stochastic block model and in the random dot product graph model, where we compare the power loss across multiple recently proposed tests in the literature. Lastly, we demonstrate the impact that shuffling can have in real-data testing in a pair of examples from neuroscience and from social network analysis.

EO634 Room K0.18 BAYESIAN NONPARAMETRICS FOR CAUSAL INFERENCE: PART II Chair: Arman Oganisian

E0431: **Understanding the spillover effects of the air pollution mixture using mobility data**
Presenter: **Joseph Antonelli**, University of Florida, United States

Estimating the causal effects of air pollution is an important problem as we require a better understanding of the nature of this relationship in order to guide future regulation. Of particular interest is the impact of air pollution mixtures, i.e. the joint impact of multiple air pollutants on health. Many cohort studies assign pollution levels to individuals based on their home zip code, but individuals travel to multiple zip codes with potentially different pollution levels. To provide a better understanding of the overall impact of the air pollution mixture, we target the effects of both exposure to pollution within an individual's zip code and exposure to pollution from other zip codes. We use a weighted average of exposure to pollution from other zip codes by incorporating cell phone mobility data. Using nonparametric Bayesian models, we then estimate the spatial spillover effect of air pollution exposure.

E0730: **A Bayesian nonparametric approach for principal causal effects**
Presenter: **Chanmin Kim**, SungKyunKwan University, Korea, South

In estimating the causal effect, principal stratification analysis is a method for interpreting the effect of treatment on the outcome based on the relationship between treatment and post-treatment (intermediate) variables. In general, modeling of an intermediate variable is required when the intermediate variable is continuous, but existing parametric modeling methods are difficult to fully capture the complex relationship between variables. Furthermore, estimating the outcome model and intermediate model separately makes it difficult to account for the uncertainty introduced by each model estimation in the final causal effect estimation. Using the Bayesian additive regression trees model, we propose a fully Bayesian method. All intermediate, outcome, and propensity score models are flexibly estimated using Bayesian nonparametric models, unlike other flexible methods in the literature. Also, the proposed method is very useful in both a specific confounding situation (referred to as targeted selection) and a broad confounding situation. A simulation study is used to demonstrate this. With the proposed method, we examine the impact of the sulfate abatement device (scrubber) installed in US coal-fired power plants on surrounding PM2.5 concentrations from various perspectives, based on the relationship between the scrubber and SO2 emissions.

E1105: **Causal framework for subgroup treatment evaluation using multivariate generalized mixed effect models**
Presenter: **Yizhen Xu**, Johns Hopkins University, United States
Co-authors: Jisoo Kim, Ami Shah, Scott Zeger

Dynamic prediction of causal effects under different treatment regimes conditional on individual's characteristics and longitudinal history is an essential problem in precision medicine. This is a challenging problem in practice because outcomes and treatment assignment mechanisms are unknown in observational studies, individual's treatment efficacy is a counterfactual, and the existence of selection bias is empirically untestable. We propose a framework for identifying the long-term individualized treatment effect adjusting for unobserved stable trait factors, using Bayesian G computation with multivariate generalized mixed effect models. Existing methods mostly focus on balancing the confounder distributions of observables between different treatments, while our proposal also accounts for a latent tendency towards each treatment due to unobserved time-invariant factors. We assume sequential ignorability conditional on unobserved stable trait factor in treatment assignment, and dynamically updates stable unobserved factors in outcomes progression as an individual's history data increases over time. Our framework naturally incorporates sensitivity analysis, providing an alternative to defining an additional sensitivity parameter for quantifying the impact of unmeasured confounding.

EO705 Room K0.20 ADVANCES IN EXTREME VALUE STATISTICS Chair: Ioannis Papastathopoulos

E1697: **Neural networks for extreme quantile regression with an application to forecasting flood risk**
Presenter: **Olivier Pasche**, University of Geneva, Switzerland
Co-authors: Sebastian Engelke

Risk assessment for extreme events requires accurate estimation of high quantiles that go beyond the range of historical observations. When the risk depends on the values of observed predictors, regression techniques are used to interpolate in the predictor space. We propose the EQRN model that combines tools from neural networks and extreme value theory into a method capable of extrapolation in the presence of complex predictor dependence. Neural networks can naturally incorporate additional structure in the data. We develop a recurrent version of EQRN that is able to capture complex sequential dependence in time series. We apply this method to forecasting of flood risk in the Swiss Aare catchment. It exploits information from multiple covariates in space and time to provide one-day-ahead predictions of return levels and exceedances probabilities. This output complements the static return level from a traditional extreme value analysis, and the predictions are able to adapt to distributional shifts as experienced in a changing climate. Our model can help authorities to manage flooding more effectively and to minimize their disastrous impacts through early warning systems.

E1961: **Modelling non-stationarity in asymptotically independent extremes**
Presenter: **Callum Murphy-Barltrop**, Lancaster University, United Kingdom

In many practical environmental applications, it is important to evaluate the joint extremal risk from two or more variables. However, the variables of interest often exhibit non-stationarity: consequently, the vast majority of approaches for multivariate extremes, where data is assumed to be identically distributed, are not applicable in this setting. Moreover, non-stationary trends often exist within marginal distributions and dependence structures simultaneously, resulting in complex data structures. Few approaches have been proposed for capturing such structures in the extremes literature to date. We propose a flexible semi-parametric modelling framework for capturing trends in asymptotically independent extremes. We show this framework is able to accurately capture a broad range of extremal dependence trends across simulate examples. We also demonstrate our approach using temperature data from the UK Climate Projections from 1980-2080. Marginal trends are first accounted for via a pre-processing technique. Our model is then applied to estimate trends in the extremal dependence that are in good agreement with empirical evidence. Finally, the fitted model is used to estimate joint extreme events up to the year 2080, allowing us to analyse of the impact of climate change on such events.

E1973: **An improved method for extreme sea level estimation**
Presenter: **Eleanor DArcy**, Lancaster University, United Kingdom
Co-authors: Jonathan Tawn

Storm surges, combined with high tide, pose an increasing risk to coastline communities. To reduce their impact, accurate return level estimates are required to provide information for coastal defence engineering. Early methods modelled sea levels directly, but this ignores the known tidal

component and results were biased due to stationary assumptions being violated. Instead, we filter out waves and remove the mean sea level trend, to consider peak tide and skew surge as the only components of sea levels. Skew surges are stochastic and define the difference between the peak tide and maximum observed sea level within a tidal cycle. They are driven meteorologically, so they are more extreme in winter. Methods currently used in practice make several restrictive and unrealistic assumptions; our approach corrects these. We model extreme skew surges using the GPD. We capture seasonality, longer-term trends and skew surge-peak tide dependence through daily, yearly and tidal covariates in the scale and rate parameters. We also account for skew surge temporal dependence using a Gaussian copula, assuming the series follows a Markov process. Since peak tides are predictable, we choose tidal samples to reflect monthly and interannual variations. To derive a distribution for the annual maximum sea levels, we combine the distributions of the skew surge and peak tide. Our return level estimates are more accurate than those currently used in the UK for coastal flood defence design.

EO342 Room S0.03 SPATIAL DATA SCIENCE Chair: Philipp Otto

E0556: **A Bayesian perspective on spatial+**
Presenter: **Isa Marques**, University of Goettingen, Germany
Co-authors: Paul Wiemann

Spatial models are used in a variety of research areas, such as environmental sciences, epidemiology, or physics. A common phenomenon in such spatial regression models is spatial confounding. This phenomenon is observed when spatially indexed covariates modeling the mean of the response are correlated with a spatial random effect included in the model, for example, as a proxy of unobserved spatial confounders. Several solutions to spatial confounding have been brought forward, including the method Spatial+ (Dupont et al., 2021). Spatial+ is based on a two-stage frequentist model. One notably absent point is the consideration of the additional uncertainty arising from the first stage estimation determining the residuals. Incorporating uncertainty can be achieved via a structural equation model. While similar point estimates will result from either method, a structural equation model has the distinct advantage of integrating all steps in one joint optimization problem that can be easily embedded in a Bayesian framework. We evaluate the performance of Spatial+ in a Bayesian framework, with special attention to uncertainty propagation. Both simulated and real datasets are considered.

E0987: **ARPALData: An R package to retrieve and analyze air quality and weather data for Lombardy**
Presenter: **Paolo Maranzano**, University of Milano-Bicocca & Fondazione Eni Enrico Mattei, Italy

ARPALData is presented: an R package developed to retrieve, manage and analyze air quality and meteorological data from the open database of the Regional Environmental Protection Agency (ARPA Lombardia) of Lombardy, Italy. ARPALData addresses several issues: (1) data and metadata provided by ARPA are entirely in the Italian language, thus excluding all non-Italian users; (2) direct collection of data from the agency's portal requires heavy data manipulation; (3) air quality and climate conditions in Lombardy are continuously raising considerable interest from researchers and technicians involved in policy evaluation. *ARPALData* package provides to the users seventeen functions, which can be re-grouped into three different categories according to their goal: 1) download, 2) classes check, and 3) data analysis and representation. The download functions aim at downloading and formatting the observations according to several filtering inputs provided by the user. The format-checking functions aim to verify that the input objects belong to appropriate object classes valid for applying the analysis functions. Eventually, the analysis and representation functions allow for the computation of descriptive statistics and graphical representation of the input data. The software (release 1.2.3) has been freely available on the R CRAN since April 2022.

E1100: **Developing spatial multi-resolution models for forestry data with Liesel**
Presenter: **Paul Wiemann**, Texas A&M University, United States
Co-authors: Isa Marques, Thomas Kneib

Liesel is a software framework for developing and estimating Bayesian models. Compared to many popular probabilistic programming languages like Stan, Liesel offers full control of the MCMC algorithm. Moreover, when a computationally efficient formulation of the model is available, it can be easily implemented in Liesel. We use these features of Liesel when developing spatial multi-resolution models for data with a specific structure often found in forestry. Here, data is collected intensively in several rather small-sized plots. The plots, however, are distant, and no observations are made between plots. The different intensities in different areas can yield single-resolution stationary spatial models – these assume the same dependence structure over the whole space – inappropriate. We present a Bayesian spatial multi-resolution approach that models separately local and global processes. Exploiting the specific structure allows us to model Gaussian random fields with full covariance matrices while keeping the model still computationally feasible. We outline the computational implementation details and compare the performance of the approach presented using simulated and real data.

EO260 Room S0.11 BAYESIAN SEMI- AND NON-PARAMETRIC METHODS III Chair: Andres Barrientos

E1125: **Sparse spatial random graphs**
Presenter: **Francesca Panero**, London School of Economics, United Kingdom
Co-authors: Francois Caron, Judith Rousseau

A model is presented to describe spatial random graphs, exploiting the so-called "graphex" setting embedded in a Bayesian nonparametric framework, that allows for flexibility and interpretable parameters. We provide a number of asymptotic results, namely that the model is able to describe both sparse and dense networks (with various levels of sparsity), is equipped with positive global and local clustering coefficients and can have a power-law degree distribution whose exponent is easily tuned. We offer a way to perform posterior inference through an MCMC algorithm. We show the results of the estimation obtained on simulated and real data from airport connections. Finally, we discuss how our proposal relates to other spatial network models in the literature.

E0985: **A semiparametric Bayesian model for biclustering**
Presenter: **Alejandro Murua**, University of Montreal, Canada
Co-authors: Fernando Quintana

Motivated by classes of problems frequently encountered in the analysis of gene expression data, we propose a semiparametric Bayesian model to detect biclusters, that is, subsets of individuals sharing similar patterns over a set of conditions. Our approach is based on the well-known plaid model. By assuming a truncated stick-breaking prior we also find the number of biclusters present in the data as part of the inference, thus freeing the traditional plaid model from the restriction of a predefined number of biclusters. The model also introduces a penalty prior that controls the size of biclusters. Evidence from a simulation study shows that the model is capable of correctly detecting biclusters and performs well compared to some competing approaches. The flexibility of the proposed prior is demonstrated with applications to the analysis of gene expression data (continuous responses) and histone modifications data (count responses).

E1021: **A class of random Bernstein copula models**
Presenter: **Alejandro Jara**, Pontificia Universidad Catolica de Chile, Chile

Copula models provide great flexibility in modeling relationships between random variables. For inference to take full advantage of this flexibility, one needs appropriately rich families of copula functions, capable of approximating any copula. One such family is the family of Bernstein copulas, which are a variety of multivariate Bernstein polynomial, and which has been shown to be dense in the space of continuous copula functions. Bernstein copulas have been used for inference before, but only using likelihood-free approximation methods. We observe a fact about

the geometry of the parameter space of Bernstein copulas, and note that it is closely related to a different class of copula known as grid uniform copulas. Based on this relationship, we propose a Bayesian model based on Bernstein copulas and an automatic MCMC algorithm capable of performing full posterior inference on the copula and marginal distributions.

EO626 Room S0.12 RECENT DEVELOPMENTS IN ROBUST MODEL SELECTION Chair: Luca Insolia

E1493: **Robust variable selection and estimation via adaptive elastic net S-estimators for linear regression**
Presenter: **David Kepplinger**, George Mason University, United States

Heavy-tailed error distributions and predictors with anomalous values are ubiquitous in high-dimensional regression problems and can seriously jeopardize the validity of statistical analyses if not properly addressed. For more reliable estimation under these adverse conditions, we propose a new robust regularized estimator for simultaneous variable selection and coefficient estimation. This estimator, called adaptive PENSE, possesses the oracle property without prior knowledge of the scale of the residuals and without any moment conditions on the error distribution. The proposed estimator gives reliable results even under very heavy-tailed error distributions and aberrant contamination in the predictors or residuals. Importantly, even in these challenging settings, variable selection by adaptive PENSE remains stable. Numerical studies on simulated and real data sets highlight improved finite-sample performance in many settings compared to other robust regularized estimators in the case of contaminated samples and competitiveness compared to classical regularized estimators in clean samples.

E0400: **Robust thin plate splines for multivariate spatial smoothing**
Presenter: **Ioannis Kalogridis**, KU Leuven, Belgium

A novel family of multivariate robust smoother is proposed based on the thin-plate (Sobolev) penalty that is particularly suitable for the analysis of spatial data. The proposed family of estimators can be expediently computed even in high dimensions, is invariant with respect to rigid transformations of the coordinate axes and can be shown to possess optimal theoretical properties under mild assumptions. The competitive performance of the proposed thin-plate spline estimators relative to their non-robust counterparts is illustrated in a simulation study and a real data example involving two-dimensional geographical data on ozone concentration.

E0627: **Robust variable selection in semiparametric regression modeling**
Presenter: **Seo-Young Park**, Sungkyunkwan University, Korea, South
Co-authors: Byungtae Seo

The penalized least squares and maximum likelihood methods have been successfully employed for simultaneous parameter estimation and variable selection. However, outlying observations can severely affect the quality of the estimator and selection performance. Although some robust methods for variable selection have been proposed in the literature, they often lose substantial efficiency. This is because the tool to gain robustness depends excessively on choosing additional tuning parameters or modifying the original objective functions. To alleviate such issues, we propose a penalized maximum likelihood method using a nonparametric Gaussian scale mixture distribution. We demonstrate that the proposed estimator has desirable theoretical properties, including sparsity and oracle properties. For the estimation, we alternatively exploit expectation-maximization and gradient-based algorithms for the parametric and nonparametric components, respectively. We also demonstrate the performance of the proposed method through numerical studies, including simulation studies and real data analysis.

EO502 Room S0.13 STATISTICS OF HIGH-FREQUENCY DATA II Chair: Carsten Chong

E0464: **Estimation of path dependent functionals under high frequency sampling**
Presenter: **Mark Podolskij**, University of Luxembourg, Luxembourg

Some recent results are presented about the (optimal) estimation of certain path-dependent functionals of Levy processes under high-frequency sampling. We mainly consider supremum, local time and occupation time measures. The focus is on the construction of estimators and the corresponding limit theorems. Furthermore, we will discuss how these estimates can be applied in practice.

E0619: **Short-time expansion of characteristic functions in a rough volatility setting with applications**
Presenter: **Viktor Todorov**, Northwestern University, United States
Co-authors: Carsten Chong, Viktor Todorov

The aim is to derive a higher-order asymptotic expansion of the conditional characteristic function of the increment of an Ito semimartingale over a shrinking time interval. The spot characteristics of the Ito semimartingale are allowed to have dynamics of general form. In particular, their paths can be rough; that is, they exhibit local behavior like that of a fractional Brownian motion, while at the same time have jumps with an arbitrary degree of activity. The expansion result shows the distinct roles played by the different features of the spot characteristics dynamics. As an application of our result, we construct a nonparametric estimator of the Hurst parameter of the diffusive volatility process from portfolios of short-dated options written on an underlying asset.

E0200: **When frictions are fractional: Rough noise in high-frequency data**
Presenter: **Carsten Chong**, Columbia University, United States

The analysis of high-frequency financial data is often impeded by the presence of noise. The motivation comes from intraday transactions data in which market microstructure noise appears to be rough, that is, best captured by a continuous-time stochastic process that locally behaves as fractional Brownian motion. Assuming that the underlying efficient price process follows a continuous Ito semimartingale, we derive consistent estimators and asymptotic confidence intervals for the roughness parameter of the noise and the integrated price and noise volatilities, in all cases where these quantities are identifiable. In addition to desirable features such as serial dependence of increments, compatibility between different sampling frequencies and diurnal effects, the rough noise model can further explain divergence rates in volatility signature plots that vary considerably over time and between assets.

EO408 Room Virtual R01 LONGITUDINAL DATA ANALYSIS Chair: Orla Murphy

E1881: **Clustering multivariate longitudinal data using matrix-variate mixture models**
Presenter: **Paul McNicholas**, McMaster University, Canada

Some model-based clustering approaches for multivariate longitudinal data are discussed. These include approaches that do not assume normality. The approaches are illustrated using real and simulated data.

E1691: **Extracting dynamic features from irregularly spaced time series**
Presenter: **Oisin Ryan**, Utrecht University, Netherlands

The advent of smartphones and wearable technology has seen an explosion in research designs which involve the collection and analysis of time-series data. In time-series analysis, tools such as autocorrelation and cross-correlation functions provide the basis for understanding the dynamics underlying this data. However, the estimation of auto- and cross-correlations typically relies on the assumption that data are equally spaced in time, and this assumption is often violated in practice: For example, in social science settings, self-report measures collected through experience sampling designs often result in high irregularly spaced measurements, either by design or due to missing measurement waves. We develop and present a statistical tool, available as an R package, which allows for the estimation of auto and cross-correlations from irregularly spaced time series. Based on generalized additive models, we assess the performance of this method in comparison to both traditional approaches and confirmatory fitting of

continuous-time models, the latter of which is vulnerable to problems of model misspecification and unobserved confounding, which the presented method avoids.

E1816: **Finite mixture models for longitudinal data with dynamic group membership**
Presenter: **Jeffrey Andrews**, University of British Columbia Okanagan, Canada
Co-authors: Liam Welsh, Ryan Browne
A compositional approach is introduced for the building and fitting of a finite Gaussian mixture model, permitting highly constrained components to be added to the model at a very low cost with respect to growth in free parameters. The explicit goal of this approach is to enable both the detection and modelling of small numbers of observations which change groups over time in longitudinal data — all under a fully unsupervised paradigm. The proposed approach can be considered an alternative to others in the literature which rely on hidden Markov models to achieve a similar effect. We provide both simulations and real data applications for illustrative purposes.

EO336 Room Virtual R02 BAYESIAN CONTRIBUTIONS TO SURVEY METHODOLOGY Chair: Brenda Betancourt

E1175: **Private tabular survey data products through synthetic microdata generation**
Presenter: **Jingchen Hu**, Vassar College, United States
Co-authors: Terrance Savitsky, Matthew Williams
Two synthetic microdata approaches are proposed to generate private tabular survey data products for public release. We adapt a pseudo posterior mechanism that downweights by-record likelihood contributions with weights in [0,1] based on their identification disclosure risks to producing tabular products for survey data. Our method applied to an observed survey database achieves an asymptotic global probabilistic differential privacy guarantee. Our two approaches synthesize the observed sample distribution of the outcome and survey weights, jointly, such that both quantities together possess a privacy guarantee. The privacy-protected outcome and survey weights are used to construct tabular cell estimates (where the cell inclusion indicators are treated as known and public) and associated standard errors to correct for survey sampling bias. Through a real data application to the Survey of Doctorate Recipients public use file and simulation studies motivated by the application, we demonstrate that our two microdata synthesis approaches to construct tabular products provide superior utility preservation as compared to the additive noise approach of the Laplace Mechanism. Moreover, our approaches allow the release of microdata to the public, enabling additional analyses at no extra privacy cost.

E1373: **Functional and structural measurement error models with global-local priors for random effects in small area estimation**
Presenter: **Xueying Tang**, University of Arizona, United States
Co-authors: Jairo Fuquene
Small area estimation (SAE) plays an important role in producing economic and public health indicators in statistical offices of developing countries. A common method of SAE utilizes the Fay-Herriot model to estimate small area means based on the direct estimates from surveys with the help of auxiliary covariates from administrative records and/or population census. Due to the long intercensal periods and the incompleteness of administrative records in developing countries, finding covariates for small area estimation is often challenging, and estimates from another survey are used as an alternative. This calls for models that account for the measurement errors of the auxiliary variables. Existing measurement error models often assume homogeneous variance for random effects across small areas. However, it has been shown that this assumption is often invalid and that taking into account the heterogeneity of random effects could increase the accuracy of SAE. We use global-local priors for the random effects in the functional and structural measurement error models to accommodate heterogeneous random effect variance. We theoretically examine the behavior of the posterior mean estimator under the proposed models. MCMC algorithms are designed to obtain posterior samples of the model parameters. We demonstrate the performance of the proposed models through a simulation study and a case study of estimating public health indicators in the municipalities of Colombia.

E1450: **A case study for subnational population estimates using a population base statistical register**
Presenter: **Jairo Fuquene**, UC Davis, United States
Population projections at subnational levels play an important role in decision-making in low-income countries. We consider a Bayesian approach with a population base statistical register to update population projections at subnational levels. We consider, for the first time to the best of our knowledge, the construction of a population base statistical register in a developing country using administrative records from the health, education and vital statistics systems and other specific but also important administrative records with information about the victims of the current armed conflict and the tax registration in Colombia. Our proposal is motivated by the need to produce more realistic population projections at sub-national levels in Colombia, where surveys and censuses are difficult to implement because of the current armed conflict in this country. Then we consider our proposal to produce population projections by age and sex groups in a municipality highly affected by forced displacement. To illustrate that the proposed procedure produces population projections with large precision, we consider an experimental population census conducted exclusively in this municipality.

EO553 Room Virtual R03 FEDERATE LEARNING AND DATA PRIVACY IN MODERN DATA ANALYSIS Chair: Xiwei Tang

E1579: **Interval privacy: A new framework for privacy-preserving data collection**
Presenter: **Jie Ding**, University of Minnesota, United States
The emerging public awareness and government regulations of data privacy motivate new paradigms of collecting and analyzing data transparent and acceptable to data owners. A new concept of privacy and related data formats, mechanisms, and theories for statistically privatizing data during data collection are introduced. The new privacy mechanisms will record each data value as a random interval (or, more generally, a range) containing it. Such mechanisms can be easily deployed through survey-based data collection interfaces, e.g., by asking a respondent whether their data value is within a randomly generated range. Using narrowed range to convey information is complementary to the popular paradigm of perturbing data. Also, the proposed mechanisms can generate progressively refined information at the discretion of individuals, naturally leading to privacy-adaptive data collection. Unique perspectives will be demonstrated, which are brought by Interval Privacy for human-centric data privacy, where individuals enjoy a perceptible, transparent, and simple way of sharing sensitive data.

E1616: **Distribution-invariant differential privacy**
Presenter: **Xuan Bi**, University of Minnesota, United States
Differential privacy is becoming one gold standard for protecting the privacy of publicly shared data. It has been widely used in biomedical sciences, data science, public health, information technology, and the U.S. decennial census. Nevertheless, to guarantee differential privacy, existing methods may unavoidably alter the conclusion of the original data analysis, as privatization often changes the sample distribution. This phenomenon is known as the trade-off between privacy protection and statistical accuracy. We mitigate this trade-off by developing a distribution-invariant privatization (DIP) method to reconcile both high statistical accuracy and strict differential privacy. As a result, any downstream statistical or machine learning task yields essentially the same conclusion as if one used the original data. Numerically, under the same strictness of privacy protection, DIP achieves superior statistical accuracy across a wide range of simulation studies and real-world benchmarks.

E1694: **A tree-based model averaging approach for personalized treatment effect estimation from heterogeneous data sources**
Presenter: **Lu Tang**, University of Pittsburgh, United States
Co-authors: Xiaoqing Tan, Chung-Chou Ho Chang, Ling Zhou

Accurately estimating personalized treatment effects within a study site (e.g., a hospital) has been challenging due to the limited sample size. Furthermore, privacy considerations and a lack of resources prevent a site from leveraging subject-level data from other sites. We propose a tree-based model averaging approach to improve the estimation accuracy of conditional average treatment effects (CATE) at a target site by leveraging models derived from other potentially heterogeneous sites, without them sharing subject-level data. To our best knowledge, there is no established model averaging approach for distributed data with a focus on improving the estimation of treatment effects. Specifically, under distributed data networks, our framework provides an interpretable tree-based ensemble of CATE estimators that joins models across study sites, while actively modeling the heterogeneity in data sources through site partitioning. The performance of this approach is demonstrated by a real-world study of the causal effects of oxygen therapy on hospital survival rates and backed up by comprehensive simulation results.

EO368 Room Virtual R04 BIOSTATISTICS IN RENAL RESEARCH Chair: Ivonne Solis-Trapala

E0789: **Graphical and multistate modelling to explore factors influencing home dialysis uptake**

Presenter: **Jessica Potts**, Keele University, United Kingdom

Co-authors: Camille Parsons, Kerry Allen, Sarah Damery, Lisa Dikomitis, James Fotheringham, Harry Hill, Mark Lambie, Louise Phillips-Darby, Iestyn Williams, Simon Davies, Ivonne Solis-Trapala

Renal replacement therapy (RRT) takes the form of home or in-centre dialysis or kidney transplantation. Home dialysis provides increased control and freedom for patients, especially in those in employment or wishing to travel. However, the use of home dialysis varies considerably across the UK and is decreasing despite attempts to encourage greater use. The Intervening to eliminate the centre-effect variation in home dialysis use (Inter-CEPt) study uses a mixed-methods approach to explore the factors driving access to home therapies to develop a cost-effective intervention. A qualitative analysis informed the design of a national survey of renal centres and the development of a chain graph model to describe the complex interrelations among patient- and centre-level factors leading to the uptake of home dialysis, based on the survey data linked to the UK Renal Registry (UKRR) which collates patient-level data from renal centres across the UK. Multistate models were developed to estimate the rates of transition from and to home, in-centre dialysis, and transplantation, and to death informed by the graphical model, for health economics analysis. We will discuss initial statistical analysis results highlighting the challenges of working with real-world data.

E1026: **Prognostic score-based methods for estimating center effects based on survival probability**

Presenter: **Douglas Schaubel**, University of Pennsylvania Perelman School of Medicine, United States

Co-authors: Youjin Lee

In evaluating the performance of facilities or centers on survival outcomes, the standardized mortality ratio (SMR), which compares the observed to expected mortality, has been widely used, particularly in the evaluation of kidney transplant centers. Despite its utility, the SMR may exaggerate center effects in settings where survival probability is relatively high. An example is one-year graft survival among U.S. kidney transplant recipients. We propose a novel approach to estimate center effects in terms of differences in survival probability (i.e., each center versus a reference population). An essential component of the method is a prognostic score weighting technique, which permits accurately evaluating centers without necessarily specifying a correct survival model. Advantages of our approach over existing facility-profiling methods include a metric based on survival probability (greater clinical relevance than ratios of counts/rates); direct standardization (valid to compare between centers, unlike indirect standardization-based methods, such as the SMR); and less reliance on correct model specification (since the assumed model is used to generate risk classes as opposed to fitted-value based expected counts). We establish the asymptotic properties of the proposed weighted estimator and evaluate its finite-sample performance under a diverse set of simulation settings. The method is then applied to evaluate U.S. kidney transplant centers.

E1414: **Modelling chronic kidney disease: the GLOMMS2 cohort**

Presenter: **Gordon Prescott**, University of Central Lancashire, United Kingdom

Co-authors: Simon Sawhney, Corri Black, Angharad Marks, Nick Fluck, Laura Clark, Adeera Levin

The Grampian Laboratory Outcomes Morbidity and Mortality Study II (GLOMMS2) is a population cohort of 70,000 adults in a single UK health authority in North-East Scotland (population 440,000). It links national and regional data sources on kidney function, comorbidities and major healthcare events (hospital episodes, renal replacement therapy and death) over time. Uniquely, all biochemistry data are processed by a single laboratory service regardless of clinical location, minimising the loss of baseline and follow-up data. The cohort includes around 20,000 with chronic kidney disease (CKD) and 20,000 with normal kidney function. Research by Aberdeen Applied Renal Research Collaboration has included the prediction of mortality, the need for renal replacement therapy within five years, renal progression, and the intermediate and 10-year prognosis of those who suffer episodes of acute kidney injury.GLOMMS2 is one of 34 multinational cohorts contributing to the CKD Prognosis Consortium. Data from more than 5 million individuals have been used to develop risk models for multiple renal outcomes. Health cohorts may be structured around regular health check data, interactions at times of poor health, or a mixture. An overview of research on the GLOMMS2 cohort will be presented. The methodological complications of combining cohorts will be discussed.

EO548 Room Virtual R05 CAUSAL MACHINE LEARNING Chair: Amir Asiaee

E1452: **Omitted variable bias in causal machine learning**

Presenter: **Carlos Cinelli**, University of Washington, United States

Co-authors: Victor Chernozhukov, Whitney Newey, Amit Sharma, Vasilis Syrgkanis

The aim is to derive general, yet simple, sharp bounds on the size of the omitted variable bias for a broad class of causal parameters that can be identified as linear functionals of the conditional expectation function of the outcome. Such functionals encompass many traditional research targets in causal inference studies, such as, for example, (weighted) average of potential outcomes, average treatment effects (including subgroup effects, such as the effect on the treated), (weighted) average derivatives, and policy effects from shifts in covariate distribution – all for general, nonparametric causal models. Our construction relies on the Riesz-Frechet representation of the target functional. Specifically, we show how the bound on the bias depends only on the additional variation that the latent variables create both in the outcome and in the Riesz representer for the parameter of interest. Moreover, in many important cases (e.g., average treatment effects and average derivatives), the bound is shown to depend on easily interpretable quantities that measure the explanatory power of the omitted variables. Therefore, simple plausibility judgments on the maximum explanatory power of omitted variables are sufficient to place overall bounds on the size of the bias. Furthermore, we use debiased machine learning to provide flexible and efficient statistical inference on learnable components of the bounds. Finally, empirical examples demonstrate the usefulness of the approach.

E1764: **Methodological advances in causal representation learning**

Presenter: **Kun Zhang**, CMU, United States

Causal representation learning aims to reveal the underlying high-level hidden causal variables and their relations. It can be seen as a special case of causal discovery, whose goal is to recover the underlying causal structure or causal model from observational data. The modularity property of a causal system implies properties of minimal changes and independent changes of causal representations, and how such properties make it possible to recover the underlying causal representations from observational data with identifiability guarantees: under appropriate assumptions, the learned representations are consistent with the underlying causal process. Various problem settings are considered involving independent and identically

distributed (i.i.d.) data, temporal data, or data with distribution shift as input, and demonstrate when identifiable causal representation learning can benefit from the flexibility of deep learning and when it has to impose suitable parametric assumptions on the causal process.

E1976: **Identifying HIV sequences that escape antibody neutralization using random forests and collaborative targeted learning**
Presenter: **David Benkeser**, Emory University, United States

Recent studies have indicated that it is possible to protect individuals from HIV infection using a passive infusion of monoclonal antibodies. However, in order for monoclonal antibodies to confer robust protection, the antibodies must be capable of neutralizing many possible strains of the virus. This is particularly challenging in the context of a highly diverse pathogen like HIV. It is, therefore, of great interest to leverage existing observational data sources to discover antibodies that are able to neutralize HIV viruses via residues where existing antibodies show modest protection. These observational data include genetic features of many diverse HIV genetic sequences, as well as in vitro measures of antibody resistance. We propose methods to analyze these data to identify important genetic features using the outcome-adaptive, collaborative targeted minimum loss-based estimation (CTMLE) approach using random forests. We demonstrate via simulation that the approach enjoys statistical benefits over existing approaches and apply the approach to the Compile, Analyze and Tally Nab Panels (CATNAP) database to identify AA positions that are potentially causally related to resistance to neutralization by several different antibodies.

EO376 Room Virtual R07 NEW STATISTICAL ADVANCES IN THE ANALYSIS OF WEARABLE DEVICE DATA Chair: Linda Valeri

E0232: **Causal identification of dynamic effects in non-stationary time series from N-of-1 observational mobile health data**
Presenter: **Xiaoxuan Cai**, Columbia University, United States

Mobile technology (e.g., mobile phones and wearable devices) enables unprecedented monitoring of social interactions, symptoms, and other health and behavioral conditions among individuals. Continuous monitoring of personal data generates multivariate time series data of outcomes, exposures, and confounding variables in an N-or-1 study. Popular methods for univariate time series or longitudinal data assume stationary time series or time-invariant treatment effects, which fail to capture the dynamic treatment effect in both short- and long-term non-stationary time series. We propose a set of causal estimands for non-stationary multivariate time series in N-of-1 studies, in order to systematically summarize time-varying treatment in the short and long term, and demonstrate their identification via the g-formula in the presence of exposure- and outcome-covariate feedbacks. The g-formula employs an innovative state space model to account for the time-varying treatment effects in non-stationary time series in an N-of-1 setting. We demonstrate the estimation of proposed estimands using a smartphone observational study of bipolar patients, the Bipolar Longitudinal Study, in which we examine both the short- and long-term effects of digital social interaction on psychiatric symptoms, as well as how these effects change over time. A novel positivity validation plot for testing the positivity assumption in time series studies is proposed.

E0769: **Smartphone-based activity recognition using movelets**
Presenter: **Emily Huang**, Wake Forest University, United States

Physical activity patterns can provide information about an individual's health profile. Traditionally, data on physical activity levels have been gathered through patient self-report in surveys. These data offer valuable firsthand accounts, but they can suffer from bias due to their subjectivity. In comparison, built-in sensors in the smartphone can measure data objectively and continuously, with less burden to the patient. There is a variety of data analysis approaches for smartphone-based human activity recognition. We applied the movelet method to classify the type of activity performed, based on smartphone accelerometer and gyroscope data. Our results show that this method has the advantages of being interpretable and transparent. A unique aspect of our movelet application is that it extracts information optimally from multiple sensors. Compared to single-sensor applications, our approach jointly incorporates the accelerometer and gyroscope sensors with the movelet method. Our findings show that combining data from the accelerometer and gyroscope can leverage their own distinct strengths, providing more accurate activity recognition than using each sensor alone.

E1970: **Modeling and estimating the effects of cumulative just-in-time treatments using data from micro-randomized trials**
Presenter: **Daniel Almirall**, University of Michigan, United States
Co-authors: Inbal Nahum-Shani, Susan Murphy

Mobile health interventions aim to provide individualized support, whenever, and wherever it is needed. This includes the provision of therapeutic support in (near) real-time, as well as the provision of prompts that support the engagement of users in the mobile health application. Micro-randomized trials (MRT) are used to address scientific questions concerning the construction of mobile health applications of this type. In an MRT, participants may be randomized hundreds of times over the course of the study. Often, MRTs are designed to learn whether–and if so, when and based on what self-report or passive/sensor data–to intervene with a prompt, suggestion, or some other form of just-in-time treatment. An important set of scientific questions for behavioral intervention scientists involves the cumulative effect of just-in-time treatments. These questions are designed to better understand the dose-response effect of accumulated just-in-time treatments, including how these effects vary over the study. A definition for "cumulative just-in-time effects" is introduced in terms of potential outcomes, which is suitable for the data arising from an MRT. It develops a regression approach for estimating these effects. The approach is illustrated using data from an MRT designed to inform the development of a just-in-time adaptive intervention for promoting real-time, real-world engagement in evidence-based self-regulatory smoking cessation strategies.

EO592 Room Virtual R08 HIGH-DIMENSIONAL PROBABILITY AND STATISTICS Chair: Tatyana Krivobokova

E1195: **Doubly debiased lasso: High-dimensional inference under hidden confounding**
Presenter: **Zijian Guo**, Rutgers University, United States
Co-authors: Domagoj Cevid

Inferring causal relationships or related associations from observational data can be invalidated by the existence of hidden confounding. We focus on a high-dimensional linear regression setting, where the measured covariates are affected by hidden confounding and propose the Doubly Debiased Lasso estimator for individual components of the regression coefficient vector. Our advocated method simultaneously corrects both the bias due to the estimation of high-dimensional parameters as well as the bias caused by the hidden confounding. We establish its asymptotic normality and also prove that it is efficient in the Gauss-Markov sense. The validity of our methodology relies on a dense confounding assumption, i.e. that every confounding variable affects many covariates. The finite sample performance is illustrated with an extensive simulation study and a genomic application.

E0787: **Higher rank signatures and filtrations**
Presenter: **Chong Liu**, ShanghaiTech University, China

Filtration is an abstract and important notion that appears naturally in stochastic analysis, which models the information flow generated by underlying stochastic processes. However, many well-known statistical methods cannot detect filtrations as they are based on weak topology. Consequently, they may lead to significant errors in those circumstances where the evolution of information plays a crucial role. We will introduce a new methodology based on the signature kernel learning approach, which can be used to give a precise description of filtrations hidden behind observed signals. We will then illustrate that this method provides a feasible statistical tool for lots of filtration–sensitive cases; in particular, it allows to

reduce highly non–linear path-and-filtration dependent functionals (e.g. the pricing of American option) to a linear regression problem, which reveals an interesting combination of (Hopf) algebra and kernel learning.

E1696: **Speeding up the Laplace approximation method for a high-dimensional Rasch model**
Presenter: **Shuhrah Alghamdi**, University of Glasgow, United Kingdom

The Laplace approximation (LA) method can be a valuable tool for researchers interested in estimating students' abilities using item response theory models. Computational time is an essential criterion for assessing the efficiency of the LA for online inference or massive datasets. The performance of the LA depends on obtaining the covariance matrix Σ by inverting the Hessian matrix H. However, computing the inverse of the H matrix is computationally expensive for high-dimensional problems (i.e. when there are many students). Two methods are discussed for reducing the computational costs of estimating students' abilities using the Rasch model. The first method is to use the idea of the block matrix, according to which the H matrix can be divided into sub-matrices, and linear algebra strategies can subsequently be used to simplify the calculations for inverting the H matrix. The second method approximates the posterior distribution using only the diagonal of the H matrix. These methods do not affect the point estimates, and as the number of students increases, the difference between the full H matrix and the diagonal H matrix becomes negligible. Compared with the standard LA method, our two proposed strategies can reduce the computational time 3- to 15-fold.

EO380 Room K2.31 (Nash Lec. Theatre) RECENT ADVANCES IN MEDIATION ANALYSIS Chair: BaoLuo Sun

E1207: **Hypothesis test for causal mediation of semicompeting risks under copula, frailty and multistate models**
Presenter: **Jih-Chang Yu**, Academia Sinica, Taiwan
Co-authors: Yen-Tsung Huang

Semicompeting risks can be formulated as a mediation model where a direct effect (DE), the effect of an exposure on the terminal outcome not through the intermediate outcome, and an indirect effect (IE), the effect on the terminal outcome mediated by the intermediate outcome. We propose testing procedures to evaluate the DE and IE under three classic semicompeting risks models: Clayton copula model, gamma frailty model and multistate model. We study the correspondence of the DE and IE with the model parameters and establish testing rules for the two effects under the three models. For statistical inference, we use the U-statistic approach for the Clayton copula model and nonparametric maximum likelihood estimation for the multistate and gamma frailty models. The simulation study shows that among the three models, the Clayton copula model attains the best statistical power if the model assumption holds, but has the potential bias caused by model misspecification; the gamma frailty model is the most robust model by sacrificing the efficiency; the multistate model balances the efficiency and robustness. We apply the proposed method to a hepatitis study. The aforementioned models unanimously suggest that both hepatitis B and C lead to a higher incidence of liver cancer by increasing liver cirrhosis incidence.

E1567: **The central role of the mediator process in mediation analysis**
Presenter: **Caleb Miles**, Columbia University, United States

Traditionally, mediation analysis involves analysis of exposure, mediator, and outcome, each observed at sequential discrete points in time. The natural direct and indirect effects are then defined based on these three-time points. Identification relies on the assumption that no effect of the exposure can cause both the mediator and the outcome. However, the mediator of interest will often be a stochastic process varying from baseline to follow-up, and its value observed at an individual point in time but a coarse measurement of this process. When the intermediate variable is a mediator, we will argue that earlier instances of the intermediate variable will often be exposure-induced confounders of the mediator at its observed time. Thus, the mediated effects defined in terms of the coarsened mediator process will not be identified. Further, we will argue that the mediated effects of greatest substantive interest are those involving the full mediator process, and that the coarsened mediator process effects can have nonsensical interpretations. To make progress, one must instead rely on strong exclusion restriction assumptions or account for the full mediator process. Lastly, we will discuss an effect decomposition relating the full mediator process's indirect effect to the coarsened mediator process's indirect effect.

E0191: **Interpretable sensitivity analysis for the Baron-Kenny approach to mediation with unmeasured confounding**
Presenter: **Peng Ding**, University of California, Berkeley, United States

Mediation analysis assesses the extent to which the treatment affects the outcome indirectly through a mediator and the extent to which it operates directly through other pathways. As the most popular method in empirical mediation analysis, the BaronKenny approach estimates the indirect and direct effects of the treatment on the outcome based on linear structural equation models. However, when the treatment and the mediator are not randomized, the estimates may be biased due to unmeasured confounding among the treatment, mediator, and outcome. Building on previous work, we propose a sharp and interpretable sensitivity analysis method for the BaronKenny approach to mediation in the presence of unmeasured confounding. We first generalize their sensitivity analysis method for linear regression to allow for heteroskedasticity and model misspecification. We then apply the general result to develop a sensitivity analysis method for the Baron-Kenny approach. To facilitate interpretation, we must express the sensitivity parameters in terms of the partial R2s that correspond to the natural factorization of the joint distribution of the direct acyclic graph. They measure the proportions of variability explained by unmeasured confounding given the observed covariates. Moreover, we extend the method to deal with multiple mediators, based on a novel matrix version of the partial R2 and a general form of the omitted variable bias formula.

EO607 Room K2.40 RECENT DEVELOPMENTS IN UNLINKED REGRESSION Chair: Fadoua Mohr

E1525: **Permuted and unlinked monotone regression in $\mathbb{R}^d$: An approach based on mixture modeling and optimal transport**
Presenter: **Martin Slawski**, George Mason Univ, United States
Co-authors: Bodhisattva Sen

Suppose the aim is to learn a map between d-dimensional inputs and d-dimensional noisy outputs, without observing (input, output)-pairs, but only separate unordered lists of inputs and outputs. We show that the notion of cyclical monotonicity of the underlying map is sufficient for identification and estimation in the unordered setting. We study restoration of the correct correspondence of (input,output)-pairs ("permutation recovery") and develop a computationally efficient and easy-to-use algorithm for denoising based on the Kiefer-Wolfowitz nonparametric maximum likelihood estimator and techniques from the theory of optimal transport. We provide explicit upper bounds on the associated mean squared denoising error under Gaussian noise. Numerical studies corroborate our theoretical analysis.

E1624: **Linear regression with unmatched data: A deconvolution perspective**
Presenter: **Mona Azadkia**, ETH Zurich, Switzerland
Co-authors: Fadoua Balabdaoui

Consider the regression problem where the response Y in R and the covariate X in R_d for $d = 1$ are unmatched. Under this scenario, we do not have access to pairs of observations from the distribution of (X,Y), but instead, we have separate data sets $\{Yi\}$ and $\{Xj\}$, possibly collected from different sources. We study this problem assuming that the regression function is linear and the noise distribution is known or can be estimated. We introduce an estimator of the regression vector based on deconvolution and demonstrate its consistency and asymptotic normality under an identifiability assumption. In the general case, we show that our estimator (DLSE: Deconvolution Least Squared Estimator) is consistent in terms of an extended l_2 norm. Using this observation, we devise a method for semi-supervised learning, i.e., when we have access to a small sample of matched pairs (Xk,Yk). Several applications with synthetic and real data sets are considered to illustrate the theory.

E1788: **Unlinked monotone regression**
Presenter: **Cecile Durot**, Univ. Paris Nanterre, France
Co-authors: Fadoua Balabdaoui, Charles Doss
The so-called univariate unlinked (sometimes decoupled, or shuffled) regression is considered when the unknown regression curve is monotone. In standard monotone regression, one observes a pair (X,Y) where a response Y is linked to a covariate X through the model $Y = m_0(X) + \varepsilon$, with m_0 the (unknown) monotone regression function and ε the unobserved error (assumed to be independent of X). In the unlinked regression setting, one only observes a vector of realizations from both the response Y and the covariate X where now, it is only known that Y has the same distribution as $m_0(X) + \varepsilon$. There is no (observed) pairing of X and Y. Despite this, it is actually still possible to derive a consistent non-parametric estimator of m_0 under the assumption of monotonicity of m_0 and knowledge of the distribution of the noise ε. We establish an upper bound on the rate of convergence of such an estimator under minimal assumptions on the distribution of the covariate X.

EO549 Room K2.41 THEORY AND METHODS IN HIGH DIMENSIONAL 'OMIC' DATA Chair: Christopher McKennan

E1120: **Modeling regulatory network topology improves genome-wide analyses of complex human traits**
Presenter: **Xiang Zhu**, The Pennsylvania State University, United States
Genome-wide association studies (GWAS) in humans have catalogued many significant associations between genetic variants and complex traits. However, most of these findings have unclear biological significance, because they often have small effects and occur in non-coding regions. Integration of GWAS with gene regulatory networks addresses both issues by aggregating weak genetic signals within regulatory programs. We develop a Bayesian hierarchical modeling framework that integrates GWAS summary statistics with gene regulatory networks to infer genetic enrichments and associations simultaneously. We implement the method with an efficient variational inference algorithm that scales well with millions of genetic variants in the human genome. Our method improves upon existing approaches by explicitly modeling network topology to assess enrichments, and by automatically leveraging enrichments to identify associations. Applying this method to 18 human traits and 38 regulatory networks shows that genetic signals of complex traits are often enriched in interconnections specific to trait-relevant cell types or tissues. Prioritizing variants within enriched networks identifies known and previously undescribed trait-associated genes revealing biological and therapeutic insights.

E1362: **Robust and accurate estimation of cellular fractions from tissue omics data via ensemble deconvolution**
Presenter: **Jiebiao Wang**, University of Pittsburgh, United States
Co-authors: Christopher McKennan, Manqi Cai, Wei Chen
Tissue-level omics data such as transcriptomics and epigenomics are average across diverse cell types. To extract cell-type-specific (CTS) signals, dozens of cellular deconvolution methods have been proposed to infer cell-type fractions from tissue-level data. However, these methods produce vastly different results under various real data settings. Simulation-based benchmarking studies showed no universally best deconvolution approaches. There have been attempts at ensemble methods, but they only aggregate multiple single-cell references or reference-free deconvolution methods. To achieve a robust estimation of cellular fractions, we proposed EnsDeconv (Ensemble Deconvolution), which adopts CTS robust regression to synthesize the results from dozens of single deconvolution methods, reference datasets, marker gene selection procedures, data normalizations, and transformations. Unlike most benchmarking studies based on simulations, we compiled four large real datasets of 4937 tissue samples in total with measured cellular fractions and bulk gene expression from different tissue types. Comprehensive evaluations demonstrated that EnsDeconv yields more stable, robust, and accurate fractions than existing methods. We illustrated that EnsDeconv estimated cellular fractions enable various CTS downstream analyses, such as differential fractions associated with clinical variables. To increase generalizability, we further extended EnsDeconv to analyze bulk DNA methylation data.

E1779: **FDR controlled multiple testing for union null hypotheses: A knockoff-based approach**
Presenter: **Ran Dai**, University of Nebraska Medical Center, United States
Co-authors: Cheng Zheng
False discovery rate (FDR) controlling procedures provide important statistical guarantees for replicability in signal identification based on multiple hypotheses testing. In many fields, FDR controlling procedures are used in high-dimensional (HD) analyses to discover features that are truly associated with the outcome. In some recent applications, data on the same set of candidate features are independently collected in multiple different studies. For example, gene expression data are collected at different facilities and with different cohorts, to identify the genetic biomarkers of multiple types of cancers. These studies provide us opportunities to identify signals by considering information from different sources (with potential heterogeneity) jointly. The focus is on providing FDR control guarantees for the tests of union null hypotheses of conditional independence. We present a knockoff-based variable selection method (Simultaneous knockoffs) to identify mutual signals from multiple independent data sets, providing exact FDR control guarantees under finite sample settings. This method can work with very general model settings and test statistics. We demonstrate the performance of this method with extensive numerical studies and real data examples.

E2037: **Tensor decomposition of longitudinal microbiome data**
Presenter: **Siyuan Ma**, Vanderbilt University Medical Center, United States
Co-authors: Hongzhe Li
A few methods available for unsupervised dimension reduction of longitudinal microbial abundance observations are discussed. Existing ones do not fully observe the distribution characteristics of such data types, namely, zero-inflation, compositionality, and overdispersion. We present a tensor decomposition model for dimension reduction of longitudinal microbiome data, by generalizing existing approaches in Gaussian data. Optimization is performed through projected gradient descent, additionally allowing interpretability constraints. Simulation studies show our method can recover low-rank structures in microbiome time course better than existing approaches. We applied our method to two existing longitudinal microbiome studies, to detect global microbial changes associated with dietary and pharmaceutical effects, as well as infant birth modes.

EC806 Room K0.50 DESIGN OF EXPERIMENTS Chair: Kalliopi Mylona

E0430: **Therapist variation within randomised trials of psychotherapy: A design of experiments perspective**
Presenter: **Steven Gilmour**, KCL, United Kingdom
Co-authors: Rebecca Walwyn
Clinical trials of psychiatric interventions present some particular challenges. One of these is that the therapist and/or other healthcare professional is an inseparable part of the treatment. Considering it as a treatment factor leads to a number of complex and non-standard factorial treatment structures. Often it is not possible for each therapist to implement each intervention of interest, so we get a nested, rather than crossed, treatment structure. Sometimes more than one type of healthcare professional is involved, when one can be crossed with interventions, but the other has to be nested. Very often, therapists can only work in one centre, so we get an unusual type of incomplete block design. In addition to this, it is often desirable to model the therapist effects as random, since the therapists in the trial can be considered a sample of a bigger population of therapists. A further complication is that it is often not possible to randomise therapists to centres. We use the principles of the design of experiments and, in particular, the link between randomisation and the derived model to untangle these different complicated structures. This has the advantage of providing models which can be justified by the trial design, which either confirms or improves on existing practice, which uses models without any clear theoretical justification.

E0434: **Optimal designs for testing pairwise differences: A game theoretic approach**
Presenter: **Arpan Singh**, Indian Institute of Technology, Hyderabad, India, India
Co-authors: Satya Prakash Singh, Ori Davidov

In a variety of experimental setups there is an interest in comparing a subset of pairs–of–treatments. Such experiments usually address one of the following two scientific questions: (1) is there a difference within any of the selected pairs of treatments? or, (2) is there a difference within all of the selected pairs of treatments? We propose optimal designs for testing the above mentioned hypotheses using a graph based game theoretic approach.

E0541: **K-optimal designs for parameters of shifted Ornstein-Uhlenbeck processes and sheets**
Presenter: **Sandor Baran**, University of Debrecen, Hungary

Continuous random processes and fields are regularly applied to model temporal or spatial phenomena in many different fields of science, and model fitting is usually done with the help of data obtained by observing the given process at various time points or spatial locations. In these practical applications sampling designs which are optimal in some sense are of great importance. We investigate the properties of the K-optimal design for temporal and spatial linear regression models driven by Ornstein-Uhlenbeck processes and sheets, respectively, and highlight the differences compared with the classical D-optimal sampling. We study the problems of the existence of K-optimal designs and also investigate the dependence of the two designs on the covariance parameters of the driving processes. This information may be crucial for an experimenter in order to increase efficiency in practical situations. Finally, we present a simulation study displaying the superiority of the K-optimal design for large parameter values of the driving random process.

CO078 Room BH (S) 1.01 Lecture Theatre 1 FINANCIAL TIME SERIES Chair: Jean-Michel Zakoian

C0909: **Optimal estimating function for weak location-scale dynamic models**
Presenter: **Christian Francq**, CREST and University Lille III, France
Co-authors: Jean-Michel Zakoian

Estimating functions provide a very general framework for the statistical inference of dynamic models under weak assumptions. We consider a class of time series modelsconsisting in the parametrization of the first two conditional moments which—by contrast with classical location scale dynamic models—do not impose further constraints on the conditional distribution/moments. Quasi-Likelihood Estimators (QLE) are obtained by solving estimating equations deduced from those two conditional moments. Conditions ensuring the existence and asymptotic properties (consistency and asymptotic normality) of such estimators are provided. We pay special attention to the optimal QLE in Godambe's sense. The particular case of the Quasi-Maximum Likelihood Estimators (QMLE) is considered. For pure location models, a data-driven procedure for optimally choosing the QLE is proposed. Our results are illustrated via Monte Carlo experiments and real financial data.

C0996: **Empirical asset pricing with score-driven conditional betas**
Presenter: **Julien Royer**, CREST, France
Co-authors: Thomas Giroux

A novel empirical asset pricing framework is introduced based on the newly introduced score-driven conditional betas model. We extend the theory of the studied conditional betas by establishing the asymptotic distribution of standard test statistics for parameter constancy in conditional regression. In particular, these tests allow for assessing the significance of a given factor in the regression. We then propose a two-step estimation procedure to recover time-varying factor risk premia from individual stock returns. We illustrate the performance of our tests and risk premia estimation procedure on simulations. Finally, we present an application where we assess the existence of a time-varying risk premium associated with a carbon risk factor in the cross-section of US industry portfolios.

C1552: **Strict stationarity and existence of moments for a family of functional GARCHs**
Presenter: **Baye Matar Kandji**, CREST, Institut Polytechnique de Paris, France

Random coefficient autoregressive models are considered with non-negative coefficients in a Banach lattice. We develop a method using functional analysis tools to establish a necessary and sufficient condition for the existence of weak-order and s-strong-order stationary solution, where $s > 0$. We apply these results to provide necessary and sufficient conditions for the existence of a stationary solution for a large family of GARCH processes, including functional GARCH models.

CO194 Room BH (SE) 1.01 RECENT DEVELOPMENTS IN MODELLING AND FORECASTING EXTREMES Chair: Ekaterina Kazak

C0799: **Which hedge funds are systemically risky, and when: A dynamic extreme value regression approach**
Presenter: **Philippe Hubner**, HEC Liege, University of Liege, Belgium
Co-authors: Julien Hambuckers

A novel approach is introduced to measure the time-varying systemic risk contribution of hedge funds at the fund level, overcoming short reporting periods in commercial databases. To do so, we extend the extreme value systemic risk model to a regression context, where marginal tail indices of hedge funds and banks are driven by a set of covariates. This formulation makes it possible to estimate systemic risk contributions by exploiting extreme value regression methods on pooled time series of hedge funds returns - in spite of the short reporting period of the funds. It also has the advantage of identifying whether a high level of systemic risk of a given fund originates from a high risk of spillovers to the banking sector, or the high level of the fund tail risk. These measures are then used to identify funds characteristics and market conditions that indicate a high systemic threat, an information of interest for regulators. Using a large sample of funds over the period 1994-2021, we find that investment strategies are clear determinants of the hedge funds' systemic risk.

C0757: **Nonparametric value-at-risk via sieve estimation**
Presenter: **Philipp Ratz**, Universite du Quebec a Montreal (UQAM), Canada

Artificial Neural Networks (ANN) have been employed for a range of modelling and prediction tasks using financial data. However, evidence of their predictive performance, especially for time-series data, has been mixed. Whereas some applications find that ANNs provide better forecasts than more traditional estimation techniques, others find that they barely outperform basic benchmarks. The aim is to guide as to when the use of ANNs might result in better results in a general setting. We propose a flexible nonparametric model and extend existing theoretical results for the rate of convergence to include the popular Rectified Linear Unit (ReLU) activation function and compare the rate to other nonparametric estimators. Finite sample properties are then studied with the help of Monte-Carlo simulations to provide further guidance. An application to estimate the Value-at-Risk of portfolios of varying sizes is also considered to show the practical implications.

C0371: **Sequentially valid tests for forecast calibration**
Presenter: **Alexander Henzi**, ETH Zurich, Switzerland

Forecasting and forecast evaluation are sequential tasks. Most predictions are issued on a regular basis, such as every hour, day, or quarter, and their accuracy can be monitored continuously. However, standard statistical tools for forecast evaluation are static, in the sense that they require the evaluation period to be fixed in advance, independent of available observations at the time of evaluation. We propose to apply sequential testing methods instead, and develop sequentially valid tests for the calibration of probabilistic forecasts for a real-valued outcome. Our methods are based on e-values, a recently introduced tool for assessing statistical significance, which generalize Wald's sequential probability ratio test. An e-value is

a non-negative random variable with expected value at most one under a null hypothesis. Large e-values give evidence against the null hypothesis, and the multiplicative inverse of an e-value is a conservative p-value. It is demonstrated that the proposed tests can yield useful insights when testing the calibration of probabilistic forecasts in practical applications.

CO098 Room BH (SE) 1.05 MACROECONOMETRICS Chair: Christos Savva

C0194: **Housing prices and inflation expectations**
Presenter: **Christos Savva**, Cyprus University of Technology, Cyprus
Co-authors: Nektarios Michail

Inflation and inflation expectations are some of the most watched macroeconomic variables, given their importance for both social and policy purposes. While usual inflation targets usually stand at two percent, households' inflation perceptions differ substantially. Using a panel VAR approach and data for all euro area countries, the focus is on identifying the main determinants of households' inflation expectations and perceptions. Our findings suggest that house price fluctuations have a primary role in the formation of public inflation expectations and perceptions. Furthermore, interest rate and GDP shock substantially affect both expectations and perceptions, while oil price disturbances appear to have a significant but short-lived impact.

C0197: **Career and non-career jobs: Dangling the carrot**
Presenter: **Demetris Koursaros**, Cyprus University of Technology, Cyprus

A model of the labor market with "career" and "non-career" jobs is developed. Non-career firms have the typical one-type job structure and pay their employees their marginal product, while in career firms, all workers start at a low rank and can be promoted to a higher rank. We show that it is optimal for career firms to compensate more the few workers that get promoted to the higher rank, thereby creating incentives for the mass of the workers in lower ranks who are paid less. We explore the macroeconomic implications of this hierarchical payment structure. We show how our model can provide interesting insights into various puzzles, such as the wage gap between men and women, the cyclicality of the labor wedge and the low volatility of the real wage relative to hours and output along the business cycle, without imposing ad-hoc nominal wage rigidities.

C0539: **Price decomposition and asymmetry in various regimes of the economy**
Presenter: **Nektarios Michail**, Central Bank of Cyprus, Cyprus
Co-authors: Christos Savva, Demetris Koursaros

The price index is decomposed into its two major constituents, markup and marginal cost, via a Markov-switching VAR with fixed transition probabilities. Since the proposed pair of variables has not been extensively analyzed, a theoretical model that derives markups and marginal costs as functions of parameters and shocks is developed to extract identifying restrictions for the VAR. In the empirical exercise, a non-linear representation of GIRFs is obtained, allowing the analysis of 3 different regimes (expansionary, recessionary, supply-shock) with potential sign or size asymmetries in the responses for each regime. We document that due to the opposite movement of markup and marginal cost in all regimes, inflation is less volatile in the recessionary state than in the expansionary state. In addition, we find that larger shocks have a lower (as a percentage of the magnitude of the shock) and less persistent effect on inflation than shocks of a lower magnitude.

CO108 Room BH (S) 2.02 EMPIRICAL ASPECTS OF CRYPTOCURRENCY MARKETS Chair: Pierangelo De Pace

C1213: **On the dynamics of cryptocurrency prices**
Presenter: **Pierangelo De Pace**, Pomona College, United States
Co-authors: Jayant Rao

The results of an extensive empirical investigation developed over the past few years are summarized. We analyze (i) instability, interrelations, and extreme realizations in cryptocurrency markets; and (ii) average associations (comovement) and associations in the distribution tails between cryptocurrency prices and between prices of cryptocurrencies and other conventional assets. Absent a solid theory of cryptocurrency prices and given our currently poor grasp of how such prices are related and evolve, the goal is to provide a comprehensive set of clear and statistically robust empirical facts about virtual currency price dynamics.

C1470: **Green vs brown cryptocurrencies: Is proof of stake the only sustainable alternative to government currency?**
Presenter: **Marco Lorusso**, Newcastle University, United Kingdom
Co-authors: Francesco Ravazzolo, Michele Costola

The purpose is to investigate whether the proof of work technology represents an increasing environmental risk that needs to be urgently addressed. We provide a forward-looking perspective by stressing the adoption rate of the main cryptocurrencies, such as Bitcoin and Ethereum to study the implications on energy consumption and carbon footprint under alternative scenarios. In this regard, we define three main scenarios: i) the baseline scenario; ii) the Proof of Work scenario and iii) the Proof of Stake scenario. The baseline scenario represents the status quo where both technologies have the same adoption level. The Proof of Work scenario analyses the evolution of energy consumption and carbon footprints according to a total adoption of the Proof of Work technology. Conversely, the Proof of Stake scenario analyses the evolution of energy consumption and carbon footprints according to a total adoption of the Proof of Stake validation process.

C1918: **Cross-section of expected cryptocurrency returns**
Presenter: **Milan Ficura**, University of Economics in Prague, Czech Republic
Co-authors: Gonul Colak

The predictive power of over 100 factors for the modelling of weekly cryptocurrency returns is analysed on a comprehensive dataset of almost 20 000 cryptocurrencies. The analysed factors include measures of momentum, illiquidity, investor attention, volatility, downside risk and systematic risk. The employed tools include univariate and multivariate portfolio-sorts, time-series regressions and Fama-MacBeth cross-sectional regressions. We show that in contrast to small and illiquid cryptocurrencies, the behaviour of large and liquid cryptocurrencies is increasingly similar to the behaviour of stocks. We identify a strong negative impact of past volatility on future cryptocurrency returns, similar to the low-volatility anomaly observed on the stock market. We further confirm the short-term momentum, long-term reversal, and short-term max-reversal effects on the cryptocurrency market, as well as the positive effect of S&P500 betas on future cryptocurrency returns. Nevertheless, the significance of these effects drops significantly in the multivariate regression tests once the low-volatility effect is taken into account.

CO581 Room BH (S) 2.03 ADVANCES IN QUANTILE REGRESSION Chair: Harry Haupt

C0319: **Decomposition of differences in distribution under sample selection and the gender wage gap**
Presenter: **Santiago Pereda-Fernandez**, Universidad de Cantabria, Spain

The decomposition of the differences between the distribution of outcomes of two groups is addressed when individuals self-select themselves into participation. We differentiate between the decomposition for participants and the entire population, highlighting how the primitive components of the model affect each of the distributions of outcomes. Additionally, we introduce two ancillary decompositions that help uncover the sources of differences in the distribution of unobservables and participation between the two groups. The estimation is done using existing quantile regression methods, for which we show how to perform uniformly valid inference. We illustrate these methods by revisiting the gender wage gap, finding that changes in female participation and self-selection have been the main drivers for reducing the gap.

C0897: **Fixed design regression quantiles for nonstationary dependent processes**
Presenter: **Harry Haupt**, University of Passau, Germany

In many time series analysis applications, data-generating processes composed of deterministic and stochastic components are available. Prominent examples are causal stationary processes that result from filtering deterministic trends and seasonal components. After a brief discussion of motivating examples from economics, medicine, and environmental science, assumptions in recent literature are improved, and Bahadur representation and Central Limit Theorem are established for linear and nonlinear regression quantiles.

C1365: **Nonparametric quantile regression interval predictions for seasonal trending time series**
Presenter: **Joachim Schnurbus**, University of Passau, Germany
Co-authors: Ida Bauer, Harry Haupt

Prediction intervals are developed based on nonparametric quantile regression for time series exhibiting seasonal patterns, possibly nonlinear trends, and trend-season interactions. Several novel bandwidth selection methods are proposed that either relate to quantile-specific measures or interval-specific measures. The performance of the proposed predictors is analyzed in a Monte Carlo simulation and for an energy demand application.

CO252 Room BH (S) 2.05 RECENT ADVANCES IN HIGH-DIMENSIONAL ECONOMETRICS Chair: Degui Li

C1519: **Estimating factor-based spot volatility matrices with noisy and asynchronous high-frequency data**
Presenter: **Degui Li**, University of York, United Kingdom

With noisy and asynchronous high-frequency data collected for an ultra-large number of assets, we estimate high-dimensional spot volatility matrices satisfying a low-rank plus sparse structure. A localised pre-averaging method is proposed to jointly tackle the microstructure noise and asynchronicity issues, and obtain uniformly consistent estimates for latent prices. We impose a continuous-time factor model with time-varying factor loadings on the price processes, and estimate the common factors and loadings via a local principal component analysis. Assuming a uniform sparsity condition on the idiosyncratic volatility structure, we combine the POET and kernel-smoothing techniques to estimate the spot volatility matrices for both the latent prices and idiosyncratic errors. Under some mild restrictions, the estimated spot volatility matrices are shown to be uniformly consistent with convergence rates affected by the estimation errors due to the microstructure noise, asynchronicity and latent factor structures. Both simulation and empirical studies are provided to assess the numerical performance of the developed methods.

C1548: **EM algorithm for high-dimensional dynamic matrix factor models**
Presenter: **Matteo Barigozzi**, University of Bologna, Italy
Co-authors: Luca Trapin

High-dimensional matrix-variate time series data are becoming increasingly popular in economics and finance. This has stimulated the development of matrix factor models to achieve significant dimension reduction. An approximate dynamic matrix factor model is proposed that accounts for the time series nature of the data, and develops an EM algorithm to perform quasi-maximum likelihood estimation of the model parameters. The algorithm is further extended to estimate the dynamic matrix factor model on a dataset with an arbitrary pattern of missing data. The finite sample properties of the proposed estimation strategies are assessed through a large simulation study and an application to a financial dataset.

C1565: **Estimating time-varying networks for high-dimensional time series**
Presenter: **Yuning Li**, University of York, United Kingdom

Time-varying networks for high-dimensional locally stationary time series are explored using the large VAR model framework with both transition and (error) precision matrices evolving smoothly over time. Two types of time-varying graphs are investigated: one containing directed edges of Granger causality linkages, and the other containing undirected edges of partial correlation linkages. Under the sparse structural assumption, we propose a penalised local linear estimation with time-varying weighted group LASSO to jointly estimate the transition matrices and identify their significant entries, and a time-varying CLIME method to estimate the precision matrix. The estimated transition and precision matrices are then used to determine the time-varying network structures. Under some mild conditions, we derive the theoretical properties of the proposed estimates, including the consistency and oracle properties. Extensive simulation studies and an empirical application are provided to illustrate the finite-sample performance of the developed methods.

CO722 Room BH (SE) 2.05 NEWS AND THE TERM STRUCTURE OF INTEREST RATES Chair: Guillaume Roussellet

C0405: **The term structure of expectations and bond yields**
Presenter: **Stefano Eusepi**, University of Texas at Austin, United States
Co-authors: Richard Crump, Emanuel Moench

Bond yields can be decomposed into expected short rates and term premiums. We directly measure the former using all available U.S. professional forecasts and obtain the latter as the difference between bond yields and survey-based expected short rates. While the behavior of nominal and real short-rate expectations is consistent with standard macroeconomic theory, term premiums account for the bulk of the cross-sectional and time series variation in yields. They also largely explain the yield curves' reaction to a host of structural economic shocks. This dramatic failure of the expectations hypothesis highlights the importance of term premiums for macro-financial transmission.

C0949: **What do bond investors learn from macroeconomic news**
Presenter: **Guillaume Roussellet**, McGill University, Canada
Co-authors: Jean-Sebastien Fontaine, Bruno Feunou

Does the macroeconomic information in data releases shape bond yields in the long term? We offer evidence that the new information embedded in high-frequency bond price changes around the release of economic data explains around 40% of yield fluctuations in the monthly horizons but that this share decreases to 20% in the long run. From the perspective of a theoretical model in which investors use data releases to learn about the path of future short rates, our results suggest that investors' expectations are updated largely based on information revealed outside of the releases. Our results cast doubt on a transparent monetary policy response function linking macroeconomic surprises to the path of interest rates.

C0958: **What moves markets**
Presenter: **Mark Kerssenfischer**, Deutsche Bundesbank, Germany

What share of asset price movements is driven by the news? We build a large, time-stamped event database covering scheduled macro news as well as unscheduled events. We find that news accounts for about 50% of all bond and stock price movements in the United States and euro area since 2002, suggesting that a much larger share of return variation can be traced back to observable news than previously thought. Moreover, we provide stylized facts about the type of news that matter most for asset prices, the persistence of news effects, and spillover effects between the US and euro area.

CO300 Room BH (SE) 2.09 DYNAMIC MULTIPLE QUANTILE MODELS Chair: Leopoldo Catania

C1289: **Structural quantile VAR identified with external instruments**
Presenter: **Sulkhan Chavleishvili**, Aarhus University, Denmark

A framework is developed for the identification of the dynamic causal effects of macroeconomic shocks on the quantiles of distribution predictions using external instruments in a multiple dynamic regression quantiles framework. The causal impulse responses are provided in terms of the

quantiles and moments of the distribution predictions. The aim is to discuss an instrumental variables regression quantiles estimator in the time-series context and show the causal effects estimator to be consistent and asymptotically normal. The weak instrument robust inference approaches are discussed. The method is applied to study the distributional effect of the US monetary policy shocks on a set of macro-financial variables. Whether tightening the monetary policy reduces the excessive increase in asset prices and moderates future macroeconomic downside risk is studied.

C1114: **Combining dynamic conditional quantile functions with a viewtowards tail risk management**
Presenter: **Pierluigi Vallarino**, Aarhus BSS, Denmark
Co-authors: Alessandra Luati, Leopoldo Catania

A new method is introduced to model the quantiles of a time series using all past information on a set of explanatory variables and on the time series interest. The resulting quantiles: do not cross over time, have dynamics which enhance the information set available to extreme quantiles, and incorporate information coming from all regions of the conditional distributions of explanatory variables. Parameters of the model are estimated through a two-stage quasi-maximum likelihood estimator (2SQMLE). Consistency and asymptotic normality of the 2SQMLE are derived, and its finite sample properties are assessed through a simulation study. An empirical analysis concerning macro-financial variables reveals a tight connection between financial and macroeconomic tail risk, and shows that the model delivers competitive density and tail risk predictions.

C1956: **Semiparametric modeling of multiple quantiles**
Presenter: **Leopoldo Catania**, Aarhus BBS, Denmark
Co-authors: Alessandra Luati

A semiparametric model is developed to track a large number of quantiles of a time series. The model satisfies the condition of non-crossing quantiles and the defining property of fixed quantiles. A key feature of the specification is that the updating scheme for time-varying quantiles at each probability level is based on the gradient of the check loss function, which forms a martingale difference sequence. Consistency of the associated M-estimator of the fixed parameters is established. The model can be applied for filtering and prediction. We also illustrate a number of possible applications, such as: i) semiparametric estimation of dynamic moments of the observables, ii) density prediction, and iii) quantile predictions.

CO276 Room BH (SE) 2.10 FINANCIAL MODELLING AND FORECASTING Chair: Ekaterini Panopoulou

C0292: **The contribution of economic policy uncertainty to the persistence of shocks to stock market volatility**
Presenter: **Theologos Pantelidis**, University of Macedonia, Greece
Co-authors: Paraskevi Tzika

The aim is to examine the contribution of the Economic Policy Uncertainty (EPU) index to the persistence of shocks to stock market volatility. An innovative approach is applied that compares the half-life of a shock in the context of a bivariate VAR model that includes the volatility of stock market returns and EPU, with the half-life of the equivalent univariate ARMA model for the stock market return volatility. The analysis is based on daily data for the UK and the US. The empirical results corroborate that EPU contributes to the persistence of shocks to stock market volatility for both countries. This contribution is higher for the US, where 14.3% of the persistence of shocks to stock market volatility can be attributed to the EPU index. Several robustness tests support the findings.

C1272: **Predicting hedge funds' returns with machine learning methods**
Presenter: **Christos Argyropoulos**, University of Essex, United Kingdom

Profitability gains are evaluated when investors select hedge funds according to machine learning methods' returns forecasts. We use three main techniques to forecast individual hedge returns: shrinkage, dimensionality reduction, and artificial neural network. We use an extended set of predictors to calculate the forecasts, including hedge fund characteristics, risk factors and other macroeconomic variables. The accuracy of the forecasts is assessed via an out-of-sample asset allocation exercise.

C1471: **Equity premium prediction: The role of information from the options market**
Presenter: **Ekaterini Panopoulou**, University of Essex, United Kingdom
Co-authors: Antonis Alexandridis, Iraklis Apergis, Nikolaos Voukelatos

The role of information from the options market in forecasting the equity premium is examined. We provide empirical evidence that the equity premium is predictable out-of-sample using a set of CBOE strategy benchmark indices as predictors. We use a range of econometric approaches to generate a point, quantile and density forecasts of the equity premium, and we find that models based on option variables consistently outperform the historical average benchmark. In addition to statistical gains, using option predictors results in substantial economic benefits for a mean-variance investor, delivering up to a fivefold increase in certainty equivalent returns over the benchmark during the 1996-2021 sample period.

CO611 Room BH (SE) 2.12 RISK, VOLATILITY AND PRICE DISCOVERY IN FINANCIAL MARKETS Chair: Robinson Kruse-Becher

C0353: **Moment conditions and time-varying risk premia**
Presenter: **Dennis Umlandt**, University of Innsbruck, Austria

A novel approach is proposed for estimating linear factor pricing models with dynamic risk premia based on a generalized method of moment framework. Time-varying risk premia follow an updating scheme driven by the influence function of the conditional moment criterion function at time t. The most informative moment for inferring risk premium dynamics stems from the cross-sectional pricing equation that is estimated in the second stage of the popular Fama-MacBeth regression approach. In addition, the procedure can accommodate time series predictors and enhance risk premium forecasts based on errors in current moment conditions. Consistency and asymptotic normality of the generalized method of moments estimators are established. The performance of the method is investigated through a Monte Carlo study. An application to a dynamic version of the Fama-French 3-Factor model reveals factor risk premia dynamics which cannot be explained by typical time series predictors.

C0421: **Real time monitoring of a change in the persistence of stochastic volatility**
Presenter: **Emily Whitehouse**, University of Sheffield, United Kingdom

Stochastic volatility models are commonly used to describe the time-varying nature of volatility in asset returns. Much financial and econometric literature has assumed stochastic volatility to be highly persistent. Still, recent research suggests that structural breaks are common in both the level and persistence of stochastic volatility and that a failure to account for these structural breaks can cause over-estimation of the true persistence in many series. We propose a real-time monitoring test procedure for structural breaks in the persistence of stochastic volatility. We exploit the autocorrelation structure of the log-squared price return series to propose several simple test statistics based on the autocorrelations of this series. A two-tailed version of a recently proposed real-time monitoring algorithm is considered to allow the detection of both increases and decreases in persistence. Monte Carlo simulations show that our test procedure has promising levels of power to detect structural breaks of this nature.

C0325: **Information shares for markets with partially overlapping trading hours**
Presenter: **Thomas Dimpfl**, University of Hohenheim, Germany
Co-authors: Karsten Schweikert

Daily information shares for markets with partially overlapping trading hours are studied. The established methodologies consider price discovery measures computed either for exactly overlapping trading hours or in sequential markets. In contrast, we develop a framework that exploits all

price information generated during a full trading day where any market can be open or closed at any time and propose a contribution-weighted information share. We apply this new method to the S&P500 ETF and E-mini futures markets and find that conventional information shares for the ETF market are overestimated. E-mini futures are traded almost continuously throughout the trading day and process additional pricing relevant information when the ETF market is closed.

CC795 Room BH (SE) 1.02 ASSET PRICING Chair: Shixuan Wang

C1393: **Anxiety in returns**

Presenter: **Sebastian Schaefer**, University of Wuppertal, Germany

Co-authors: Uta Pigorsch

Empirical evidence is provided that risk-averse investors avoid stocks with signs of increasing uncertainty, missing 1.02 percentage points in next-month returns. The observed effect counteracts short-term reversal and supports convex risk aversion. Moreover, anxiety predicts cross-sectional returns in out-of-sample tests, strongly suggesting that empirical risk premia are driven by predictable risk-averse investors' preferences.

C1935: **Random preferences, truncated distributions, and the pricing kernel: A note**

Presenter: **Maria Magdalena Vich Llompart**, Washington College, United States

Co-authors: Luiz Vitiello

An asset pricing model is developed where risk aversion is random and has a truncated-normal distribution. We show that the distributional parameters related to risk aversion, and the limits of integration of the truncated distribution change the slope of the pricing kernel. That is, while some parameters rotate the pricing kernel clockwise, others rotate it anti-clockwise, which may have implications for asset and derivative pricing. We also show that our pricing kernel contains several others as special cases.

CC800 Room BH (SE) 1.06 REGIME SWITCHING Chair: Mohammad Jahan-Parvar

C1704: **Identification and forecasting of bull and bear markets using multivariate returns**

Presenter: **Jia Liu**, Saint Mary's University, Canada

Bull and bear market identification generally focuses on a broad index of returns through univariate analysis. A new approach is proposed to identify and forecast bull and bear markets through multivariate returns. The model assumes all assets are directed by a common discrete state variable from a hierarchical Markov switching model. The hierarchical specification allows the cross-section of state-specific means and variances to differ over bull and bear markets. We investigate several empirically realistic specifications that permit feasible estimation even with 100 assets. Our results show that the multivariate framework provides competitive bull and bear regime identification and improves portfolio performance and density prediction compared to several benchmark models, including univariate Markov switching models.

C1712: **Machine learning utility-maximising market regime classifications**

Presenter: **Richard McGee**, University College Dublin, Ireland

An unsupervised machine learning methodology is proposed to classify market periods into Bull or Bear classifications, conditional on a set of widely adopted market and macroeconomic bellwether variables. The classification of regimes is determined by decision trees that optimise the long-term investor utility of a market timing strategy, switching between Bull and Bear portfolios, whose optimal portfolio weights are themselves estimated over the sets of returns associated with each classification. Analysis covers Bull and Bear regimes across a range of thematic investment factors such as the market index, size, value and momentum.

C1940: **Regime causality**

Presenter: **Florian Ielpo**, Centre Economie de la Sorbonne, France

Co-authors: Jerome Collet

A new class of embedded Markov switching models is introduced, allowing for one time series to see its regimes influence the probability of occurrence of other regimes. This mirrors the intuition that certain economic circumstances can have an influence on financial markets' behavior. We detail this new model alongside the necessary econometrics to estimate its parameters. We then show via Monte Carlo tests the precision with which its parameter can be estimated as a function of the sample size. Finally, we provide two empirical applications, one which ties to the selection of economic data when building a nowcasting indicator, and another testing the ability of widely used economic data to influence returns on equities.

Sunday 18.12.2022 17:15 - 19:20 Parallel Session K – CFE-CMStatistics

EI025 Room Safra Lecture Theatre SPECIAL INVITED SESSION IN MEMORY OF DAVID COX Chair: Christiana Kartsonaki

E0312: **Using the Cox model for the three tasks of health data science**
Presenter: **Ruth Keogh**, London School of Hygiene and Tropical Medicine, United Kingdom
The Cox model was described by David Cox in 1972 as a model for the analysis of survival data, specifically to relate explanatory variables to the hazard of an event via hazard ratios. It has since become the most widely used model for analysis of health data and beyond. In several papers, it has been discussed how the research questions addressed in medical statistics, and health data science can be broadly classified into three tasks: description, prediction, and causal investigation. we will discuss the use of the Cox model in performing these three tasks. We will focus especially on the use of the Cox model in causal investigations, including addressing criticisms of the hazard ratio as a measure for describing causal relationships. Examples will be given from medical applications.

E0313: **D. R. Cox: Aspects of scientific inference**
Presenter: **Heather Battey**, Imperial College London, United Kingdom
A different exposition of some of D. R. Cox's scientific contributions to those presented in the original papers is given. A synthesis of some aspects will be attempted, based on insights from his other work.

E0314: **Being inspired by David R. Cox**
Presenter: **Nanny Wermuth**, Chalmers University of Technology, Sweden
I am describing our early encounters, how we started and continued to cooperate for thirty years and mention results which can be seen as building on our joint work.

EO707 Room S-2.23 RECENT STATISTICAL ADVANCES IN IMAGING Chair: Israel Almodovar Rivera

E1015: **Statistical inference for mean functions of 3D functional objects**
Presenter: **Lily Wang**, George Mason University, United States
Co-authors: Yueying Wang, Brandon Klinedinst, Guannan Wang, Auriel Willette
Functional data analysis has become a powerful tool for the statistical analysis of complex objects, such as curves, images, shapes, and manifold-valued data. Among these data objects, 2D or 3D images obtained using medical imaging technologies have been attracting researchers' attention. In general, 3D complex objects are usually collected within the irregular boundary, whereas the majority of existing statistical methods have been focused on a regular domain. To address this problem, we model the complex data objects as functional data and propose trivariate spline smoothing based on tetrahedralizations for estimating the mean functions of 3D functional objects. The asymptotic properties of the proposed estimator are systematically investigated where consistency and asymptotic normality are established. We also provide a computationally efficient estimation procedure for covariance function and corresponding eigenvalue and eigenfunctions and derive uniform consistency. Motivated by the need for statistical inference for complex functional objects, we then present a novel approach for constructing simultaneous confidence corridors to quantify estimation uncertainty. Extension of the procedure to a two-sample case is discussed together with numerical experiments and a real-data application using Alzheimer's Disease Neuroimaging Initiative database.

E1443: **Distribution-free clustering diagnostic for outlier detection**
Presenter: **Israel Almodovar Rivera**, University of Puerto Rico, United States
Finding groups in the presence of scatter can be challenging. Scatter (or outliers) observations in clustering are referred to as those observations that do not necessarily belong or fit into any cluster. We proposed a distribution-free approach to perform a diagnostic to a clustering solution to find potential outliers in it. Our approach uses a k-means solution to find the potential outliers in a homogeneous spherical group. The method uses a smooth estimation of the distribution function of the normed residuals from a given clustering solution. Further, we propose a rule-of-thumb method to compute an estimate of the smoothing parameter for the estimation of the distribution function. Then, we study the proposed diagnostic tool in several experiments with the presence of outliers in a homogeneous spherical group. Our diagnostic tool is, in general, a top performer in finding the potential outliers in these groups. Finally, we apply the distribution-free diagnostics in a functional Magnetic Resonance Imaging study to determine activated regions in a single-subject single-task experiment.

E1456: **A two-stage approach to image segmentation in forensic footwear comparisons**
Presenter: **Adam Pintar**, National Institute of Standards and Technology, United States
Co-authors: Steven Lund, Rishi Venkatasubramanian
In forensic footwear comparisons, a shoe print examiner compares an image of a shoe impression from a crime scene (the Q image) to an image of a shoe impression made in a laboratory (the K image) and reaches a conclusion such as identification, exclusion, or inconclusive. Processing the Q image, the K image, or both may aid the comparison. One type of processing, a form of noise reduction, is segmenting the shoe contact from the background in the Q image. However, the complex background observed in many Q images can make segmentation difficult. A two-stage approach to solving the problem is presented. In the first stage, simple linear iterative clustering (SLIC) is used to partition the image into regions. Where regions follow meaningful boundaries between contact and background, they may be used to quickly segment a portion of the Q image. The segmented portion yields training data to feed into a convolutional neural network algorithm for image segmentation known as U-net. The final product is a method to create a single-use neural network for the complete segmentation of one Q image.

E1461: **Reduced-rank tensor-on-tensor regression and tensor-variate analysis of variance**
Presenter: **Carlos Llosa**, Sandia National Laboratories, United States
Co-authors: Ranjan Maitra
Fitting regression models with many multivariate responses and covariates can be challenging, but such responses and covariates sometimes have a tensor-variate structure. We extend the classical multivariate regression model to exploit such structure in two ways: first, we impose four types of low-rank tensor formats on the regression coefficients. Second, we model the errors using the tensor-variate normal distribution that imposes a Kronecker separable format on the covariance matrix. We obtain maximum likelihood estimators via block-relaxation algorithms, and derive their computational complexity and asymptotic distributions. Our framework enables us to formulate a tensor-variate analysis of variance (TANOVA) methodology. This methodology, when applied in a one-way TANOVA layout, enables us to identify cerebral regions significantly associated with the interaction of suicide attempters or non-attemptor ideators and positive-, negative- or death-connoting words in a functional Magnetic Resonance Imaging study. Another application uses three-way TANOVA on the Labeled Faces in the Wild image dataset to distinguish facial characteristics related to ethnic origin, age group and gender.

EO378 Room S-2.25 MODERN TOPICS IN STATISTICAL LEARNING Chair: Radu Craiu

E0276: **Modal Grids in MIDAS estimation with large number of regressors**
Presenter: **Lorenzo Frattarolo**, European Commission Joint Research Centre (JRC), Italy
MIDAS estimation handles regressors with lower frequency using temporal aggregation with a parametrized weight distribution. Once the aggrega-

tion is done, estimation is equivalent to OLS. The proposed method exploits this feature and, given the weight function, computes a grid of weights such that each set of weights has its mode on a different lag. Then aggregation is performed for each set of weights and each regressor, resulting in a number of new aggregated regressors equal to the number of original regressors multiplied by the number of weight sets. The selection of aggregated regressors is then performed using the generalized least squares screening (GLSS). Values of parameters of the weight function originating the most significant aggregated regressors are stored and reused as initial values in a final maximum-likelihood estimation of the MIDAS regression. This methodology allows pre-selection among a large number of variables while maintaining contributions from a wide distribution of lags in the final estimation.

E0850: **Adaptively exploiting d-separators with causal bandits**
Presenter: **Blair Bilodeau**, University of Toronto, Canada
Co-authors: Linbo Wang, Daniel Roy

Multi-armed bandit problems provide a framework to identify the optimal intervention over a sequence of repeated experiments. Without additional assumptions, minimax optimal performance (measured by cumulative regret) is well-understood. With access to additional observed variables that d-separate the intervention from the outcome (i.e., they are a d-separator), recent "causal bandit" algorithms provably incur less regret. However, in practice, it is desirable to be agnostic as to whether observed variables are a d-separator. Ideally, an algorithm should be adaptive; that is, perform nearly as well as an algorithm with oracle knowledge of the presence or absence of a d-separator. We formalize and study this notion of adaptivity, and provide a novel algorithm that simultaneously achieves (a) optimal regret when a d-separator is observed, improving on classical minimax algorithms, and (b) significantly smaller regret than recent causal bandit algorithms when the observed variables are not a d-separator. Crucially, our algorithm does not require any oracle knowledge of whether a d-separator is observed. We also generalize this adaptivity to other conditions, such as the front-door criterion.

E1004: **Directional testing in astrophysics**
Presenter: **Yanbo Tang**, Imperial College London, United Kingdom
Co-authors: Nancy Reid

A previously proposed directional test for vector parameters of interest is discussed. This test is obtained through the tangent exponential model constructed by embedding the suitably re-parametrized initial model within a linear exponential model. It is shown to be highly accurate both theoretically and numerically in the small sample setting compared to other omnibus tests. We then discuss an ongoing astrophysics project in which we use the proposed directional test to determine if observed X-ray emissions from black holes are constant over time.

E1534: **Learning extremal graphical structures in high dimensions**
Presenter: **Michael Lalancette**, Technical University of Munich, Germany
Co-authors: Sebastian Engelke, Stanislav Volgushev

Multiple characterizations and models exist for extremal dependence, the dependence structure of multivariate data in unobserved tail regions. However, statistical inference for extremal dependence uses merely a fraction of the available data, drastically reducing the effective sample size and creating challenges even in moderate dimensions. Recently introduced graphical models for multivariate extremes allow for enforced sparsity in moderate- to high-dimensional settings, reducing the effective dimension. We propose a novel, scalable method for the selection of extremal graphical models that makes no assumption on the underlying graph structure, as opposed to existing approaches. It exploits existing tools for Gaussian graphical model selection, such as the graphical lasso and neighborhood selection. Model selection consistency is established in sparse regimes where the dimension is allowed to be exponentially larger than the effective sample size.

E1584: **Adapting to failure of the IID assumption**
Presenter: **Jeffrey Negrea**, University of Chicago, United States
Co-authors: Blair Bilodeau

Assumptions on data are used to develop statistical methods with optimistic performance guarantees. Even if these assumptions do not hold, we often believe that if our models are nearly correct, our methods will perform similarly to those optimistic guarantees. How can we use models that we know to be wrong, but expect to be nearly correct, in a way that is robust and reliable? We will discuss work on the canonical problem of statistical aggregation, i.e., combining predictions from a large number of models or experts. We define a continuous spectrum of relaxations of the IID assumption for prediction problems with sequential data, with IID data at one extreme and mechanisms that select worst-case responses to one's actions at the other. We develop methods for statistical aggregation with sequential data that adapt to the level of failure of the IID assumption. We quantify the difficulty of statistical aggregation in all scenarios along the spectrum we introduce, demonstrate that the prevailing methods do not adapt to this spectrum, and present new methods that are adaptively minimaxed optimal. More broadly, it is shown that it is possible to develop methods that are both adaptive and robust: they realize the benefits of the IID assumption when it holds, without ever compromising performance when the IID assumption fails, and without having to know the degree to which the IID assumption fails in advance.

EO190 Room S-1.01 ADVANCEMENTS IN THE ANALYSIS OF HIGH-DIMENSIONAL AND COMPLEX DATA Chair: Eugen Pircalabelu

E0674: **A divide-and-conquer approach for covariate-adjusted Gaussian graphical models**
Presenter: **Ensiyeh Nezakati Rezazadeh**, Catholic University of Louvain, Belgium
Co-authors: Eugen Pircalabelu

Analysis of massive data sets is challenging due to the limited capacity of available machines. To overcome this limitation, various distributed frameworks for statistical estimation and inference have been proposed. We provide statistical guarantees and asymptotic properties of the lasso estimation for covariate-adjusted Gaussian graphical models using a divide-and-conquer approach. Covariate-adjusted graphical models have many applications in the real world, especially in genomic studies when the graph structure of thousands of microRNAs is affected by thousands of DNA gene covariates. We propose a new approach to aggregate all local parallel estimators of the adjusted graphical models into a final estimator by maximizing the pseudo-log-likelihood function, which comes from the asymptotic distribution of the local debiased estimators. The asymptotic behavior of this estimator is provided when the number of parameters, covariates and machines all grow with the sample size. Due to the asymptotic distribution, statistical inference based on the final estimator is also proposed. A simulation study and a real data example are used to compare the performance of the proposed estimator relative to the naive average-based estimators.

E0815: **Optimal compressed sparse regression**
Presenter: **Alexander Munteanu**, TU Dortmund, Germany

Low-distortion random projections are at the heart of fast data compression algorithms for statistical problems such as linear regression. However, their target size is crucially at least linear in the data dimension d. A natural assumption for high dimensional regression problems is that useful solutions lie in the span of only a relatively small number of, at most, k columns. We show how the sparsity assumption helps to compress data for various regression models to within essentially tight $\Theta(k\log(d)/\varepsilon^2)$ bounds, while preserving the regression loss up to a $(1+\varepsilon)$ factor. Similar results were known for the related sparse recovery problem studied in compressed sensing, which surprisingly turns out to be a strictly easier problem allowing for smaller compression size. Finally, we show similar compression bounds for LASSO regression, a popular convex relaxation and heavily used heuristic for sparse regression.

E1645: **Rank and factor loadings estimation in time series tensor factor model by pre-averaging**
Presenter: **Weilin Chen**, London School of Economics and Political Science, United Kingdom
Co-authors: Clifford Lam

As a major dimension reduction tool, the idiosyncratic components of a tensor time series factor model can exhibit serial correlations, especially in financial and economic applications. This rules out a lot of state-of-the-art methods that assume white idiosyncratic components, or even independent/Gaussian data. While the traditional higher-order orthogonal iteration (HOOI) is proved to be convergent to a set of factor loading matrices, the closeness of them to the true underlying factor loading matrices is, in general, not established, or only under i.i.d. Gaussian noises. Under the presence of serial and cross-correlations in the idiosyncratic components and time series variables with only bounded fourth-order moments, we propose a pre-averaging method that accumulates information from tensor fibres for better estimating all the factor loading spaces. The estimated directions corresponding to the strongest factors are then used for projecting the data for a potentially improved re-estimation of the factor loading spaces themselves, with theoretical guarantees and rate of convergence spelt out. We also propose a new rank estimation method which utilises correlation information from the projected data. Extensive simulations are performed and compared to other state-of-the-art or traditional alternatives. A set of matrix-valued portfolio return data is also analysed.

E1352: **Optimal discriminant analysis in high-dimensional latent factor models**
Presenter: **Xin Bing**, University of Toronto, Canada

In high-dimensional classification problems, a commonly used approach is to first project the high-dimensional features into a lower-dimensional space, and base the classification on the resulting lower-dimensional projections. We formulate a latent-variable model with a hidden low-dimensional structure to justify this two-step procedure and to guide which projection to choose. We propose a computationally efficient classifier that takes certain principal components (PCs) of the observed features as projections, with the number of retained PCs selected in a data-driven way. A general theory is established for analyzing such two-step classifiers based on any projections. We derive explicit rates of convergence of the excess risk of the proposed PC-based classifier. The obtained rates are further shown to be optimal up to logarithmic factors in the minimax sense. Our theory allows the lower dimension to grow with the sample size and is also valid even when the feature dimension (greatly) exceeds the sample size. Extensive simulations corroborate our theoretical findings. The proposed method also performs favorably relative to other existing discriminant methods on three real data examples.

E1801: **Soft principal component regression using Bernstein matrix polynomials**
Presenter: **Keith Knight**, University of Toronto, Canada

Ridge regression and principal component (PC) regression are useful in cases where the number of predictors exceeds the number of observations or where the predictors are highly collinear. We explore a "compromise" between these two methods (soft PC regression), which uses Bernstein matrix polynomials to downweight PCs with smaller variances without eliminating them. A modification of de Casteljau's algorithm is used to compute the soft PC estimates.

EO352 Room S-1.06 ADVANCES IN MULTIVARIATE ANALYSIS: CLUSTERING, FACTOR MODELS, AND MORE Chair: Joshua Cape

E0309: **Inference after latent variable estimation for single-cell RNA sequencing data**
Presenter: **Lucy Gao**, University of British Columbia, Canada

In the analysis of single-cell RNA sequencing data, researchers often first characterize the variation between cells by estimating a latent variable, representing some aspect of the individual cell state. They then test each gene for association with the estimated latent variable. If the same data are used for both of these steps, then standard methods for computing p-values and confidence intervals in the second step will fail to achieve statistical guarantees such as Type 1 error control or nominal coverage. Furthermore, approaches such as sample splitting that can be fruitfully applied to solve similar problems in other settings are not applicable in this context. We introduce count splitting, an extremely flexible framework that allows us to carry out valid inference in this setting, for virtually any latent variable estimation technique and inference approach, under a Poisson assumption. We demonstrate the Type 1 error control and power of count splitting in a simulation study, and apply count splitting to a dataset of pluripotent stem cells differentiating to cardiomyocytes.

E0488: **Conditional probability tensor decompositions for multivariate categorical response regression**
Presenter: **Aaron Molstad**, University of Florida, United States
Co-authors: Xin Zhang

In many modern regression applications, the response consists of multiple categorical random variables whose probability mass is a function of a common set of predictors. We propose a new method for modeling such a probability mass function in settings where the number of response variables, the number of categories per response, and the dimension of the predictor are large. We introduce a latent variable model which implies a low-rank tensor decomposition of the conditional probability tensor. This model is based on the connection between the conditional independence of responses, or lack thereof, and the rank of their conditional probability tensor. Conveniently, our model can be interpreted in terms of a mixture of regressions and can thus be fit using maximum likelihood. We derive an efficient and scalable penalized expectation maximization algorithm to fit this model and examine its statistical properties. We demonstrate the encouraging performance of our method through both simulation studies and an application to modeling the functional classes of genes.

E0504: **A Generalized Latent Factor Model Approach to Mixed-data Matrix Completion with Entrywise Consistency**
Presenter: **Xiaoou Li**, University of Minnesota, United States
Co-authors: Yunxiao Chen

Matrix completion is a class of machine learning methods that concerns the prediction of missing entries in a partially observed matrix. We study matrix completion for mixed data, i.e., data involving mixed types of variables (e.g., continuous, binary, ordinal). We formulate it as a low-rank matrix estimation problem under a general family of non-linear factor models and then propose entrywise consistent estimators for estimating the low-rank matrix. Tight probabilistic error bounds are derived for the proposed estimators. The proposed methods are evaluated by simulation studies and real-data applications for collaborative filtering and large-scale educational assessment.

E1168: **Projection-type priors for structured orthogonal matrices**
Presenter: **Michael Jauch**, Florida State University, United States

A family of projection-type priors for structured (sparse or smooth) orthogonal matrices is introduced that leads to tractable posterior inference for a wide variety of statistical models built from matrix decompositions. Let $\boldsymbol{Z} = (z_{i,j})$ be a $p \times k$ matrix with i.i.d. real entries having mean zero, unit variance, and finite fourth moments; let $\boldsymbol{\Omega}$ be a $p \times p$ correlation matrix, and set $\boldsymbol{X} = \boldsymbol{\Omega}^{1/2}\boldsymbol{Z}$. The proposed prior is the distribution of $\boldsymbol{Q}_X$, the projection of $\boldsymbol{X}$ onto the Stiefel manifold obtained via the polar decomposition. It turns out that features of the distribution of $\boldsymbol{X}$ are inherited by the distribution of $\boldsymbol{Q}_X$. Most significantly, the distribution of finitely many elements of $\sqrt{p}\boldsymbol{Q}_X$ converges weakly to the distribution of the corresponding elements of $\boldsymbol{X}$ as $p, k \to \infty$ with $k/p \to 0$. Thus, if we want the prior distribution for a tall and skinny orthogonal matrix parameter to reflect structural assumptions, we can build these features into the distribution of $\boldsymbol{X}$. We illustrate the proposed prior through applications to real data and make connections to recent literature on projection-type priors as well as high-dimensional probability and random matrix theory.

E1200: **Factor analysis in high dimensional biological data with dependent observations**
Presenter: **Christopher McKennan**, University of Pittsburgh, United States

Factor analysis is a critical component of high-dimensional biological data analysis. However, modern biological data contain two key features that irrevocably corrupt existing methods. First, these data, which include longitudinal, multi-treatment and multi-tissue data, contain samples that break critical independence requirements necessary for the utilization of prevailing methods. Second, biological data contain factors with large, moderate and small signal strengths, and therefore violate the ubiquitous "pervasive factor" assumption essential to the performance of many methods. We develop a novel statistical framework and the first set of provably accurate estimators to perform factor analysis and interpret its results in dependent data with factors whose signal strengths span several orders of magnitude. We prove that this methodology can be used to solve many important and previously unsolved problems that routinely arise when analyzing dependent biological data, including high dimensional covariance estimation, subspace recovery, latent factor interpretation and data denoising. Additionally, we show that my estimator for the number of factors overcomes both the notorious "eigenvalue shadowing" problem, as well as the biases due to the pervasive factor assumption that plague existing estimators. Simulated and real data demonstrate the superior performance of my methodology in practice.

EO350 Room S-1.22 BEYOND PROPORTIONAL HAZARDS AND STANDARD SURVIVAL Chair: Dennis Dobler

E0409: **K-sample omnibus non-proportional hazards tests based on right-censored data**
Presenter: **Malka Gorfine**, Tel Aviv University, Israel

Novel and powerful tests for comparing non-proportional hazard functions are presented, which are based on sample-space partitions. Right censoring introduces two major difficulties, which make the existing sample-space partition tests for uncensored data non-applicable: (i) the actual event times of censored observations are unknown and (ii) the standard permutation procedure is invalid in case the censoring distributions of the groups are unequal. We overcome these two obstacles, introduce invariant tests, and prove their consistency. Extensive simulations reveal that under non-proportional alternatives, the proposed tests are often of higher power compared with existing popular tests for non-proportional hazards. Efficient implementation of our tests is available in the R package KONPsurv, which can be freely downloaded from CRAN.

E0553: **Factorial survival analysis for treatment effects under dependent censoring**
Presenter: **Takeshi Emura**, Kurume University, Japan
Co-authors: Dennis Dobler, Marc Ditzhaus

Factorial analyses offer a powerful nonparametric means to detect main or interaction effects among multiple treatments. For survival outcomes, e.g. from clinical trials or animal experiments, such techniques can be adopted for comparing reasonable quantifications of treatment effects. The key difficulty to solve in survival analysis concerns the proper handling of censoring. So far, all existing factorial analyses in survival data were developed under the independent censoring assumption, which is too strong for many applications. As a solution, the central aim is to develop new methods in factorial survival analyses under quite a general dependent censoring regimes. This will be accomplished by combining existing results for factorial survival analyses with techniques developed for survival copula models. As a result, we will present an appealing F-test which exhibits good performance in our simulation study and is illustrated in a real data analysis.

E0664: **A multiple kernel testing procedure for non-proportional hazards in factorial designs.**
Presenter: **Tamara Fernandez**, Universidad Adolfo Ibanez, Chile
Co-authors: Marc Ditzhaus, Nicolas Rivera

A Multiple kernel testing procedure is presented to infer survival data when several factors (e.g. different treatment groups, gender, medical history) and their interaction are of interest simultaneously. Our method is able to deal with complex data and can be seen as an alternative to the omnipresent Cox model when assumptions such as proportionality cannot be justified. Our methodology combines well-known concepts from Survival Analysis, Machine Learning and Multiple Testing: differently weighted log-rank tests, kernel methods and multiple contrast tests. By that, complex hazard alternatives beyond the classical proportional hazard set-up can be detected. Moreover, multiple comparisons are performed by fully exploiting the dependence structure of the single testing procedures to avoid a loss of power. In all, this leads to a flexible and powerful procedure for factorial survival designs whose theoretical validity is proven by martingale arguments and the theory for V-statistics. We evaluate the performance of our method in an extensive simulation study and illustrate it by a real data analysis.

E0483: **Benefit-risk assessment via generalized pairwise comparisons**
Presenter: **Brice Ozenne**, University of Copenhagen, Denmark
Co-authors: Esben Budtz-Joergensen, Julien Peron

The benefit-risk balance is critical when evaluating a new treatment, especially in oncology, where side effects may outweight small gains in survival. Combining traditional approaches, e.g. using a hazard ratio to summarize gains in survival and a chi-squared test to compare toxicity, typically leads to an obscure estimand. Instead, the net benefit has been proposed, as an extension of the Mann-Whitney parameter to multiple outcomes. It can be interpreted as the probability for a random patient in the treatment group to have a better overall outcome than a random patient in the control group (e.g. difference of at least two months in survival, or, if equivalent survival, a lower grade adverse event), minus the probability of the opposite situation. We present how estimation and uncertainty quantification can be performed in the presence of right-censoring, leveraging results from the U-statistic theory. In particular, we derive the first-order H-decomposition of the statistic. This enables us to quantify the uncertainty in sensitivity analyses, e.g. when varying the threshold for survival. These developments are implemented in the R package BuyseTest that is available on CRAN.

E0248: **Bootstrapping complex survival models: From type-II censoring to causal inference**
Presenter: **Sarah Friedrich**, University of Augsburg, Germany
Co-authors: Jasmin Ruehl

Bootstrap methods are frequently applied to derive confidence intervals in complex survival settings. The classical nonparametric bootstrap by Efron, however, relies on the independence assumption, which is not always fulfilled. For example, randomised clinical trials with time-to-event endpoints are frequently stopped after a pre-specified number of events has been observed. This practice leads to dependent data and non-random censoring, though, which can generally not be solved by conditioning on the underlying baseline information. Matters are further complicated by staggered study entry. Our simulations show that a martingale-based wild bootstrap approach still provides reasonable estimates, while Efron's classical bootstrap may lead to biased results. In the context of causal effect estimates in competing risks settings, bootstrap approaches are also often applied. We investigate the asymptotic validity of the bootstrap in these settings and compare different approaches by means of simulations.

EO578 Room S-1.27 ADVANCES IN STATISTICAL METHODOLOGY FOR THE ANALYSIS OF LONGITUDINAL DATA Chair: Samuel Manda

E0378: **The weighted least squares method for heteroscedastic interval censored survival data**
Presenter: **Najmeh Nakhaeirad**, University of Pretoria, South Africa
Co-authors: Ding-Geng Chen

In clinical and epidemiological research, the failure times are observed exactly or within certain intervals, such as in HIV infection, which is called as partly interval-censored data. If an individual takes frequent visits, then the occurrence of an asymptomatic event can be determined with sufficient accuracy, while when the visits are infrequent, the occurrence of an asymptomatic event is known to lie within an interval that may be too broad to be treated as exact. In the analysis of time-to-event data, many approaches have been proposed to estimate the parameter of the accelerated failure time model (AFT), which is popular due to its simplicity and ease of interpretability. In the classical AFT model, the random errors are

assumed independent and identically distributed, and independent of the covariates despite the fact that in practice, the random errors depend on the covariates which exhibit heteroscedasticity. In this regard, the semi-parametric weighted least squares method is extended to accommodate heteroscedastic partly interval-censored survival data. A resampling method is developed to estimate the variance of the parameter estimates. A simulation study is conducted to assess the performance of the proposed approach in the presence of interval-censored data and to evaluate the resampling method. Finally, a real dataset is analyzed for illustration.

E0383: **Modelling non-homogeneous censored time-to-event data using semiparametric accelerated failure time model**
Presenter: **Iketle Maharela**, University of Pretoria, South Africa
Co-authors: Din Chen, Lizelle Fletcher

In survival analysis, classical accelerated failure time (AFT) models provide a useful alternative to the usual proportional hazards models in analysing the associations between covariates and time-to-event data. In most cases, the basic assumption is that the event time data being analysed are homogeneous in that some covariates influencing the hazard function for an individual may not be observed or measured. Semi-parametric AFT model has recently been developed to analyse both homogeneous and heterogeneous survival data. We illustrate the performances of these models with an extensive simulation study and with application to real datasets.

E0420: **A marginal structural model for longitudinal observational data with multiple outcomes**
Presenter: **Halima Twabi**, University of Malawi, Malawi
Co-authors: Samuel Manda, Dylan Small, Hans-Peter Kohler

Causal inference methods for observational studies are available for a single outcome and often under time-invariant treatment exposure and confounders. We propose a causal inference method for longitudinal observational data with multiple outcomes using a Marginal Structural Model (MSM) and derive an Inverse Probability Weighting (IPW) estimator for balancing the time-varying confounders. We illustrate the proposed methodology by estimating the causal effect of awareness of HIV-positivity on condom use and multiple sexual partners using individuals enrolled in the Malawi Longitudinal Study of Families and Health (MLSFH). Awareness of HIV-positivity was negatively associated with multiple partners but positively associated with condom use. We have demonstrated the considerable potential of marginal structured models for estimating causal effects in longitudinal observational studies with multiple outcomes.

E0512: **Prediction of COVID-19 incidence in Africa using Bayesian a hierarchical smooth transition autoregressive model**
Presenter: **Samuel Manda**, University of Pretoria, South Africa
Co-authors: Geoffrey Singini

Disease incidence forecasting informs disease control policies, resource allocation and preparedness level of the health care system. A common approach for forecasting infectious disease incidences is based on linear time series models such as the autoregressive moving average (ARMA)-type models. Due to the time dynamics of infectious diseases, these linear time series are limited in scope. Using a simulation study, we show the performance of nonlinear smooth transition autoregressive (STAR) models in capturing the nonlinear dynamics in infectious disease data in comparison to linear time series models. The capabilities of STAR-type models are demonstrated with an application to forecasting COVID-19 incidence in African countries, with the country-specific nonlinear dynamics captured by a logistic transition function. Both in-sample and out-sample COVID-19 incidence predictions are used. The parameters of the resulting model are estimated using the Bayesian hierarchical modelling approach.

E1930: **Exploring the bidirectional pathway between intimate partner violence and depression from a cluster randomized trial**
Presenter: **Nada Abdelatif**, South African Medical Research Council, South Africa
Co-authors: Enat Chirwa, Andrew Gibbs, Samuel Manda

Intimate partner violence (IPV) predominately affects women and involves physical, sexual, emotional or psychological abuse by an intimate partner. IPV affects women of different cultures, cuts across geographical boundaries and settings, with approximately one-third of women worldwide having experienced at least one form of violence. Furthermore, it has been shown that there is a strong association between IPV and mental health, such as depression and depressive symptoms. Women who experience IPV are at higher risk of experiencing PTSD, anxiety disorders and suicidal ideation; but it has also been shown that those with depressive symptoms are at greater risk of being victims of IPV. To explore this bi-directional relationship between IPV and depression, women's experience of IPV and depression will be analyzed from a longitudinal cluster randomized trial conducted in South Africa. This will be done using longitudinal structural equation modeling. This will allow us to explore the reciprocal relationship between IPV and depression. The consistency of the model will then be tested in the men's cohort, by looking at men's perpetration of IPV and depression, and seeing whether the theoretical model can be compared between women and men.

EO630 Room K0.16 MODELING OF COMPLEX HIGH DIMENSIONAL DATA IN NEUROSCIENCE Chair: Ani Eloyan

E1025: **Tensor quantile regression with application to association between neuroimages and human intelligence**
Presenter: **Cai Li**, St. Jude Children's Research Hospital, United States
Co-authors: Heping Zhang

Human intelligence is usually measured by well-established psychometric tests through a series of problem-solving. The recorded cognitive scores are continuous but usually heavy-tailed with potential outliers and violating the normality assumption. Motivated by association studies between MRI images and human intelligence, we propose a tensor quantile regression model, which is a general and robust alternative to the commonly used scalar-on-image linear regression. Moreover, we take into account rich spatial information of brain structures, incorporating low-rankness and piecewise smoothness of imaging coefficients into a regularized regression framework. We formulate the optimization problem as a sequence of penalized quantile regressions with a generalized Lasso penalty, based on tensor decomposition, and develop a computationally efficient algorithm to estimate the model components. Extensive numerical studies are conducted to examine the empirical performance of the proposed method and its competitors. Finally, we apply the proposed method to a large-scale important dataset, the Human Connectome Project. We are able to identify the most activated brain subregions associated with quantiles of human intelligence. The prefrontal and anterior cingulate cortex are found to be mostly associated with lower and upper quantiles of fluid intelligence. The insular cortex associated with the median of fluid intelligence is a rarely reported region.

E1182: **Similarity-based multimodal regression for integrated analysis of data with complex structures**
Presenter: **Haochang Shou**, University of Pennsylvania, United States

To better understand complex human phenotypes, large-scale studies have increasingly collected multimodal data across domains such as imaging, mobile health, and physical activity. The properties of each data type often differ substantially and require either separate analyses or extensive processing to obtain comparable features for a combined analysis. Multimodal data fusion enables certain analyses on matrix-valued and vector-valued data, but it generally cannot integrate modalities of different dimensions and data structures. For a single data modality, multivariate distance matrix regression provides a distance-based framework for regression accommodating a wide range of data types. However, no distance-based method exists to handle multiple complementary types of data. We propose a novel distance-based regression model, which we refer to as Similarity-based Multimodal Regression (SiMMR), that enables simultaneous regression of multiple modalities through their distance profiles. We demonstrate through simulation, imaging studies, and longitudinal mobile health analyses that our proposed method can detect associations in

multimodal data of differing properties and dimensionalities, even with modest sample sizes. We perform experiments to evaluate several different test statistics and provide recommendations for applying our method across a broad range of scenarios.

E1499: **Randomness and statistical inference of shapes via the smooth euler characteristic transform**
Presenter: **Kun Meng**, Brown University, United States

The mathematical foundations for the randomness of shapes and the distributions of smooth Euler characteristic transform are provided. Based on these foundations, we propose an approach for testing hypotheses on random shapes. Simulation studies are provided to support our mathematical derivations and show the performance of our proposed hypothesis testing framework. Our discussions connect the following fields: algebraic and computational topology, probability theory and stochastic processes, Sobolev spaces and functional analysis, statistical inference, and medical imaging.

E1495: **Imaging and clinical biomarker estimation in Alzheimers disease**
Presenter: **Ani Eloyan**, Brown University, United States

Estimation of biomarkers related to disease classification and modelling of its progression is essential for treatment development for Alzheimer's Disease (AD). The task is more daunting for characterizing relatively rare AD subtypes such as the early-onset (AD) and others. We will describe the Longitudinal Alzheimers Disease Study (LEADS), intending to collect and publicly distribute clinical, imaging, genetic, and other types of data from people with EOAD, as well as cognitively normal (CN) controls and people with early-onset non-amyloid positive (EOnonAD) dementias. We will discuss factor-analytic methods for the estimation of clinical biomarkers of AD and their use for modeling differences in longitudinal trajectories of clinical deterioration between CN, EOAD, and EOnonAD groups in LEADS. Finally, we will discuss leveraging magnetic resonance imaging and positron emission tomography data to characterize distributions of white matter hyperintensities in people with EOAD and to obtain imaging-based biomarkers of disease trajectories of AD subtypes.

EO180 Room K0.18 ROBUST STATISTICS: A DATA DEPTH APPROACH Chair: Alicia Nieto-Reyes

E0945: **Variations of the depth based Liu-Singh two-sample test including functional spaces**
Presenter: **Felix Gnettner**, Otto-von-Guericke-Universitaet Magdeburg, Germany
Co-authors: Claudia Kirch, Alicia Nieto-Reyes

Statistical depth functions provide measures of the outlyingness, or centrality, of the elements of a space with respect to a distribution. It is a non-parametric concept applicable to spaces of any dimension, for instance, multivariate and functional. A multivariate two-sample test exists based on depth-ranks. The objective is to improve the power of the associated test statistic and incorporate its applicability to functional data. In doing so, we obtain a more natural test statistic that is symmetric in both samples. We derive the null asymptotic of the proposed test statistic, also proving the validity of the testing procedure for functional data. Finally, the finite sample performance of the test with several different depth functions for multivariate as well as functional data is illustrated by means of a simulation study.

E0649: **Robust fitting of wrapped models to multivariate torus data**
Presenter: **Luca Greco**, University G. Fortunato of Benevento, Italy
Co-authors: Claudio Agostinelli, Giovanni Saraceno

Multivariate circular data arise commonly in many different fields. Depending on the situation, observations can be thought of as points on the surface of a p-dimensional torus. The peculiarity of multivariate torus data is periodicity, which reflects in the boundedness of the sample space and often of the parametric space. Multivariate torus data are not immune to the occurrence of outliers, such as unexpected angles or directions that do not share the main pattern of the bulk of the data. Hence, a likelihood-based estimation can be badly affected, leading to unreliable results. Therefore, robust methods are needed to handle such data inadequacies with a twofold aim: lead to a robust parametric fit that is reliable under contamination and provide a testing strategy to detect outliers. Robust estimation is pursued according to a general CEM-type algorithm. In the CE step, data are suitably unwrapped on a flat torus, then the M-step is enhanced by the computation of a set of data-dependent weights aimed to down-weight outliers and mitigate their effect on the fit. We discuss and compare different strategies to measure outlyingness and evaluate weights. On the other hand, outliers detection can rely on formal rules and be based on the inspection of robust distances, rather than on weights, stemming from the robust fit.

E1126: **Tukey depth for compact and convex random sets**
Presenter: **Luis Gonzalez-De La Fuente**, University of Cantabria, Spain
Co-authors: Alicia Nieto-Reyes, Pedro Teran

The focus is on a statistical data depth with respect to compact convex random sets. It is consistent with the multivariate Tukey depth and the Tukey depth for fuzzy sets. It also provides a different perspective to an existing Tukey depth definition with respect to compact convex random sets. A series of properties are studied for the statistical data depth. These properties are based on the axiomatic notions of depth in multivariate, functional and fuzzy cases.

E1138: **Clustering via local depth functions**
Presenter: **Giacomo Francisci**, George Mason University, United States

Depth functions have been used to identify the median of multivariate distributions. Local depth functions (LDFs) involve an additional tuning parameter and are used to identify local features of the distribution, such as peaks and valleys. When the tuning parameter converges to zero, (rescaled) LDFs converge to the underlying density and are applied in a modal-like clustering algorithm. We show that, as the sample size increases, empirical clusters converge to the corresponding population clusters.

E0650: **Statistical data depth aimed for text data**
Presenter: **Alicia Nieto-Reyes**, Universidad de Cantabria, Spain

Text data has particular characteristics when transformed into quantitative data. In particular, it can be high dimensional with each datum having many zero-value components. Generally, statistical data depth functions (multivariate or functional) do not perform well under this scenario. Thus, some transformations of these functions will be proposed to tackle the problem. The performance will be shown through an application to health care text data.

EO188 Room K0.20 MACHINE LEARNING FOR EXTREMES Chair: Stephane Girard

E0554: **Partially-interpretable neural networks for extreme quantile regression**
Presenter: **Jordan Richards**, King Abdullah University of Science and Technology, Saudi Arabia
Co-authors: Raphael Huser, Emanuele Bevacqua, Jakob Zscheischler

Quantile regression is a powerful tool for modelling environmental data which exhibits spatio-temporal non-stationarity in its marginal behaviour. If our interest lies in quantifying risk associated with particularly extreme events, we may want to estimate conditional quantiles that lie outside the range of observable data; it is practical to describe characteristics of the data using a parametric extreme value model with its parameters represented as functions of predictors. Classical approaches rely on linear, or additive, functions and such models suffer in either their predictive capabilities or computational efficiency in high dimensions. Neural networks can capture complex structures in data and scale well to high dimensions, but statisticians may choose to forego their use as a result of their "black box" nature; However, they facilitate highly accurate prediction, statistical

inference with neural networks is difficult due to their abundance of estimable parameters. We propose a framework for performing extreme quantile regression using partially-interpretable neural networks, which combine semi-/parametric methods with deep learning and facilitate both high predictive accuracy and statistical inference. We use our approach to estimate extreme quantiles for a relatively high-dimensional dataset and gain insights into the drivers of extreme wildfires occurring within, and around, the Mediterranean Basin.

E1098: **Estimation of extreme quantiles from heavy-tailed distributions with neural networks**
Presenter: **Michael Allouche**, Ecole Polytechnique, France
Co-authors: Stephane Girard, Emmanuel Gobet

New parametrizations for neural networks are proposed in order to estimate extreme quantiles in both non-conditional and conditional heavy-tailed settings. All proposed neural network estimators feature a bias correction based on an extension of the usual second-order condition to an arbitrary order. The convergence rate of the uniform error between extreme log-quantiles and their neural network approximation is established. The finite sample performances of the non-conditional neural network estimator are compared to other bias-reduced extreme-value competitors on simulated data. It is shown that our method outperforms them in difficult heavy-tailed situations where other estimators almost all fail. The source code is available at github. Finally, conditional neural network estimators are implemented to investigate the behavior of extreme rainfalls as functions of their geographical location in the southern part of France.

E1158: **Cross-validation for extreme value analysis**
Presenter: **Anne Sabourin**, MAP5, UMR 8145, Universite Paris-Cite / formerly at Talacom Paris, France
Co-authors: Patrice Bertail, Anass Aghbalou, Francois Portier

A non-asymptotic study is conducted for the Cross Validation (CV) estimate of the generalization risk for learning algorithms dedicated to extreme regions of the covariates space. In this Extreme Value Analysis context, the risk function measures the algorithm's error given that the norm of the input exceeds a high quantile. The main challenge within this framework is the negligible size of the extreme training sample with respect to the full sample size and the necessity to re-scale the risk function by a probability tending to zero. We open the road to a finite sample understanding of CV for extreme values by establishing two new results: an exponential probability bound on the K-fold CV error and a polynomial probability bound on the leave-p-out CV. Our bounds are sharp in the sense that they match state-of-the-art guarantees for standard CV estimates while extending them to encompass a conditioning event of small probability. We illustrate the significance of our results regarding high dimensional classification in extreme regions via a Lasso-type logistic regression algorithm. The tightness of our bounds is investigated in numerical experiments.

E1862: **Extremal random forests**
Presenter: **Sebastian Engelke**, University of Geneva, Switzerland
Co-authors: Nicola Gnecco, Edossa Merga Terefe

Quantile regression relies on minimizing the conditional quantile loss. This has been extended to flexible regression functions such as the gradient forest. These methods break down if the quantile of interest lies outside of the range of the data. Extreme value theory provides the mathematical foundation for the estimation of such extreme quantiles. A common approach is to approximate the exceedances over a high threshold by the generalized Pareto distribution. For conditional extreme quantiles, one may model this distribution's parameters as the predictors' functions. Up to now, the existing methods are either not flexible enough or do not generalize well in higher dimensions. We develop a new approach for extreme quantile regression based on random forests that estimates the parameters of the generalized Pareto distribution flexibly, even in higher dimensions. This estimator outperforms classical quantile regression methods and methods from extreme value theory in simulation studies. We illustrate the methodology with the example of U.S. wage data.

E1749: **Physics-informed max-stable spatial processes for inference in regions with no-observations**
Presenter: **Jose Blanchet**, Stanford University, United States
Co-authors: Ali Hasan, Vahid Tarokh

Max-stable distributions are used for inference about extreme future realizations based on relatively limited observations. These multivariate distributions represent statistical laws that can be calibrated based on observations collected in a given geographical region. We explore the use of physical laws (encoded via partial differential equations) to extend the inference from areas in which observations are collected to areas in which no observations are collected, but for which physical principles can be applied. This leads to new definitions of solutions to PDEs with random input, which are appropriate for extremes. We illustrate the method in several applications, including extreme heat estimation.

EO262 Room S0.03 SPATIO-TEMPORAL HEALTH MODELING: DEVELOPMENTS Chair: Andrew Lawson

E0359: **New multivariate spatio-temporal P-spline models for areal count data**
Presenter: **Tomas Goicoa**, Universidad Publica de Navarra, Spain
Co-authors: Maria Dolores Ugarte, Gonzalo Vicente

Univariate spatio-temporal models for estimating risk or rates have been extensively used in disease mapping, mainly to study certain diseases such as cancer. However, and despite their potential, multivariate models are not so widespread in practice due to computational burden and difficulties in implementation. We propose multivariate spatio-temporal P-spline models for areal count data with special emphasis on crimes against women, a public health problem of epidemic proportions according to the World Health Organization. The joint modelling of the different crimes improves the precision of the estimates in comparison to univariate models and provides between-crime correlations. More precisely, correlations between the spatial patterns and the temporal trends of the different crimes are obtained. The models are fitted using Integrated nested Laplace approximations (INLA) and are implemented in R using the generic construction in the package R-INLA. The methodology is used to analyze four crimes against women in the Indian state of Maharashtra during the period 2001-2013.

E0596: **Bayesian spatial modeling of misaligned data using INLA and SPDE**
Presenter: **Paula Moraga**, King Abdullah University of Science and Technology (KAUST), Saudi Arabia

Spatially misaligned data are becoming increasingly common due to advances in data collection and management. We present a Bayesian geostatistical model for the combination of data obtain at different spatial resolutions. The model assumes that underlying all observations, there is a spatially continuous variable that can be modeled using a Gaussian random field process. The model is fitted using the integrated nested Laplace approximation (INLA) and the stochastic partial differential equation (SPDE) approaches. In order to allow the combination of spatially misaligned data, a new SPDE projection matrix for mapping the Gaussian Markov random field from the observations to the triangulation nodes is proposed. We show the performance of the new approach by means of simulation and an application of PM2.5 prediction in USA. The approach presented provides a useful tool in a wide range of situations where information at different spatial scales needs to be combined.

E0809: **A Bayesian surveillance metric to predict emerging high risk cluster regions of infectious disease**
Presenter: **Joanne Kim**, Medical University of South Carolina, United States
Co-authors: Andrew Lawson

Detection of the high-risk disease cluster has been the main research area for spatial disease mapping researchers. Numerous clustering methods have been developed to map the high-risk areas of the disease of interest. However, previous spatial disease mapping research focused on identifying the current hotspot of the elevated risk area. Still, it did not provide information about where the next high-risk cluster is likely to occur, given the existing hotspot. We will introduce a novel Bayesian metric to predict the occurrence of new clusters of the elevated risk areas for the infectious

disease outbreak. Our novel metric is based on the Bayesian spatio-temporal hierarchical model and extended the posterior exceedance probability, which is commonly used for hotspot clustering. Specifically, we predict the next high-risk neighboring area given the existing hotspot by 1) using both exceedance probability and exceedance level and 2) effective information sharing between the areas' own risk with its risk trend over time and its neighborhood risk. We evaluate the performance of our metric with a simulation study based on different infectious disease outbreak situations and the real data of COVID-19 outbreak. We expect our novel metric would contribute to the public health surveillance of the infectious disease outbreak by providing a novel perspective for the high-risk area cluster prediction.

E1336: **Modelling spatially misaligned disease count data with multiple severities**
Presenter: **Craig Anderson**, University of Glasgow, United Kingdom
Co-authors: Duncan Lee

Population-level disease risk varies between communities, and public health professionals are interested in mapping this spatial variation to monitor the locations of high-risk areas and the magnitudes of health inequalities. Almost all of these risk maps relate to a single severity of disease outcome, such as hospitalisation, which thus ignores any cases of disease of different severity, such as a mild case treated in a primary care setting. These spatially-varying risk maps are estimated from spatially aggregated disease count data, but the set of areal units to which these disease counts relate often varies by severity. Thus, the statistical challenge is to provide spatially comparable inference from multiple sets of spatially misaligned disease count data, and additional complexity is that the spatial extents of the areal units for some severities are partially unknown. Thus, a novel spatial realignment approach is proposed for multivariate misaligned count data, and it is applied to the first study delivering spatially comparable inference for multiple severities of the same disease. The inference is via a novel spatially smoothed data augmented Markov chain Monte Carlo algorithm, and the methods are motivated by a new study of respiratory disease risk in Scotland in 2017.

E1526: **Coupled Markov switching count models for the detection and forecasting of COVID-19 outbreaks in Quebec hospitals**
Presenter: **Alexandra Schmidt**, McGill University, Canada

COVID-19 has greatly strained the hospital system in Quebec since the first cases emerged in February 2020. We develop a novel Bayesian Markov switching model to understand better the emergence and persistence of COVID-19 outbreaks in the 30 largest Quebec hospitals and to detect/forecast the outbreaks within each hospital. We assume each hospital switches between outbreak and non-outbreak periods through a series of coupled nonhomogeneous hidden Markov chains. We allow the probability of an outbreak emerging, or persisting, in a hospital to depend on space-time covariates, such as lagged COVID-19 test positivity rates and lagged mobility data. We also allow the probability of an outbreak emerging to depend on the outbreak status of other hospitals previously, which allows the outbreaks to spread between hospitals. We assume the effects of outbreak spread can change over space and time to account for differing levels of connectivity in the hospital network. We assume incidence in the endemic period is stable and predictable overtime, following a log-linear negative binomial model with simple seasonal and time trends. During the epidemic period, we assume incidence follows a log-linear autoregressive negative binomial model to allow the cases to rise or fall rapidly.

EO162 Room S0.11 STATISTICAL OPTIMAL TRANSPORT (VIRTUAL) Chair: Asaf Weinstein

E0969: **Nonparametric Pitman efficient distribution-free testing using optimal transport**
Presenter: **Nabarun Deb**, Columbia University, United States

In recent years, the problem of optimal transport has received significant attention in statistics and machine learning due to its powerful geometric properties. We introduce the optimal transport problem and present concrete applications of this theory in statistics. In particular, we will propose a general framework for distribution-free nonparametric testing in multi-dimensions, based on a notion of multivariate ranks defined using the theory of optimal transport. We demonstrate the applicability of this approach by constructing exactly distribution-free tests for two classical nonparametric problems: (i) testing for the equality of two multivariate distributions, and (ii) testing for mutual independence between two random vectors. We investigate the consistency and asymptotic distributions of these tests, both under the null and local contiguous alternatives. We further study their local power and asymptotic (Pitman) efficiency, and show that a subclass of these tests achieve attractive efficiency lower bounds that mimic previous remarkable efficiency results. Finally, we also study the rates of convergence of the estimated optimal transport maps and show that the natural plugin estimators for these maps achieve minimax optimal rates of convergence without any tuning parameters.

E0947: **On universally consistent and fully distribution-free rank tests of vector independence**
Presenter: **Hongjian Shi**, Technical University of Munich, Germany
Co-authors: Marc Hallin, Mathias Drton, Fang Han

Rank correlations have found many innovative applications in the last decade. In particular, suitable rank correlations have been used for consistent and distribution-free tests of independence between pairs of random variables. However, the traditional concept of ranks relies on ordering data and is, thus, tied to univariate observations. As a result, it has long remained unclear how one may construct distribution-free yet consistent tests of independence between random vectors. We will discuss how this problem can be addressed via a general framework for designing multivariate dependence measures and associated test statistics based on the recently introduced concept of center-outward ranks and signs, a multivariate generalization of traditional ranks. In this framework, we obtain new multivariate Hajek asymptotic representation results and use them for local power analyses that demonstrate the statistical efficiency of our tests.

E2039: **Distributional regression via optimal transport**
Presenter: **Victor Panaretos**, EPFL, Switzerland
Co-authors: Laya Ghodrati

A framework is presented for performing regression when both covariate and response are probability distributions on a compact real interval. Our regression model is based on the theory of optimal transportation and links the conditional Frechet mean of the response distribution to the covariate distribution via an optimal transport map. We define a Frechet-least-squares estimator of this regression map, establish its consistency, and show it attains the minimax rate of convergence to the true map. The computation of the estimator is shown to reduce to an isotonic regression problem, and thus our regression model can be implemented with ease. We illustrate our methodology using real and simulated data.

E0225: **Entropy regularized optimal transport independence**
Presenter: **Zaid Harchaoui**, University of Washington, United States

An independence criterion is introduced based on entropy-regularized optimal transport. Our criterion can be used to test for independence between two samples. We establish non-asymptotic bounds for our test statistic and study its statistical behavior under both the null hypothesis and the alternative hypothesis. The theoretical results involve tools from U-process theory and optimal transport theory. We also offer a random feature type approximation for large-scale problems, as well as a differentiable program implementation for deep learning applications. We present experimental results on existing benchmarks for independence testing, illustrating the interest of the proposed criterion to capture both linear and nonlinear dependencies in synthetic data and real data.

EO545 Room S0.12 PREDICTING AND FORECASTING FOR COMPLEX DATA Chair: Matus Maciak

E0417: **Testing dependencies in high dimensions**
Presenter: **Marie Huskova**, Charles University, Czech Republic

Co-authors: Zdenek Hlavka, Simos Meintanis

The focus is on various tests for independencies in multivariate as well as functional sequences of time series. Both serial, as well as mutual procedures, are considered. The focus is on tests formulated as L2-type criteria based on characteristic empirical functions. Presented are asymptotic properties, computational aspects, simulation results as well as application to some data.

E0492: **Semi-continuous time series for sparse data with volatility clustering**

Presenter: **Michal Pesta**, Charles University, Czech Republic

Co-authors: Sarka Hudecova

Time series containing a non-negligible portion of possibly dependent zeros, whereas the remaining observations are positive, are considered. They are regarded as GARCH processes consisting of non-negative values. The aim lies in the estimation of the omnibus model parameters taking into account the semi-continuous distribution. The hurdle distribution, together with dependent zeros, causes the classical GARCH estimation techniques to fail. Two different likelihood-based approaches are derived, namely the maximum likelihood estimator and a new quasi-likelihood estimator. Both estimators are proved to be strongly consistent and asymptotically normal. Predictions with bootstrap add-ons are proposed. The empirical properties are illustrated in a simulation study, which demonstrates the computational efficiency of the methods employed. The developed techniques are presented through an actuarial problem concerning sparse insurance claims.

E1280: **Goodness-of-fit tests for Gaussian random processes**

Presenter: **Daniel Hlubinka**, Univerzita Karlova, Czech Republic

Co-authors: Zdenek Hlavka

Goodness-of-fit tests are introduced for families of Gaussian random processes parametrized by finite-dimensional parameters. The tests are based on the Cramer-von Mises distance of characteristic functionals. The main advantage of the characteristic functional approach is the sensitivity of the tests to violation of Gaussianity contrary to the classical tests based on covariance operators. We show several examples, including Ornstein-Uhlenbeck processes, fractional Brownian motion and Cox-Ingersoll-Ross processes.

E1466: **On the stability and generalization of the privacy-preserving decentralized learning**

Presenter: **Yafei Wang**, University of Alberta, Canada

The stochastic decentralized optimization that minimizes a finite sum of expected losses over a topology network diagram has found tremendous success within distributed and parallel learning due to its natural relevance to sophisticated computing and large-scale optimization. With the communication across the nodes through the network system, the concerns of privacy risk have motivated the development of private variants of learning algorithms for many complex inference and training tasks. We discuss a novel formulation of operator splitting schemes that solve complicated monotone inclusions and stochastic optimization problems built by alliteratively updating each piece of decomposition. In particular, leveraging the decentralized learning procedure to train models under privacy constraints, we propose a general framework of privacy-preserving stochastic decentralized operator iteration algorithms and show that the proposed algorithm retains the performance guarantee in terms of stability, generalization, and finite sample performance. We further investigate the impact of the local privacy-preserving computation on global differential privacy through the composition theorem.

E0681: **Online changepoint test in a nonlinear expectile model**

Presenter: **Matus Maciak**, Charles University, Czech Republic

Co-authors: Gabriela Ciuperca, Michal Pesta

An automatic data-driven changepoint test is presented to detect specific structural changes within an underlying stochastic model. The proposed methodology is based on a nonlinear (parametric) regression framework which ensures relatively large flexibility of the overall model. Conditional expectiles—well-known in econometrics for being the only coherent and elicitable risk measure—induce some robustness in the model estimation, and the proposed statistical test is proved to be consistent while the distribution of the test statistic under the null hypothesis does not depend on the functional form of the underlying model neither on the unknown parameters. Therefore, relatively easy and straightforward practical application is guaranteed. Important theoretical details are discussed and finite sample empirical properties (simulations and real data example) are presented.

EO222 Room S0.13 PROJECTION PURSUIT: PREDICTION Chair: Nicola Loperfido

E0639: **Supervised projection pursuit for discriminant analysis through machine learning**

Presenter: **Donald Jacobs**, University of North Carolina at Charlotte, United States

Co-authors: Tyler Grear, Chris Avery

Supervised projection pursuit is implemented as a neural network used for dimensionality reduction and building classification models. The architecture consists of a single layer of interacting neurons that serve as both input and output layers. Each neuron has access to the mean and covariance of the input data, and represents a basis vector that projects data to obtain two emergent features (mean and variance). Efficacy measures how well data clusters in a mode feature space plane. Because efficacy is linearly separable across all projections, a stochastic process prioritizes Jacobi or Cayley rotations to monotonically increase net efficacy, while the complete and orthonormal basis set is maintained. Each k-th projection contributes to the efficacy associated with a rectifying unit as the product of two conjugate variables: $E(k) = Q(k)S(k)$, where $E(k)$ is efficacy, $S(k)$ is a projection index, and $Q(k)$ is the quality of clustering. Once converged, multivariate differences and similarities are quantified by projections within discriminant and indifferent subspaces. The application to molecular function recognition is described. The input is a time series of atomic motions over many molecular trajectories of functional and non-functional molecules. The results identify functional dynamics that describe biochemical mechanisms important for drug discovery.

E1002: **Projection pursuit supervised classification**

Presenter: **Natalia da Silva**, Universidad de la Republica, Uruguay

Co-authors: Di Cook, Eun-Kyung Lee

Projection pursuit random forest (PPF) is a new ensemble learning method for classification problems, built from trees utilizing combinations of predictors. PPF builds a forest from many projection pursuit trees (PPtree); trees are constructed by splitting on linear combinations of randomly chosen variables. Projection pursuit is used to find the linear combination of variables that best separates groups, and many different rules to make the actual split are provided. Utilizing linear combinations of variables to separate classes takes the correlation between variables into account, which allows PPF to outperform a traditional random forest when separations between groups occur in combinations of variables. PPF can be used in multi-class problems and is implemented into an R package PPforest. Some extensions of the individual trees in PPF are explored to make the classifier more flexible, to tackle more complex problems, while maintaining interpretability.

E1123: **Projection pursuit regression versus generalized additive model for location scale and shape an application in health**

Presenter: **Jose Pereira**, Universidade do Porto, Portugal

Co-authors: Teresa Oliveira, Luzia Mendes

The relationship of periodontal probing depth (PPD) with age, high-density lipoproteins (HDL) and diabetic status (DS) was addressed using two different methods, the projection pursuit regression model (PPR) and the generalized additive model for location scale and shape (GAMLSS). The first is non-parametric and non-linear by nature, and the second is a distributional semi-parametric regression method. In the gamlss model was assumed a truncated exponential modified Gaussian distribution of PPD, with three distribution parameters ($\mu, \nu and \tau$) to be modeled as a function

of data. The results were similar, with both models yielding the same r squared (0.31), uncovering a curve shape relationship between PPD and HDL and DS, and their effects on the dependent variable are of the same sign, allowing for similar conclusions. From the user's perspective, the advantages of PPR over GAMLSS are that interactions between predictors do not need to be explained, and a probabilistic distribution for the dependent variable is not assumed a priori. Moreover, for a sufficiently large number of terms, it can approximate any continuous function in $\mathbb{R}_p$. However, the interpretation of the PPR is not as intuitive as GAMLSS models.

E0936: **Prediction of random variables by excursion metric projections**
Presenter: **Vitalii Makogin**, Ulm University, Germany
Co-authors: Evgeny Spodarev

The concept of excursions is used for the prediction of random variables without any moment existence assumptions. To do so, an excursion metric on the space of random variables is defined, which appears to be a kind of weighted L_1 distance. Using equivalent forms of this metric and the specific choice of excursion levels, we formulate the prediction problem as a minimization of a certain target functional, which involves the excursion metric. The existence of the solution and the weak consistency of the predictor are discussed. An application to the extrapolation of stationary heavy-tailed random functions illustrates the use of the aforementioned theory. Numerical experiments with the prediction of Gaussian, alpha-stable and further heavy-tailed time series demonstrate a good performance of our method.

E1941: **On the efficacy of neural networks as good tools for the process of the statistical projection pursuit**
Presenter: **Mahmoud Mansour**, The British University in Egypt, Egypt

Reducing multidimensional data sets is a must for obtaining a good statistical analysis, like Partial least squares regression and Minimum expected posterior loss. In contrast to earlier dimension reduction techniques, we aim to present a new algorithm for a neural network that enjoys good features adapted to statistical analysis techniques. Through the examination of two models and data sets, we demonstrate the effectiveness of the proposed dimension reduction algorithm.

EO254 Room Virtual R02 GEOSPATIAL HARMONIZATION: DYNAMIC PREDICTION AND MAPPING Chair: Eleni-Rosalina Andrinopoulou

E1143: **Geospatial harmonization with multimodal geomarkers and R package developments**
Presenter: **Erika Rasnick**, Cincinnati Children's Hospital Medical Center, United States
Co-authors: Emrah Gecili, Anushka Palipana, Patrick Ryan, Eleni-Rosalina Andrinopoulou, Pedro Miranda Afonso, Ruth Keogh, John Clancy, Rhonda Szczesniak, Cole Brokamp

It is known that environmental and place-based characteristics, termed geomarkers, play a role in the development and exacerbation of respiratory diseases such as cystic fibrosis (CF). However, geomarker data comes from many different sources, exists in diverse file formats, and is available at varying spatial and temporal extents and resolutions. We harmonized a set of geospatial variables, including air pollution exposure estimates from land use regression models, weather data, roadway proximity, neighborhood deprivation and crime, land use characteristics, and access to care. For each geomarker, we developed methods for estimating an individual's exposure based on their geocoded residential address, as well as postal ZIP code estimation for scenarios when full addresses are not available. We also created R packages that make geomarker data publicly available and easily accessible and make our geomarker workflows reproducible. The harmonized geomarker dataset and R packages were used for further analyses of CF and environmental and place-based exposures.

E1663: **Acute exposure to ambient particulate matter and pulmonary exacerbations: A case-crossover simulation study**
Presenter: **Stephen Colegate**, Cincinnati Children's Hospital Medical Center, United States
Co-authors: Cole Brokamp, Rhonda Szczesniak, Marepalli Rao

Cystic fibrosis (CF) is an autosomal recessive disorder characterized by chronic lung infections and recurrent respiratory symptoms called pulmonary exacerbations (PEx). Individual PEx symptoms may substantially vary according to changes in lung function, nutrition and other clinical factors, and there is no consensus on timely treatment initiation. The Early Intervention in Cystic Fibrosis Exacerbation Study (eICE) determined whether early treatment of PEx events with home monitoring devices is beneficial. We implemented a case-crossover design to evaluate whether PEx events were associated with daily ambient fine particulate matter (PM2.5). The results indicate an increase in PEx cases with higher PM2.5, although this association depends on the selection of a suitable washout period between two consecutive PEx events. We simulate data based on our applied eICE study to estimate the bias and precision of the odds ratio (OR) estimate for various washout periods. Our case-crossover simulation shows how a short washout period introduces bias while also demonstrating how a conservative washout period affects the precision of the OR estimate. We strongly recommend a washout period of 14 days for the case-crossover design modeling PEx outcomes. The simulation study illustrates a broader implication for the development and application of the case-crossover design.

E1210: **Granularity in deriving environmental exposures & community characteristics: Impacts on predictive accuracy**
Presenter: **Anushka Palipana**, Cincinnati Children's Hospital Medical Center, United States
Co-authors: Emrah Gecili, Rhonda Szczesniak, Erika Rasnick, Andrew Vancil, Daniel Ehrlich, Teresa Pestian, Eleni-Rosalina Andrinopoulou, Pedro Miranda Afonso, Ruth Keogh, Yizhao Ni, John Clancy, Patrick Ryan, Cole Brokamp

Nearly 50% of the variability in lung function measurements from individuals with cystic fibrosis (CF) is attributable to environmental influences. Little is known about how the resolution with which these environmental exposures and community characteristics (geomarkers) are measured leads to biased and/or imprecise estimates and predictions in longitudinal modeling. Although differing resolution has been shown to yield biased associations with health outcomes, research has focused on a limited number and type of geomarkers. In this empirical study, we evaluate geomarker measurements derived for a local CF center cohort ($n = 148$, aged 6 to 20 years, followed from 2012 to 2017) to determine the extent to which geomarker granularity, coded based on US postal zip code or residential address, impacts the accuracy of dynamic prediction modeling of lung function decline. We employ a stochastic linear mixed effects model with target functions tailored to prediction of clinically relevant thresholds of rapid lung function decline in CF. A novel Bayesian selection approach for clinical and geomarker covariates is presented. Findings from the two different derivation types (zip code and residential address) are compared for the real-world CF clinical data and simulation settings. Implications and trade-offs of each derivation are discussed.

E1725: **Machine learning methodologies for the prediction of rapid lung function decline**
Presenter: **Judith Dexheimer**, Cincinnati Children's Hospital, United States

Cystic Fibrosis (CF) affects more than 70,000 individuals worldwide. Patients have an average of 9 visits per year. We collected data from the US Cystic Fibrosis Foundation Patient Registry which contains clinical encounter-level data obtained from patients at accredited care centers. We developed machine learning methodologies to predict the decrease in lung function, FEV1-Indicated Exacerbation Signal (FIES). We developed five machine learning models: Logistic Regression (LR), Random Forest (RF), K-Nearest Neighbors (KNN), Artificial Neural Network (ANN), and a Recurrent Neural Network (RNN). Outcomes included area under the receiver operating characteristic curve (AUROC), sensitivity, specificity, recall, and precision. From the 20,153 patients with data in the registry, 13,285 were included in the analysis. Patients were split into a training and validation cohort. Using 10-fold cross validation, the RNN and RF performed best with AUROC = 0.900 and AUROC=0.885 respectively. Findings across all methods will be compared, and implications of recall and precision will be discussed including accounting for longitudinal dependencies in electronic health record data.

E1600: **Jointly modeling lung function decline, nutritional evolution and pulmonary exacerbation onset**
Presenter: **Pedro Miranda Afonso**, Erasmus University Medical Center, Netherlands
Co-authors: Rhonda Szczesniak, Dimitris Rizopoulos, Grace Zhou, John Clancy, Anushka Palipana, Erika Rasnick, Cole Brokamp, Patrick Ryan, Ruth Keogh, Eleni-Rosalina Andrinopoulou
Cystic fibrosis (CF) is an inherited disease primarily affecting the lungs and gastrointestinal tract. It is of clinical interest to simultaneously investigate the association between the risk of recurrent pulmonary exacerbations (PEx), lung function measured as percent-predicted of forced expiratory volume in 1 second (FEV1) decline, nutritional status (BMI) evolution, and the risk of lung transplant/death. Previous work has been limited to continuous longitudinal markers and time-to-first PEx, and ignored the spatial variability among individuals. This was mainly due to the unavailability of appropriate and robust statistical methods and software. We propose a Bayesian hierarchical model for jointly modeling multiple longitudinal markers, and recurrent and terminal event processes. We account for the individual geographical location and explore different forms of association between the markers and the events of interest. The developed model is available in the R statistical package JMbayes2. Full MCMC algorithm implementation in C++ enables model fit in a timely fashion, despite its complexity. The proposed multivariate joint model allows more efficient use of all available data. It thereby brings new insights into CF disease progression across different geographical regions and enhances our understanding of risks posed by PEx.

EO312 Room Virtual R03 STATISTICAL METHODS IN CAUSAL INFERENCE AND REINFORCEMENT LEARNING Chair: Chengchun Shi

E1582: **Constructing stabilized dynamic surveillance rules for optimal monitoring schedules**
Presenter: **Yingqi Zhao**, Fred Hutchinson Cancer Research Center, United States
Dynamic surveillance rules (DSRs) are sequential surveillance decision rules informing monitoring schedules in clinical practice, which can adapt over time according to a patient's evolving characteristics. In many clinical applications, it is desirable to identify and implement optimal stabilized DSRs, where the parameters indexing the decision rules are shared across different decision points. We propose a new criterion for DSRs that accounts for the benefit-cost tradeoffs during the course of disease surveillance. We develop two methods to estimate the stabilized DSRs optimizing the proposed criterion, and establish asymptotic properties for the estimated parameters of biomarkers indexing the DSRs. The first approach estimates the optimal decision rules for each individual at every stage via regression modeling, and then estimates the stabilized DSRsvia a classification procedure with the estimated time-varying decision rules as the response. The second approach proceeds by optimizing a relaxation of the empirical objective, where a surrogate function is utilized to facilitate computation. Extensive simulation studies are conducted to demonstrate the superior performances of the proposed methods. The methods are further applied to the Canary Prostate Active Surveillance Study (PASS).

E0209: **Testing an elaborate theory of a causal hypothesis in an observational study**
Presenter: **Dylan Small**, University of Pennsylvania, United States
Co-authors: Bikram Karmakar
When R.A. Fisher was asked what can be done in observational studies to clarify the step from association to causation, he replied, Make your theories elaborate: when constructing a causal hypothesis, envisage as many different consequences of its truth as possible and plan observational studies to discover whether each of these consequences is found to hold. William Cochran called this multi-phasic attack one of the most potent weapons in observational studies. Statistical tests for the various pieces of the elaborate theory help to clarify how much the causal hypothesis is corroborated. In practice, the degree of corroboration of the causal hypothesis has been assessed by a verbal description of which of the several tests provides evidence for which of the several predictions. This verbal approach can miss quantitative patterns. We develop a quantitative approach to making statistical inference about the amount of the elaborate theory that is supported by evidence.

E1351: **A robust whole-genome Mendelian randomization approach for improved estimation and inference of causal effects**
Presenter: **Haoyu Zhang**, National Cancer Institute, United States
Co-authors: Zhonghua Liu, Xihong Lin
Mendelian randomization (MR) uses genetic variants as instrumental variables (IV) to assess causal associations between modifiable risk factors and diseases. The method only requires summary-level statistics from genome-wide association studies (GWAS), reducing logistical load and making it widely used. MR makes several strong assumptions, which can be violated in practice and lead to biased estimates. For example, MR estimates can be biased due to weak IVs, pleiotropic effects, or sample overlaps. To address these issues, we develop a whole-genome MR method (WMR) that accounts for linkage disequilibrium (LD) across genetic variants. We assume pleiotropic effects follow distributions with mean 0 and a fixed variance, and estimate the variance of pleiotropiceffects using LD-score regression. Our simulation analyses mimic real LD patterns using 1000 Genomes Project European haplotype data. We simulate different proportions of causal SNPs, levels of pleiotropic effects, and sample overlap proportions. We find that WMR is robust to weak IVs and different levels of pleiotropic effects. Meanwhile, WMR provides a smaller empirical standard error than alternative approaches since WMR uses more correlated IVs. We apply the proposed methods along with other approaches to many traits, including BMI, lipids-related traits, cardiovascular disease, breast cancer, etc. The standard error of WMR estimates is smaller than alternative approaches, consistent with simulations.

E1635: **Learning Bellman complete representations for offline policy evaluation**
Presenter: **Nathan Kallus**, Cornell University, United States
Co-authors: Jonathan Chang, Kaiwen Wang, Wen Sun
Representation learning for offline reinforcement learning is studied, focusing on the task of off-policy evaluation (OPE). Recent work shows that, in contrast to supervised learning, realizability of the Q-function is not enough for learning it. Two sufficient conditions for sample-efficient OPE are Bellman completeness and coverage. Prior work often assumes that representations satisfying these conditions are given, with results being mostly theoretical in nature. We propose BCRL, which directly learns from data an approximately linear Bellman complete representation with good coverage. With this learned representation, we perform OPE using Least Square Policy Evaluation (LSPE) with linear functions in our learned representation. We present an end-to-end theoretical analysis, showing that our two-stage algorithm enjoys polynomial sample complexity provided some representation in the rich class considered is linear Bellman complete. Empirically, we extensively evaluate our algorithm on challenging, image-based continuous control tasks from the Deepmind Control Suite. We show that our representation enables better OPE compared to previous representation learning methods developed for off-policy RL. BCRL achieves competitive OPE error with the state-of-the-art method Fitted Q-Evaluation (FQE), and beats FQE when evaluating beyond the initial state distribution. Our ablations show that both linear Bellman complete and coverage components of our method are crucial.

E0242: **Sensitivity analysis of individual treatment effects: A robust conformal inference approach**
Presenter: **Ying Jin**, Stanford University, United States
Co-authors: Zhimei Ren, Emmanuel Candes
A model-free framework is proposed for sensitivity analysis of individual treatment effects (ITEs), building upon ideas from conformal inference. For any unit, our procedure reports the Gamma-value, a number which quantifies the minimum strength of confounding needed to explain away the evidence for ITE. Our approach rests on the reliable predictive inference of counterfactuals and ITEs in situations where the training data is confounded. Under a previous marginal sensitivity model, we characterize the shift between the distribution of the observations and that of the counterfactuals. We first develop a general method for predictive inference of test samples from a shifted distribution; we then leverage this to

construct covariate-dependent prediction sets for counterfactuals. No matter the value of the shift, these prediction sets (resp. approximately) achieve marginal coverage if the propensity score is known exactly (resp. estimated). We describe a distinct procedure also attaining coverage, however, conditional on the training data. In the latter case, we prove a sharpness result showing that for certain classes of prediction problems, the prediction intervals cannot possibly be tightened. We verify the validity and performance of the new methods via simulation studies and apply them to analyze real datasets.

EO744 Room Virtual R04 STATISTICAL LEARNING IN MODERN COMPLEX DATA ANALYSIS Chair: Guanqun Cao

E0486: **Efficient nonparametric estimation of Toeplitz covariance matrices**
Presenter: **Tatyana Krivobokova**, University of Vienna, Austria

A new nonparametric estimator for Toeplitz covariance matrices based on a periodic smoothing spline estimator of the log-spectral density function is proposed. This estimator is positive definite by construction, fully data-driven and computationally very fast. Moreover, the estimator is shown to be minimax optimal under the spectral norm for a large class of Toeplitz matrices. These results are readily extended to inverses of Toeplitz covariance matrices. Also, an alternative version of the Whittle likelihood for the spectral density based on the Discrete Cosine Transform (DCT) is proposed.

E0610: **Minimax nonparametric multi-sample test under smoothing**
Presenter: **Pang Du**, Virginia Tech, United States
Co-authors: Xin Xing, Zuofeng Shang, Ping Ma, WenXuan Zhong, Jun Liu

The problem of comparing probability densities among multiple groups is considered. A new probabilistic tensor product smoothing spline framework is developed to model the joint density of two variables. Under such a framework, the probability density comparison is equivalent to testing the presence/absence of interactions. We propose a penalized likelihood ratio test for such interaction testing and show that the test statistic is asymptotically chi-square distributed under the null hypothesis. Furthermore, we derive a sharp minimax testing rate based on the Bernstein width for nonparametric multi-sample tests and show that our proposed test statistic is minimax optimal. In addition, a data-adaptive tuning criterion is developed to choose the penalty parameter. Simulations and real applications demonstrate that the proposed test outperforms the conventional approaches under various scenarios.

E0834: **Oracle-efficient estimation for variance function of dense functional data with a smooth simultaneous confidence band**
Presenter: **Suojin Wang**, Texas A and M University, United States
Co-authors: Li Cai, Suojin Wang

A new two-step reconstruction-based moment estimator and an asymptotically correct smooth simultaneous confidence band (SCB) is proposed for the heteroscedastic variance function of dense functional data. Step one involves spline smoothing for individual trajectory reconstructions, and step two employs kernel regression on the individual squared residuals to estimate each trajectory variability. Then by the method of moment, an estimator for the variance function of functional data is constructed. The estimation procedure is innovative by synthesizing spline smoothing and kernel regression together, which allows one not only to apply the fast computing speed of spline regression but also to employ the flexible local estimation and the extreme value theory of kernel smoothing. The resulting estimator for the variance function is shown to be oracle-efficient in the sense that it is uniformly as efficient as the ideal estimator when all trajectories were known by "oracle". As a result, an asymptotically correct SCB as a global inference tool for the variance function is established. Simulation results support our asymptotic theory with fast computation. As an illustration, the proposed method is applied to the analyses of two real data sets leading to a number of discoveries.

E1226: **Estimation of dynamic diarrhea effects on childhood growth with latent subgroup**
Presenter: **Jianhui Zhou**, University of Virginia, United States
Co-authors: Tonghao Zhang, Jennie Ma, William Petri

Existing studies treat diarrhea effect on childhood growth as constant over time, and also assume the effect is the same across the population. However, in practice, diarrhea episodes may have different clinical severity depending on the types of viruses infected and on the diarrhea onset time. Moreover, the individual reaction to a diarrhea episode may vary across children due to their different health characteristics. As a result, the diarrhea effect on children growth is heterogeneous across the cohort and is dynamic across time. We propose a semi-parametric model with latent subgroup to model the heterogeneity and dynamics of the diarrhea effect. To accommodate the heterogeneity of the effect, natural growth and diarrhea effect are modeled as random effect curves and the mixture distributions are adopted. The latent subgroup in the mixture distribution model helps to explain individual characteristics such as household socioeconomic status, hygiene level, and other driving factors behind diarrhea. Simulation studies show that our model achieves simultaneous identification of subgroups in growth pattern and diarrhea vulnerability and estimation of the dynamic diarrhea effect curves. The estimator variance is also validated by empirical coverage probability. The proposed model is applied to a dataset collected from children in Bangladesh, to illustrate the application of the proposed methods.

E1380: **Inter-subject correlation analysis for heterogeneous functional data**
Presenter: **Ping-Shou Zhong**, University of Illinois at Chicago, United States
Co-authors: Hongnan Wang

A focus of the inter-subject correlation (ISC) analysis is to understand the correlation among individuals' brain activities to identify the brain regions that respond similarly to the same real-life stimuli. It plays an important role in neuroscience research. The aim is to develop a consistent test for the ISC analysis with fMRI data. We explore the benefit of using nonparametric smoothing in the ISC test and propose a nonparametric test procedure for testing the existence of the inter-subject correlation. More specifically, testing whether the covariance matrix among subjects is diagonal. Our proposed test is applicable under subject heteroscedasticity and temporal heteroscedasticity. We establish the asymptotic distributions of the proposed test statistics under the null hypothesis and a series of local alternative hypotheses. Numerical studies show that the proposed test procedure performs better than the commonly used methods in the ISC studies and cross-sectional dependence tests, including the adjusted Lagrange multiplier test, Pesarans cross-sectional dependence (CD) test, and the adjusted Pesarans CD test.

EO570 Room Virtual R05 RECENT ADVANCES IN STOCHASTIC MODELS II Chair: Anna Panorska

E1848: **Modelling of anomalous diffusion processes with random parameters**
Presenter: **Agnieszka Wylomanska**, Wroclaw University of Science and Technology, Poland

Anomalous diffusion processes are observed in various phenomena. One of the classical anomalous diffusion model is the fractional Brownian motion (FBM), a Gaussian non-Markovian self-similar process with stationary long-correlated increments. The correlation and diffusion properties of this random motion are fully characterized by its Hurst exponent. However, recent biological experiments revealed highly complicated anomalous diffusion phenomena that can not be attributed to a class of self-similar random processes. Inspired by these observations, we study the process which preserves the properties of FBM at a single trajectory level; however, the Hurst index randomly changes from trajectory to trajectory. We provide a general mathematical framework for analytical, numerical and statistical analysis of FBM with random Hurst exponent. The explicit formulas for probability density function, mean square displacement and autocovariance function of the increments are presented for three generic distributions of the Hurst exponent, namely two-point, uniform and beta distributions. The important features of the process studied here are accelerating diffusion and persistence transition, which we demonstrate analytically and numerically.

E1854: **Identification and validation of periodic autoregressive model with additive noise: Finite-variance case**
Presenter: **Wojciech Zulawinski**, Wroclaw University of Science and Technology, Poland
Co-authors: Aleksandra Grzesiek, Radoslaw Zimroz, Agnieszka Wylomanska

The problem of modelling data containing periodic autoregressive (PAR) time series and additive noise is discussed. In most cases, such data are processed under noise-free model (i.e., without additive noise) assumptions that cannot be accepted in real life. The first two steps in PAR model identification are order selection and period estimation; thus, the main attention is paid to those issues. Finally, the model should be validated; thus the procedure for analysis of the residuals, considered here as multidimensional vectors, is proposed. Both issues (order and period selection, as well as model validation) are implemented here by using the characteristic function (CF) of the residual series. The CF is applied to receive the probability density function used in the information criterion. In the case of residuals analysis, it is used for the residuals distribution testing. To complete the PAR model, the procedure for estimation of the coefficients is required; however, this problem is just recalled, as it is a separate issue (prepared in parallel). The presented methodology can be considered as the general framework for analysis of data with periodically non-stationary characteristics disturbed by finite-variance external noise. The original contribution is related to optimal model order and period identification as well as analysis of residuals. All these findings have been inspired by our earlier work on machine condition monitoring using PAR modelling.

E1953: **Ornstein - Uhlenbeck process driven by alpha-stable process and its Gamma subordination**
Presenter: **Aleksandra Grzesiek**, Wroclaw Univeristy of Science and Technology, Poland
Co-authors: Agnieszka Wylomanska, Janusz Gajda

The variety and diversity of phenomena surrounding us and easy access to empirical data require either new and more complicated models that allow capturing features to resemble the data. We study the Ornstein-Uhlenbeck (OU) process driven by the alpha-stable Levy process delayed by the Gamma subordinator. The considered model captures the essential features of the parent process, i.e., the OU process with heavy-tailed-based distribution; however, it also possesses some characteristics that are not adequate to the model without the subordination scenario. Thus, it can be beneficial for real data with very specific behavior. The considered model can be regarded as the natural extension of the variance Gamma process that arises as the ordinary Brownian motion time-changed by the Gamma process.

E2004: **Slash distributions, generalized convolutions, and extremes**
Presenter: **Marek Arendarczyk**, University of Wroclaw, Poland
Co-authors: Tomasz Kozubowski, Anna Panorska

An α-slash distribution built upon a random variable X is a heavy-tailed distribution corresponding to $Y = X/U^{1/\alpha}$, where U is a standard uniform random variable, independent of X. We point out and explore a connection between α-slash distributions, which are gaining popularity in statistical practice, and generalized convolutions, which come up in probability theory in connection with generalizations of the standard concept of convolution of probability measures. Our theoretical results are illustrated by several examples involving standard and novel probability distributions and extremes.

E2003: **Tempered fractionally integrated process with stable noise as a transient anomalous diffusion model**
Presenter: **Krzysztof Burnecki**, Wroclaw University of Science and Technology, Poland
Co-authors: Farzad Sabzikar, Jinu Kabala

The autoregressive tempered fractionally integrated moving average (ARTFIMA) process is presented, which is obtained by taking the tempered fractional difference operator of the non-Gaussian stable noise. The tempering parameter makes the ARTFIMA process stationary for a wider range of the memory parameter values than for the classical autoregressive fractionally integrated moving average, and leads to semi-long range dependence and transient anomalous behavior. We investigate ARTFIMA dependence structure with stable noise and construct Whittle estimators. Finally, we illustrate the usefulness of the ARTFIMA process on a trajectory from a single-particle tracking experiment.

EO723 Room Virtual R06 CAUSAL INFERENCE IN NETWORK SETTINGS Chair: Subhadeep Paul

E0387: **Identification and estimation of causal peer effects using double negative controls for unmeasured network confounding**
Presenter: **Naoki Egami**, Columbia University, United States
Co-authors: Eric Tchetgen Tchetgen

Scientists have been interested in estimating causal peer effects to understand how people's behaviors are affected by their network peers. However, it is well known that identification and estimation of causal peer effects are challenging in observational studies for two reasons. The first is the identification challenge due to unmeasured network confounding, for example, homophily bias and contextual confounding. The second issue is the network dependence of observations, which one must take into account for valid statistical inference. Negative control variables, also known as placebo variables, have been widely used in observational studies, including peer effect analysis over networks, although they have been used primarily for bias detection. We establish a formal framework which leverages a pair of negative control outcomes and exposure variables (double negative controls) to nonparametrically identify causal peer effects in presence of unmeasured network confounding. We then propose a generalized method of moments estimator for causal peer effects, and establish its consistency and asymptotic normality under an assumption about ψ-network dependence. Finally, we provide a network heteroskedasticity and autocorrelation consistent variance estimator. Our methods are illustrated with an application to peer effects in education.

E0453: **Network influence with latent homophily and measurement error**
Presenter: **Subhadeep Paul**, The Ohio State University, United States

Modeling social influence on outcomes of network-connected individuals is a central research question in several scientific disciplines. However, network influence cannot be identified from observational data because it is confounded with unobserved homophily. We propose a latent homophily-adjusted Spatial Autoregressive model (SAR) for networked responses to identify the causal contagion effects. The latent homophily is estimated from the spectral embedding of the network's adjacency matrix. We further develop maximum likelihood estimators for the parameters of the SAR model when covariates are measured with error. The bias-corrected MLE enjoys statistical consistency and asymptotic normality properties. We combine the estimated latent homophily with the bias-corrected MLE in the SAR model to estimate network influence. Our simulations show that the methods perform well in finite samples. Applying our methodology to a data set of female criminal offenders in a therapeutic community (TC), we provide causal estimates of network influence on graduation from the TC.

E0557: **HODOR: Hold-Out Design for Online A/B testing with lurking variables**
Presenter: **Srijan Sengupta**, North Carolina State University, United States
Co-authors: Nicholas Larsen, Jonathan Stallrich, Srijan Sengupta

A/B tests are common tools for estimating the average treatment effect in online controlled experiments (OCEs). Classical OCE theory is based on the Stable Unit Treatment Value Assumption, which holds that individual user responses are determined solely by the assigned treatment and not by the treatments of others. This assumption is violated when users are subjected to network interference, which is a common occurrence on social media platforms and other online testing platforms. Standard methods for estimating the average treatment effect typically fail to account for network effects, resulting in highly biased results. Furthermore, unobserved user covariates that influence user response and network structure, such as offline information or variables hidden due to privacy restrictions, can bias current estimators of the average treatment effect. The aim is to show how network-influential lurking variables can heavily bias popular network clustering-based methods, rendering them unreliable. To

address this issue, we propose HODOR (Hold-Out Design for Online Randomized experiments), a two-stage design and estimation technique. We demonstrate that HODOR is unbiased for the average treatment effect and has low variance. Remarkably, HODOR provides reliable estimation even when the underlying network is partially unknown or uncertain.

E0560: **Network interference in micro-randomized trials**
Presenter: **Shuangning Li**, Harvard University, United States
Co-authors: Stefan Wager

The micro-randomized trial (MRT) is an experimental design that can be used to develop optimal mobile health interventions. In MRTs, interventions in the form of notifications or messages are sent through smartphones to individuals, targeting a health-related outcome such as physical activity or weight management. Often, mobile health interventions have a social media component; an individual's outcome could thus depend on other individuals' treatments and outcomes. We study the micro-randomized trial in the presence of such cross-unit interference. We model the cross-unit interference with a network interference model; the outcome of one individual may affect the outcome of another individual if and only if they are connected by an edge in the network. Assuming the dynamics can be represented as a Markov decision process, we analyze the behavior of the outcomes in large sample asymptotics and show that they converge to a mean-field limit when the sample size goes to infinity. Based on the mean-field result, we give characterization results and estimation strategies for various causal estimands, including the short-term direct effect of a binary intervention, its long-term direct effect and its long-term total effect.

E1242: **Estimating the prevalence of peer effects and other spillovers**
Presenter: **David Choi**, Carnegie Mellon University, United States

In randomized experiments with arbitrary and unknown interference, we show that hypothesis tests for the sharp null of no effect can be inverted with no assumptions on interference, producing one-sided interval estimates (or lower bounds) – not for the treatment effect, but rather for the number of units who were affected by treatment. Similarly, tests for the null of no interference can be inverted with no assumptions beyond randomization to estimate the number of units that were affected by the treatment of others. This does not fully identify the treatment effect, but may be used to show that a peer effect exists, and to estimate whether it is widely prevalent.

EO158 Room Virtual R07 BAYESIAN SEMI- AND NON-PARAMETRIC METHODS II Chair: Andrea Cremaschi

E1828: **Unsupervised tree boosting for learning probability distributions**
Presenter: **Li Ma**, Duke University, United States
Co-authors: Naoki Awaya

An unsupervised tree boosting algorithm is proposed for inferring the underlying sampling distribution of an i.i.d. sample based on fitting additive tree ensembles in a fashion analogous to supervised tree boosting. Integral to the algorithm is a new notion of "addition" on probability distributions that leads to a coherent notion of "residualization", i.e., subtracting a probability distribution from an observation to remove the distributional structure from the sampling distribution of the latter. We show that these notions arise naturally for univariate distributions through cumulative distribution function (CDF) transforms and compositions due to several "group-like" properties of univariate CDFs. While the traditional multivariate CDF does not preserve these properties, a new definition of multivariate CDF can restore these properties, thereby allowing the notions of "addition" and "residualization" to be formulated for multivariate settings as well. This then gives rise to the unsupervised boosting algorithm based on forward-stagewise fitting of an additive tree ensemble, which sequentially reduces the Kullback-Leibler divergence from the truth. The algorithm allows the analytic evaluation of the fitted density and outputs a generative model that can be readily sampled from. We enhance the algorithm with scale-dependent shrinkage and a two-stage strategy that separately fits the marginals and the copula.

E0582: **Hierarchical and time-dependent random partitions based on the shrinkage partition distribution**
Presenter: **David Dahl**, Brigham Young University, United States
Co-authors: Richard Warr, Thomas Jensen

Bayesian nonparametric models often rely on clustering models that borrow strength within and between groups. In a scenario where researchers have some notion of the clustering composition, we propose the shrinkage partition distribution, which allows for tractable posterior analysis based on the researchers' prior knowledge. The shrinkage partition distribution (SPD) shrinks any baseline random partition distribution towards a baseline partition. An extension to any sequentially-allocated partition model, SPD is extremely flexible with relatively inexpensive posterior simulation. We show several distinct advantages over the existing methods, including the ability to model dependent random partitions. Specifically, we show that the SPD can hierarchically model a collection of random partition distributions and can also model time-dependent random partitions.

E0743: **Bayesian nonparametric marked Hawkes processes for earthquake modeling**
Presenter: **Athanasios Kottas**, University of California Santa Cruz, United States
Co-authors: Hyotae Kim

The Hawkes process is a versatile stochastic model for point processes that exhibit self-excitation, that is, the property that the occurrence of an event increases the rate of occurrence for some period of time in the future. Including extensions to incorporate marks, Hawkes processes have been successfully applied in several scientific areas. We will present a nonparametric Bayesian modeling approach for marked Hawkes processes. The prior models for the functions that define the point process conditional intensity are constructed to achieve flexible, computationally efficient inference, utilizing the Hawkes process branching structure. The motivating application involves earthquake data modeling, where the mark is given by the earthquake magnitude. The methodology builds from a prior for the Hawkes process excitation function that allows flexible shapes for mark-dependent offspring densities. In the context of the application, the modeling approach enables the estimation of aftershock densities that vary with the magnitude of the main shock, thus significantly expanding the inferential scope of existing self-exciting point process models for earthquake occurrences. The methodology will be studied empirically with data on earthquake occurrences from Japan.

E1917: **Bayesian pyramids: Identifiable multilayer discrete latent structure models for discrete data**
Presenter: **Yuqi Gu**, Columbia University, United States
Co-authors: David Dunson

High-dimensional categorical data are routinely collected in biomedical and social sciences. It is of great importance to building interpretable parsimonious models that perform dimension reduction and uncover meaningful latent structures from such discrete data. Identifiability is a fundamental requirement for valid modeling and inference in such scenarios, yet is challenging to address when there are complex latent structures. We propose a class of identifiable multilayer (potentially deep) discrete latent structure models for discrete data, termed Bayesian pyramids. We establish the identifiability of Bayesian pyramids by developing novel transparent conditions on the pyramid-shaped deep latent directed graph. The proposed identifiability conditions can ensure Bayesian posterior consistency under suitable priors. As an illustration, we consider the two-latent-layer model and propose a Bayesian shrinkage estimation approach. Simulation results for this model corroborate the identifiability and estimability of model parameters. Applications of the methodology to DNA nucleotide sequence data uncover useful discrete latent features that are highly predictive of sequence types. The proposed framework provides a recipe for interpretable unsupervised learning of discrete data, and can be a useful alternative to popular machine learning methods.

EO284 Room Virtual R08 RECENT DEVELOPMENTS IN OPTIMAL DESIGNS Chair: Saumen Mandal

E1611: **Bayesian and maximin A-optimal designs for spline regression models**
Presenter: **Julie Zhou**, University of Victoria, Canada
Co-authors: Isaac Rankin

Optimal designs for spline regression models with multiple unknown knots are studied using an A-optimality criterion. Locally A-optimal designs are constructed, which depend on the true values of the knots. However, in practice, the knots are never known exactly, but it may be reasonable to assume a prior distribution for these knots. Using the prior distribution, we apply a Bayesian or maximin efficiency criterion to construct optimal designs. Several theoretical results are derived. We also propose to use a numerical method for computing the optimal designs. The key to using the numerical method is to transform the design problems into convex optimization problems, and the method is very fast to compute optimal designs on discrete design spaces. Examples of Bayesian A-optimal and maximin efficiency optimal designs are presented.

E1835: **Bayesian analysis and follow-up experiments for supersaturated multistratum designs**
Presenter: **Po Yang**, University of Manitoba, Canada

Supersaturated multistratum designs are applied to identify important factors in experiments in which the run order cannot be completely randomized. Since supersaturated multistratum designs have small run sizes and large numbers of factors, problems of model uncertainty exist. A drawback of the stepwise regression analysis commonly used in the literature is that it only produces a single model and, thus, is not suitable for dealing with model uncertainty. We propose a Bayesian approach for analyzing the data collected from supersaturated multistratum designs. Instead of producing a single model, the Bayesian analysis reports several competing models and, thus, provides an opportunity for the experimenters to explore potentially important factors. To further reduce uncertainty, we suggest conducting follow-up experiments and develop a generalized model-discrimination criterion for selecting follow-up supersaturated designs that are effective in reducing ambiguity in the analysis results.

E1899: **Optimal designs for minimizing some correlation-based criteria**
Presenter: **Saumen Mandal**, University of Manitoba, Canada

Optimal design theory can be applied to solve a variety of optimization problems, which demand the calculation of one or more optimizing probability distributions. We consider one such problem based on some correlation-based criteria. In some regression models, it is desired to estimate certain parameters as independently of each other as possible. We construct such designs by minimizing the squared correlations between the estimators of parameters or linear combinations of the parameters of a linear model. This optimization problem is quite challenging as the criterion is neither convex nor concave. We first determine the optimality conditions for such criterion, and then construct optimal designs for some regression models. As a second problem, we construct optimal designs by minimizing the A-optimality criterion (average variance) subject to achieving zero correlation among certain parameter estimators. In this way, we can achieve two goals with one optimization problem. We achieve the goals by using the Lagrangian theory and then solving the problem by means of a simultaneous optimization technique. We consider some regression models of interest, and report the optimal designs.

E1837: **CV, ECV, and robust CV designs for replications under a class of linear models in factorial experiments**
Presenter: **Subir Ghosh**, University of California, United States

A class of linear models is considered for describing the data collected from an experiment so that any two models have some common and uncommon parameters. The uncommon parameters play a significant role in discriminating between any two models. We propose a common variance (CV) design for collecting the data so that all the uncommon parameters are estimated with as similar variances as possible in all models. We get the variance equality for a CV design when there is one uncommon parameter for any two models within the class. We introduce a new concept, Robust CV designs for replications, having the possibility of replicated observations. We present the conditions for a CV design having no replicated observations to be robust for general replicated observations. A CV design having no replicated observations is always robust for equally replicated observations. In the class of linear models considered for factorial experiments, the common parameters for all models correspond to the general mean and main effects, and the other parameters correspond to two-factor interactions. We present two general CV designs for three-level factorial experiments. We also present examples of Efficient CV (ECV) designs and Robust CV designs for general replicated observations.

EO685 Room BH (S) 2.05 ADVANCES IN BAYESIAN FACTOR ANALYSIS Chair: Pantelis Samartsidis

E0297: **Latent variable model for graph-estimation in multivariate stationary time series**
Presenter: **Arkaprava Roy**, University of Florida, United States

Vector-valued multivariate time series data are routinely collected in many application areas. Although stationarity, causality and invertibility are very useful modeling assumptions for any time series data, methodological developments are limited under these assumptions for multivariate time series. Under some assumptions on the autocovariance matrices, we achieve those properties for a new class of Gaussian multivariate time series. In this proposed class, the normalized multivariate time series is assumed to be some orthogonal rotation of a set of independent latent univariate time series. To capture the graphical dependence structure among the variables, we also propose to sparsely estimate the marginal precision matrix and develop related computational methodologies. An efficient Markov Chain Monte Carlo (MCMC) algorithm is developed for posterior computation. We also study theoretical consistency properties. We show excellent performance in simulations and GDP data applications.

E1177: **Structured factorization for single-cell gene expression data**
Presenter: **Lorenzo Schiavon**, University of Padova, Italy
Co-authors: Luisa Galtarossa, Lorenzo Schiavon, Antonio Canale, Davide Risso

Single-cell gene expression experiments yield count data characterised by both high dimensionality and high complexity, with tens of thousands of cells and genes. In this context, factorization models represent a powerful tool to condense the available information through a sparse decomposition into lower-rank matrices. We adapt and implement a recent Bayesian class of generalized factorization models to count data and, specifically, to model the covariance between genes. The developed methodology also allows one to include exogenous information about genes within the prior, such that recognition of covariance structures between similar genes is favoured. We use biological pathways as external information to induce sparsity patterns within the loadings matrix. This also helps to assign a meaning to the loadings columns and, as a consequence, to the corresponding latent factors, which can be interpreted as unobserved cell covariates. We apply the model to sc-RNAseq data, collected on lung adenocarcinoma cell lines, showing promising results about the role of the pathways in characterizing the relations between genes and extracting valuable insights about unobserved cell traits.

E1523: **Bayesian correlated clustering of multiple high-dimensional datasets using mixtures of factor analysers**
Presenter: **Johan van der Molen Moris**, University of Cambridge, United Kingdom

Factor analysis is commonly used to perform dimensionality reduction to pre-process data for clustering. More recently, Bayesian mixtures of factor analysers have been proposed to perform clustering and dimensionality reduction jointly in a more principled manner. We extend this model to the Bayesian integrative clustering setting where multiple high-dimensional datasets are clustered simultaneously, by using multiple dependent mixtures of factor analysers. This is particularly relevant in molecular precision medicine, where we would like to identify disease subtypes on the basis of multiple omics data layers. Each observation unit (e.g. patients) has data coming from multiple high-dimensional data sources, such as gene or protein expression. These data sources provide complementary information, and thus it is critical to integrate them into the analysis. However, Bayesian model-based clustering using high-dimensional data presents challenges even in the single-dataset case, due to

difficulties exploring multiple posterior modes, and these challenges are exacerbated when we are dealing with multiple high-dimensional datasets. We present a Bayesian correlated clustering mixture of factor analysers model for addressing these points, and further elucidate the challenges of performing MCMC-based parameter inference in such models.

E1767: **Bayesian cross-study models: From epidemiological to genomics applications**
Presenter: **Roberta De Vito**, Brown University, United States

One of the most important challenges in biological sciences today is to elucidate how complex experiments, which measure hundreds of thousands of variables, generate consistent answers when repeated; and how these answers can be learned by analyzing several studies together. We start from the premise that genuine biological patterns are more likely than spurious patterns to be consistently present in multiple studies. Our challenge is to systematically and reliably identify the consistent biological patterns shared among studies and remove variation that lacks such reproducibility. To meet this challenge, we propose a novel analytical concept by upgrading a very widely used statistical technique known as factor analysis. Our Bayesian Multi-study Factor model is able to handle multiple studies simultaneously in a Bayesian shrinkage prior approach by using a Gibbs Sampling algorithm for parameter estimates. We present several different biological: nutritional epidemiological data in seven different countries, microarray gene expression in 4 different studies, and 12 brain regions in tissue studies.

E1960: **Multi-study factor regression models for heterogenous data: Applications to cancer genomics and nutritional epidemiology**
Presenter: **Alejandra Avalos Pacheco**, Vienna University of Technology, Austria
Co-authors: David Rossell, Roberta De Vito, Jack Jewson, Richard S Savage

Data integration of multiple studies can be key to understanding and gaining knowledge in statistical research. However, such data present both biological and artifactual sources of variation, also known as covariate effects. Covariate effects can be complex, leading to systematic biases. We will present novel sparse latent factor regression (FR) and multi-study factor regression (MSFR) models to integrate such heterogeneous data. The FR model provides a tool for data exploration via dimensionality reduction and sparse low-rank covariance estimation while correcting for a range of covariate effects. MSFR are extensions of FR that enable us to jointly (i) capture common components across studies, (ii) isolate the sources of variation that are unique to each study, and (iii) correct for non-biological sources of variation. We will discuss the use of several sparse priors (local and non-local) to learn the dimension of the latent factors. The approach provides a flexible methodology for sparse factor regression, which is not limited to data with covariate effects. We will present several examples, with a focus on bioinformatics applications. The results show an increase in the accuracy of the dimensionality reduction, with non-local priors substantially improving the reconstruction of factor cardinality. The results of our analyses illustrate how failing to account for covariate effects properly can result in unreliable inference.

EO214 Room K2.31 (Nash Lec. Theatre) MODERN METHODS FOR CAUSAL INFERENCE Chair: Luke Keele

E0272: **An automated solution to causal inference in discrete settings**
Presenter: **Dean Knox**, UPenn Wharton, United States

When causal quantities cannot be point-identified, researchers often pursue partial identification to quantify the range of possible values. However, the peculiarities of applied research conditions can make this analytically intractable. We present a general and automated approach to causal inference in discrete settings. We show causal questions with discrete data reduce to polynomial programming problems, and we present an algorithm to automatically bound causal effects using efficient dual relaxation and spatial branch-and-bound techniques. The user declares an estimand, states assumptions, and provides data (however incomplete or mismeasured). The algorithm then searches over admissible data-generating processes and outputs the most precise possible range consistent with available information -i.e., sharp bounds-including a point-identified solution if one exists. Because this search can be computationally intensive, our procedure reports and continually refines non-sharp ranges that are guaranteed to contain the truth at all times, even when the algorithm is not run to completion. Moreover, it offers an additional guarantee we refer to as ε-sharpness, characterizing the worst-case looseness of the incomplete bounds. Analytically validated simulations show the algorithm accommodates classic obstacles, including confounding, selection, measurement error, noncompliance, and nonresponse.

E0495: **A devils bargain? Repairing a difference in differences parallel trends assumption with an initial matching step**
Presenter: **Luke Miratrix**, Harvard University, United States
Co-authors: Dae Woong Ham

The Difference in Difference (DiD) estimator is a popular estimator built on the "parallel trends" assumption that the treatment group, absent treatment, would change "similarly" to the control group over time. To increase the plausibility of this assumption, a natural idea is to match treated and control units prior to a DiD analysis. We characterize the bias of such matching under a class of linear structural models with both observed and unobserved confounders with time-varying effects. We find matching on baseline covariates generally reduces the bias associated with these covariates. We further find that additionally matching on pre-treatment outcomes has both cost and benefit. First, matching on pre-treatment outcomes will partially balance unobserved confounders, which mitigates some bias. This reduction is proportional to the outcome's reliability, a measure of how coupled the outcomes are with the latent covariates. Matching on pre-treatment outcomes also undermines the second "difference" in a DiD estimate. This injects bias into the final estimate, creating a bias-bias tradeoff. We extend our results to multivariate confounders with multiple pre-treatment periods and find similar results. We summarize with heuristic guidelines on whether to match prior to a DiD analysis, along with a rough bias reduction estimator. We illustrate our guidelines by reanalyzing a recent empirical study of principal turnover on student achievement.

E0535: **Approximate balancing weights for clustered observational studies**
Presenter: **Luke Keele**, University of Pennsylvania, United States

Many interventions are in settings where treatments are applied to groups. For example, a reading intervention may be implemented for all students in some schools and withheld from students in other schools. When such treatments are non-randomly allocated, outcomes across the treated and control groups may differ due to the treatment or due to baseline differences between groups. When this is the case, researchers can use statistical adjustment to make treated and control groups similar in terms of observed characteristics. Recent work in statistics has developed matching methods designed for contexts where treatments are clustered. In this article, we develop an alternative to multilevel matching based on approximate balancing weights. Moreover, these methods can automatically balance interactions between group and unit-level covariates. Using simulations, we show that weighting estimators based on these methods outperform matching in terms of both balance and root-mean-squared error. We conclude with two empirical applications.

E1391: **Minimax optimal subgroup identification**
Presenter: **Matteo Bonvini**, Carnegie Mellon University, United States
Co-authors: Edward Kennedy, Luke Keele

Quantifying treatment effect heterogeneity is a crucial task in many areas of causal inference, e.g. optimal treatment allocation and estimation of subgroup effects. We study the problem of estimating the level sets of the conditional average treatment effect (CATE), identified under the no-unmeasured-confounders assumption. Given a user-specified threshold, the goal is to estimate the set of all units for whom the treatment effect exceeds that threshold. The estimator that we study follows the plug-in principle and consists of simply thresholding a good estimator of the CATE. While many CATE estimators have been recently proposed and analysed, how their properties relate to those of the corresponding level set estimators remains unclear. Our first goal is thus to fill this gap by deriving the asymptotic properties of level set estimators depending on which estimator of the CATE is used. Next, we identify a minimax optimal estimator in a model where the CATE, the propensity score and the outcome

model are Holder-smooth of varying orders. We consider data-generating processes that satisfy a margin condition governing the probability of observing units for whom the CATE is close to the threshold. We investigate the performance of the estimators in simulations and illustrate our methods on a dataset from REFLUX, a multi-center study that aimed to compare the effectiveness of surgery to treat Gastro-Oseophageal Reflux Disease.

E1942: **Policy learning with asymmetric utilities**
Presenter: **Eli Ben-Michael**, Carnegie Mellon University, United States
Co-authors: Kosuke Imai, Zhichao Jiang

Data-driven decision-making plays an important role in high-stakes settings like medicine and public policy. Learning optimal policies from observed data requires a careful formulation of the utility function whose expected value is maximized across a population. Although researchers typically use utilities that depend on observed outcomes alone, in many settings, the decision maker's utility function is more properly characterized by the joint set of potential outcomes under all actions. For example, the Hippocratic principle to "do no harm" implies that the cost of causing death to a patient who would otherwise survive without treatment is greater than the cost of forgoing life-saving treatment. We consider optimal policy learning with asymmetric utility functions of this form. We show that asymmetric utilities lead to an unidentifiable social welfare function, and so we first partially identify it. Drawing on statistical decision theory, we then derive minimax decision rules by minimizing the maximum regret relative to alternative policies. We show that one can learn minimax decision rules from observed data by solving intermediate classification problems. We also establish that the finite sample regret of this procedure is bounded by the mis-classification rate of these intermediate classifiers. We apply this conceptual framework and methodology to the decision about whether or not to use right heart catheterization for patients with possible pulmonary hypertension.

EO268 Room K2.40 NEW FRONTIERS IN COMPLEX AND FUNCTIONAL DATA ANALYSES (VIRTUAL) Chair: Kuang-Yao Lee

E1798: **Feature representation learning in neural networks guided by the sliced inverse regression**
Presenter: **Su-Yun Huang**, Academia Sinica, Taiwan
Co-authors: Yan-Bin Chen

Due to demands from recognition of digital biomedical images, a lot of interest is being paid to finding feature representations in neural networks. We will share a neural network-based feature representation learning method guided by the notion of Sliced Inverse Regression (SIR). The proposed method works on deep neural networks with the SIR as a neural network layer. The proposed SIR layer is expected to improve image classification accuracy. We will present some numerical examples to demonstrate the usage of SIR-guided feature representation learning.

E1376: **Bayesian spatially varying coefficient models with functional predictors**
Presenter: **Yehua Li**, University of California at Riverside, United States

Reliable prediction for crop yield is crucial for economic planning, food security monitoring, and agricultural risk management. The aim is to develop a crop yield forecasting model at large spatial scales using meteorological variables closely related to crop growth. The influence of climate patterns on agricultural productivity can be spatially inhomogeneous due to local soil and environmental conditions. We propose a Bayesian spatially varying functional model (BSVFM) to predict county-level corn yield for five Midwestern states, based on annual precipitation and daily maximum and minimum temperature trajectories modeled as multivariate functional predictors. The proposed model accommodates spatial correlation and measurement errors of functional predictors, and respects the spatially heterogeneous relationship between the response and associated predictors by allowing the functional coefficients to vary over space. The model also incorporates a Bayesian variable selection device to expand its capacity to accommodate spatial heterogeneity further. The proposed method is demonstrated to outperform other highly competitive methods in corn yield prediction, owing to the flexibility of allowing spatial heterogeneity with spatially varying coefficients in our model. Our study provides further insights into understanding the impact of climate change on crop yield.

E0518: **Ultrahigh dimensional variable selection for Bayesian mixed type multivariate generalized linear models**
Presenter: **Hsin-Hsiung Huang**, University of Central Florida, United States
Co-authors: Shao-Hsuan Wang, Ray Bai

In recent years, the literature on Bayesian high-dimensional variable selection has rapidly grown. It is increasingly important to understand whether these Bayesian methods can consistently estimate the model parameters. To this end, we develop a multivariate Bayesian model with shrinkage priors (MBSP) model to mixed-type response generalized linear models (MRGLMs), and we consider a latent multivariate linear regression model associated with the observable mixed-type response vector through its link function. Under our proposed model (MBSP-GLM), multiple responses belonging to the exponential family are simultaneously modeled and mixed-type responses are allowed. We show that the MBSP-GLM model achieves strong posterior consistency when p grows at a subexponential rate with n. Furthermore, we quantify the posterior contraction rate at which the posterior shrinks around the true regression coefficients and allow the dimension of the responses q to grow as n grows. This greatly expands the scope of the MBSP model to include response variables of many data types, including binary and count data. To address the non-conjugacy concern, we propose an adaptive sampling algorithm via a Polya-gamma data augmentation scheme for the MRGLM estimation. We provide simulation studies and real data examples.

E1674: **Learning healthcare delivery network with longitudinal electronic health records data**
Presenter: **Jiehuan Sun**, University of Illinois at Chicago, United States

Knowledge networks such as the healthcare delivery network (HDN), describing relationships among different medical encounters, are useful summaries of state-of-art medical knowledge. The increasing availability of longitudinal electronic health records (EHR) data promises a rich data source for learning HDN. Most existing methods for inferring knowledge networks are based on co-occurrence patterns that do not account for temporal effects or patient-level heterogeneity. Building upon the multivariate Hawkes process (mvHP), we propose a flexible covariate-adjusted random effects (CARE) mvHP modeling strategy for HDN construction. Our model allows for patient-specific time-varying background intensity functions via random effects, which can also adjust for effects of important covariates. We adopt a penalized approach to select fixed effects, yielding a sparse network structure, and removing unnecessary random effects from the model. Through extensive simulation studies, we show that our proposed method performs well in recovering the network structure and that it is essential to account for patient heterogeneities. We further illustrate our CARE mvHP method to an EHR study of type 2 diabetes patients to learn an HDN for these patients and demonstrate that our results are consistent with current clinical practice in healthcare systems.

E1829: **Functional-directed acyclic graphs**
Presenter: **Kuang-Yao Lee**, Temple University, United States
Co-authors: Lexin Li, Bing Li

A new method is introduced to estimate a directed acyclic graph (DAG) from multivariate functional data. We build on the notion of faithfulness that relates a DAG with a set of conditional independences among the random functions. We develop two linear operators, the conditional covariance operator and the partial correlation operator, to characterize and evaluate the conditional independence. Based on these operators, we adapt and extend the PC-algorithm to estimate the functional directed graph, so that the computation time depends on the sparsity rather than the full size of the graph. We study the asymptotic properties of the two operators, derive their uniform convergence rates, and establish the uniform consistency of the estimated graph, all of which are obtained while allowing the graph size to diverge to infinity with the sample size. We demonstrate the

efficacy of our method through both simulations and an application to a time-course proteomic dataset.

EO575 Room K2.41 METHODS AND ALGORITHMS IN CONTEMPORARY DATA ANALYSIS Chair: Zhihua Su

E1733: **Measurement error in linear regression models with fat tails and skewed errors**
Presenter: **Jiyoun Myung**, California State University, East Bay, United States
Co-authors: Mahmoud Torabi, Malay Ghosh, Mark Steel

Linear regression models, which account for skewed error distributions with fat tails, have been previously studied and often observed in real data analyses. Covariates measured with error also happen frequently in the observational data set-up. As a motivating example, wind speed as a covariate is usually used, among other covariates, to estimate particulate matter, which is one of the most critical air pollutants and has a major impact on human health and the environment. However, the wind speed is measured with error, and the distribution of particulate matter is neither symmetric nor normally distributed. Ignoring the issue of measurement error in covariates may produce bias in model parameters estimate and lead to wrong conclusions. A hierarchical Bayesian approach is implemented for properly studying linear regression models where the covariates are measured with error and error distribution is skewed with fat tails. The performance of the proposed approach is evaluated through a simulation study and also by a real data application.

E1867: **fastkqr: A fast algorithm for kernel quantile regression**
Presenter: **Boxiang Wang**, University of Iowa, United States
Co-authors: Qian Tang, Yuwen Gu

Quantile regression is a powerful tool to model certain quantiles of the response and has been widely used in many application areas, including economics, finance, social sciences, and engineering, among others. The computation of quantile regression is typically expensive due to its nonsmooth loss function. We propose a major advance to the computation of quantile regression in reproducing kernel Hilbert spaces. We develop a novel and efficient algorithm called fastkqr for computing the exact solution path of kernel quantile regression. To improve the computation speed, we develop a fast implementation strategy to carefully reuse the matrix computations in fastkqr. Extensive simulation studies and benchmark applications demonstrate orders of magnitude speedup of fastkqr over the existing algorithms of quantile regression with almost the same algorithm accuracy.

E1354: **Two-stage adaptive designs of single arm clinical trials based on median event time test**
Presenter: **Yeonhee Park**, University of Wisconsin, United States

The Phase II clinical trials aim to assess the therapeutic efficacy of a new drug. The therapeutic efficacy has been often quantified by response rate such as overall response rate or survival probability in the Phase II setting. However, there is a strong desire to use survival time, which is the gold standard endpoint for the confirmatory Phase III study, when investigators set the primary objective of the Phase II study and test hypotheses based on the median survivals. We propose a method for median event time test (METT) to provide the sample size calculation and decision rule of testing. The decision rule is simple and straightforward in that it compares the observed median event time to the identified threshold. Moreover, it is extended to optimal two-stage design for practice, which extends the idea of Simons optimal two-stage design for survival endpoint. We investigate the performance of the proposed methods through simulation studies. The proposed methods are applied to redesign a trial based on median event time for trial illustration, and practical strategies are given for application of proposed methods.

E1879: **An augmented likelihood approach incorporating error-prone auxiliary data into a survival analysis**
Presenter: **Noorie Hyun**, Medical College of Wisconsin, United States
Co-authors: Pamela Shaw

Substantial clinical data collection is available from large healthcare community studies or electronic health records (EHR) in health systems. However, data accuracy can vary according to measurement methods. For example, self-reported medical history can include bias, such as recall bias or response bias. In contrast, biomarkers from a laboratory test are less likely to be biased. We are motivated to study what benefit we can gain by augmenting error-prone self-reported and biomarker-based disease diagnoses in regression for time-to-disease onset. The proposed model addresses left-truncation and interval-censoring in time-to-disease onset outcomes. Also, self-reported disease diagnosis errors are corrected using sensitivity and specificity parameters in the joint likelihood. Compared to other models using biomarker or self-reported data, comprehensive simulation studies found appealing finite sample properties of the proposed augmenting model, including the smallest mean square error. The proposed model is applied to the Hispanic Community Health Study/ Study of Latino data to quantify risk factors associated with diabetes onset.

E1921: **Nonlinear envelope model for nonparametric regression**
Presenter: **Zhihua Su**, University of Florida, United States

An envelope model is introduced for parsimonious nonparametric multivariate regression, where the regression function is assumed to lie in a reproducing kernel Hilbert space. This model extends earlier works on envelope models from the linear to the nonlinear case. The conditional independence relations offered by the envelope structure allow us to effectively reduce the dimensions of both the predictor and response while performing the regression. Along with the estimation procedure, we also developed inference tools to construct confidence intervals, prediction intervals, and for conducting hypothesis test based on the asymptotic distribution of the envelope estimate. Simulation studies show that our model achieves substantial efficiency gains compared with standard nonparametric regression, principal component regression, and partial least squares. We applied our method to two data sets in chemometrics and breast cancer applications.

EC384 Room S-1.04 STATISTICAL MODELLING II Chair: Jochen Einbeck

E0388: **Nonparametric predictive inference for multiple future ordinal observations**
Presenter: **Abdulmajeed Alharbi**, Durham University, United Kingdom

In the theory of classical probability, a single precise probability satisfying Kolmogorov's axioms is used to quantify uncertainty. Imprecise probability in uncertainty quantification constitutes an appropriate and more general alternative approach when the information is incomplete or vague. Nonparametric Predictive Inference (NPI) is one of the statistical methodologies that have been developed to quantify uncertainty using imprecise probabilities, and it is based only on an exchangeability assumption for future and past observations. NPI has been developed for ordinal data, which are a type of categorical data with ordered categories. Examples of such data include pain level, satisfaction rating and education level. NPI methods for ordinal data use an assumed underlying data representation, with latent variables on the real-line falling into intervals which represent the categories. NPI for ordinal data was developed with attention restricted to a single future observation; our aim is to generalise this to multiple future observations and develop statistical methods based on this. Lower and upper probabilities for events involving multiple ordinal future observations are presented. This development forms the basis for a range of possible applications to be considered later, such as an application to the reproducibility of statistical tests for ordinal data.

E1537: **Parametric predictive bootstrap**
Presenter: **Abdulrahman Aldawsari**, Durham University, United Kingdom

Bootstrap methods are used to quantify the uncertainty of sample estimates, they have been applied to a wide range of statistical problems due to their simplicity and efficiency in giving good estimates. There are two main bootstrap methods: parametric and nonparametric. The parametric bootstrap method uses available data to estimate the parameters of the assumed distribution and then generates a number of parametric bootstrap

samples from the assumed distribution with the estimated parameters. A new bootstrap method is proposed, especially for predictive inference, the parametric predictive bootstrap, which does use an assumed parametric model. It is evaluated in a variety of scenarios that have been used with other bootstrap methods in order to investigate its performance in estimation and prediction inference. The proposed bootstrap method is compared to different types of bootstrap methods in terms of the coverage probability through simulations. Confidence intervals and prediction intervals based on the bootstrap technique are used to examine the PP-B's performance in estimation and prediction inference. The explicitly predictive nature of PP-B provides good performance for predictive inference.

E1574: **Nonparametric predictive inference for 2x2 contingency tables**
Presenter: **Reid Alotaibi**, Durham University, United Kingdom

Nonparametric predictive inference (NPI) is a frequentist statistics approach that makes few assumptions, enabled by using lower and upper probabilities to quantify uncertainty, and explicitly focuses on future observations. NPI has been developed for a range of data types, and for a variety of applications and problems in statistics. In statistics, data in the form of so-called contingency tables occur in many applications. Such tables show the distribution of one variable in rows and another variable in columns, and are typically used to study the dependence between the two variables (with generalization to more than two variables). The most basic version is the 2x2 table, where both variables are binary. In addition to studying the dependence, one may be interested in further inference based on such data. NPI method for such data has been developed where the inferences are restricted to only one future observation; however, considering more general inferences about multiple future observations for a 2x2 contingency which will be briefly presented.

E1627: **A generalised normal distribution with interpretable parameters for location, body-shape, skewness, and tail-weight**
Presenter: **Matthias Wagener**, University of Pretoria, South Africa
Co-authors: Andriette Bekker, Mohammad Arashi

The multivariate normal distribution is a foundational model in statistics. There are many generalised forms for modelling non-normal and irregular data. However, very few of these generalisations have shape parameters with clear roles that determine, for instance, skewness and tail shape. Here, we add a skewness parameter for the body-tail generalised normal distribution, which yields the flexible and interpretable normal distribution FIN with parameters for location, scale, body-shape, skewness, and tail weight. Basic statistical properties of the FIN are provided, such as the density function, cumulative density function, moments, and likelihood equations. The FIN density is extended to a multivariate setting using a student t-copula, yielding the multivariate FIN distribution MFIN. The MFIN is applied to stock returns data where it is compared to the t-copula multivariate sinh-arcsinh, skew-t, and hyperbolic distribution.

E2001: **Small-sample test for comparing Cohen's d effect sizes between groups**
Presenter: **Florin Vaida**, University of California San Diego, United States
Co-authors: Anya Umlauf

The fosuc is on comparing treatment effects between two or more different studies, or populations. Two instances when this question arises are meta-analysis, when determining whether treatment effects are homogeneous or heterogeneous over the studies; and neurocognitive testing, when comparing group effects (e.g., the effect of sex or race) between two different populations. We assume that for each study, the treatment effect is measured by Cohen's d, which is the difference between the means in the treatment and control groups relative to the common standard deviation. An approximately unbiased estimator of this effect is given by Hedges's g. When comparing two studies, a large-sample approximation to the distribution of the difference of Hedges's g, g1-g2, is currently used. We show that a finite-sample approximation for the distribution of this statistic, suitably modified, can be obtained. This is done using second-order expansions. The improvement of the proposed method is demonstrated via statistical simulations. The use is illustrated in an application from neurocognitive research in HIV.

EC827 Room K0.19 APPLIED STATISTICS Chair: Bojana Milosevic

E1480: **Probabilistic forecasting of weather-driven faults on electricity distribution networks**
Presenter: **Daniela Castro-Camilo**, University of Glasgow, United Kingdom
Co-authors: Jethro Browell

Electricity networks are exposed to the weather, and severe weather may cause faults that result in power cuts. Predicting the occurrence of faults in a region on time scales from hours to days ahead can increase preparedness and accelerate the response to weather-related faults, and ultimately reduce the duration of power cuts. Furthermore, these predictions should quantify uncertainty so that planners may assess risk and distribute limited resources accordingly. We present a method for probabilistic fault prediction that leverages ensemble numerical weather prediction and methods from extreme value theory for discrete processes. Data describing network topology and vulnerability, such as elevation and proximity to vegetation, are combined with meteorological data to model the occurrence of faults, which is stochastic and may be heavy-tailed. In addition, forecasts of future weather conditions are required, and associated uncertainty is quantified via ensemble numerical weather prediction, which requires statistical post-processing. Finally, we discuss the communication of the resulting complex forecast information to decision-makers.

E1998: **Predicting biodiversity with the generalised functional response model**
Presenter: **Shaykhah Aldossari**, University of Glasgow, United Kingdom
Co-authors: Dirk Husmeier, Jason Matthiopoulos

Biodiversity is a measure of variability and is widely used to describe the variation in different fields. The Shannon entropy score is the most frequently used measure of biodiversity in ecology. It summarises the information about species abundance within a sample or a group. We use the Shannon entropy score to investigate three different things. First, we use the entropy score to assess the transferability of the generalized function response (GFR) model by measuring the information content in the dataset under study. Second, we observe the relationship between biodiversity and land cover types using the GFR model and various recent extensions. Finally, we investigate the legacy effect using the GFR models of land cover types on biodiversity. The large-scale North American Breeding Bird Survey (BBS) dataset was used for this purpose. We discuss how the information in the dataset affects the predictive ability of the model. Our finding is that our recent extensions of the GFR model double the biodiversity prediction accuracy compared to the standard generalised linear model (GLM). Moreover, biodiversity in the BBS dataset is found not to occur with time lags in response to land cover covariates using the GFR models.

E1817: **Bivariate modelling of rainfalls and temperature**
Presenter: **Giovanni De Luca**, University of Naples Parthenope, Italy
Co-authors: Giorgia Rivieccio

In the context of the undeniable ongoing climate change, rainfall is one of the main factors causing uncertainty in crops and damage to civil structures. Crops are heavily penalized by extreme climatic events that induce anomalous rainfall. Civilian structures, such as bridges and roads, can be hit and, in turn, cause massive damage and death. The study of rainfalls aimed at creating an early warning system cannot ignore the relationship with the temperature. The analysis of the interdependence between these two variables is crucial and effective modeling is therefore required. The use of copula functions is relevant since it is a flexible tool since the marginal distributions are not constrained by the bivariate distribution. A copula-based bivariate model for daily rainfalls and temperatures in some southern Italian cities is analyzed.

E0219: **Doubly enhanced medicaid partnership annuities (DEMPANs): Long-term care for seniors in the Medicaid penumbra**
Presenter: **Colin Ramsay**, University of Nebraska-Lincoln, United States

A 2019 NAIC long-term care insurance guide estimates about 70% of US residents age 65 are expected to need long-term care, and about 35% are expected to enter a nursing home at least once in their lifetime. U.S. retirees often can access long-term care services via Medicaid, which is a means-tested program geared to lower-income Americans. But, to quickly qualify for Medicaid, many retirees may transfer their assets to family members and incur Medicaid penalties. To improve access to long-term care, most U.S. States developed Partnership for Long-Term Care Program insurance policies that provide access to Medicaid while sheltering some or all of a retiree's assets. We propose a doubly enhanced Medicaid Partnership annuity (DEMPAN) that combines an annuity with long-term care insurance that is integrated within the framework of a qualified Partnership policy. We use a multi-state model of long-term care with health states based on a retiree's ability to perform activities of daily living (ADLs), instrumental activities of daily living (IADLs), and cognitive ability. We explicitly assume the quality of long-term care affects the transition probabilities used in the multi-state model.

E0545: **Calibration of wind speed ensemble forecasts using truncated GEV based EMOS approach**
Presenter: **Marianna Szabo**, University of Debrecen, Hungary
Co-authors: Sandor Baran, Patricia Szokol

Probabilistic ensemble weather forecasting is an operatively used method of prediction at all major weather prediction centres. These forecasts are obtained from multiple runs of numerical weather prediction models with different initial conditions or model parametrizations. However, to account for the under-dispersive or biased nature of the ensemble forecasts, some kind of post-processing is applied. One of the most popular parametric statistical post-processing techniques is the ensemble model output statistics (EMOS), which provides a full predictive distribution of the weather quantity. We propose a novel EMOS model for calibrating wind speed ensemble forecasts, where the predictive distribution is a generalized extreme value (GEV) distribution left truncated at zero (TGEV). The truncation corrects the disadvantage of the GEV distribution-based models occasionally predicting negative wind speed values, without affecting its favourable properties. The new model is tested on four data sets of wind speed ensemble forecasts provided by three different ensemble prediction systems, covering various geographical domains and time periods. The forecast skill of the TGEV EMOS model is compared with the predictive performance of the truncated normal, log-normal and GEV methods and the raw and climatological forecasts as well. The results confirm the favourable properties of the novel TGEV EMOS approach.

EC817 Room BH (S) 2.03 BAYESIAN STATISTICS I Chair: Pier Giovanni Bissiri

E0457: **Approximate Bayesian inference in epidemic models: A focus on nowcasting and the time-varying reproduction number**
Presenter: **Oswaldo Gressani**, Hasselt University, Belgium

In epidemiology, mathematical models play a determinant role in the analysis of infectious disease data. Statistical methods and their underlying algorithms form the core backbone to compute estimates of key epidemiological parameters and to quantify their associated uncertainty, thereby providing a robust toolbox to understand the disease dynamics resulting from the transmission of a pathogen in a population. When inference is carried out under the Bayesian paradigm, Markov chain Monte Carlo (MCMC) methods often require a large computational budget resulting in a prohibitively slow estimation process. Building upon the synergy between Laplace approximations and P-splines, a flexible methodology is proposed as a lightning-fast alternative to simulation-based methods. This new toolbox is illustrated in the context of nowcasting (i.e. the real-time assessment of the current epidemic situation corrected for imperfect data information caused by delays in reporting) and in the recently proposed EpiLPS framework for estimating the time-varying reproduction number with applications on data of SARS-CoV-2.

E1765: **Type I Tobit Bayesian additive regression trees for censored outcome regression**
Presenter: **Eoghan O Neill**, Erasmus University Rotterdam, Netherlands

Type 1 Tobit Bayesian Additive Regression Trees (TOBART-1) are introduced. Simulation results and applications to real datasets demonstrate that TOBART-1 produces more accurate predictions than competing methods. TOBART-1 provides posterior probabilities of censoring, posterior intervals for the conditional expectation, and estimates of heterogeneous treatment effects. TOBART-1 is also combined with a Dirichlet Process mixture of normal distributions to provide a fully nonparametric censored outcome regression method (TOBART-1-NP).

E1825: **Smaller p-values in genomics studies using distilled auxiliary information**
Presenter: **Jordan Bryan**, Duke University, United States
Co-authors: Peter Hoff

Medical research institutions have generated massive amounts of biological data by genetically profiling hundreds of cancer cell lines. In parallel, academic biology labs have conducted genetic screens on small numbers of cancer cell lines under custom experimental conditions. In order to share information between these two approaches to scientific discovery, a "frequentist assisted by Bayes" (FAB) procedure is proposed for hypothesis testing that allows auxiliary information from massive genomics datasets to increase the power of hypothesis tests in specialized studies. The exchange of information takes place through a novel probability model for multimodal genomics data, which distills auxiliary information pertaining to cancer cell lines and genes across a wide variety of experimental contexts. If the relevance of the auxiliary information to a given study is high, then the resulting FAB tests can be more powerful than the corresponding classical tests. If the relevance is low, then the FAB tests yield as many discoveries as the classical tests. Simulations and practical investigations demonstrate that the FAB testing procedure can increase the number of effects discovered in genomics studies while still maintaining strict control of type I error and false discovery rate.

E1932: **Fisher *s* noncentral hypergeometric distribution for population size estimation**
Presenter: **Veronica Ballerini**, University of Florence, Italy
Co-authors: Brunero Liseo

A method is introduced to make inferences on the subgroups sizes of a heterogeneous population using survey data, even in the presence of a single list. To this aim, we use Fisher s noncentral hypergeometric distribution, which allows us to account for the possibility that capture heterogeneity is related to key survey variables. We propose a Bayesian approach for estimating the population sizes posterior distributions, exploiting both extra-experimental information, e.g., coming from administrative data, and the computational efficiency of MCMC and ABC methods. The motivating case study deals with the size estimation of the population of Italian youngsters who are not employed one year after graduating by gender and degree program. We account for the possibility that surveys' response rates differ according to individuals' employment status, implying a not-at-random missing data scenario. We find that employed persons are generally more inclined to answer the questionnaire; this behavior might imply the overestimation of the employment rate.

EP027 Room Posters Virtual Room 1 POSTER SESSION II (ONLY VIRTUAL) Chair: Cristian Gatu

E1667: **Goodness-of-fit tests for variance function in regression models**
Presenter: **Sandie Ferrigno**, INRIA Nancy and University Nancy Lorraine, France
Co-authors: Marie-Jose Martinez

Many goodness-of-fit tests have been developed to assess the different assumptions of a regression model. Most of them are "directional" in that they detect departures from a given assumption of the model. Other tests are "global" in that they assess whether a model fits a dataset on all its assumptions. We focus on the task of choosing the structural part of the variance function in the (possibly heteroscedastic) regression model. We consider two nonparametric "directional" tests and one nonparametric "global" test, all based on generalizations of the Cramer-von Mises statistic.

A simulation study is carried out to compare the three test methods in terms of statistical significance and power function. The implementation of such statistical tests requires the use of wild bootstrap methods.

E2012: **Skewed normal classification in high-dimensional data**
Presenter: **Haesong Choi**, Florida state university, United States

A considerable number of studies have been devoted to high-dimensional classification models under the assumption of normality. However, it tends to be restrictive in applications. Data transformation is one alternative way, but it may affect the distinctive characteristics of the original data. Motivated by the data set that exhibits asymmetry, including environmental, financial, and biomedical ones, we propose a high-dimensional discriminant analysis model called the SKNC model (short for SKewed Normal Classification). By incorporating the skewed normal model, the SKNC model inherits all properties of the normal distribution and improves its flexibility on skewed data in classification. Theoretical results rigorously show that the SKNC model achieves variable selection, penalized estimation, and prediction consistency, especially in high-dimensional settings. We empirically demonstrate the superior performance of the SKNC model over existing methods in simulated and real datasets.

C1539: **The effect of microfinance on countries and individuals**
Presenter: **Joseph Tarrant**, Rose-Hulman Institute of Technology, United States

The concept of microfinance goes back to the 1970s, though its use has increased over the past twenty years. Many have heralded microfinance as a way for people in poverty to have the liquidity needed to start businesses or to keep from reaching a point of desperation in their personal finances. Many claim that microfinance allows the poor a way out of the generational poverty they have experienced. Yet the rates on microfinance loans are often exorbitant. Can there be proof that microfinance helps individuals? Does microfinance lead to a smaller shadow economy as individuals can move to the formal economy? Can it be seen to help national economies? We look at the effect of growing microfinance on country's GDP, on the size of the underground economy, and on Gini coefficients.

E1588: **Modeling flexible trajectories and related outcomes using a three-level enriched Dirichlet process mixture**
Presenter: **Natalie Burns**, University of Florida, United States
Co-authors: Michael Daniels, Elizabeth Widen

Dirichlet process mixture (DPM) models can be used for density estimation and clustering. When jointly modeling a response and covariates, the enriched Dirichlet process mixture (EDPM) overcomes the possibility that the covariate measurements will dominate the response measurements in the clustering structure induced by the DPM. A further extension of the EDPM is proposed, the three-level EDPM (EDP3) with flexible trajectories. In the motivating example, there are three sets of variables: many measurements of gestational weight gain (GWG); several neonatal size outcomes; and a number of other maternal variables. The EDP3 induces a three-level nested clustering structure on the data. Nesting the clustering of the neonatal outcomes within the top-level GWG trajectory clusters, combined with relabeling of the MCMC output to address the label-switching problem, allows us to make meaningful interpretations of the top-level components to identify various GWG trajectory classes, which characterize different patterns of weight gain, and analyze the distributions of the neonatal outcomes within each GWG trajectory class. Further, including the maternal covariates in the third level of clustering ensures that the random partitions of subjects within each GWG trajectory class rely sufficiently on the neonatal outcomes and are not dominated by the maternal characteristics.

E1535: **On the robustness of machine learning methods for genomic prediction**
Presenter: **Vanda Lourenco**, NOVA University of Lisbon and NOVA MATH, Portugal
Co-authors: Piepho Hans-Peter, Joseph O. Ogutu

The accurate prediction of genomic breeding values is central to genomic selection in both plant and animal breeding studies. Genomic prediction (GP) involves the use of thousands of molecular markers spanning the entire genome and therefore requires methods able to efficiently handle high dimensional data. Machine learning (ML) methods, which encompass different groups of supervised and unsupervised learning methods, are becoming widely advocated for and used in GP studies. Although several studies have compared the predictive performances of individual methods, studies comparing the predictive performance of different groups of methods are rare. This is also the case of studies that assess the predictive performance of methods when data are contaminated. However, such studies are crucial for (i) identifying groups of methods with superior predictive performance, and (ii) assessing the merits and demerits of such groups of methods relative to each other and to the established classical methods when the phenotypic data are and are not contaminated. We comparatively evaluate, in terms of predictive accuracy and prediction errors, the genomic predictive performance and robustness of several groups of supervised ML methods. Specifically, regularized, ensemble, and instance-based methods, using one simulated dataset (animal breeding population; three distinct traits).

CI021 Room BH (S) 1.01 Lecture Theatre 1 MACHINE LEARNING AND MACROECONOMIC FORECAST Chair: Anna Simoni

C0204: **Bayesian modeling of time-varying parameters using regression trees**
Presenter: **Florian Huber**, University of Salzburg, Austria
Co-authors: Niko Hauzenberger, James Mitchell, Gary Koop

In light of widespread evidence of parameter instability in macroeconomic models, many time-varying parameter (TVP) models have been proposed. A nonparametric TVP-VAR model is proposed using Bayesian Additive Regression Trees (BART). The novelty of this model arises from the law of motion driving the parameters being treated nonparametrically. This leads to great flexibility in the nature and extent of a parameter change, both in the conditional mean and in the conditional variance. In contrast to other nonparametric methods that are black box, structural inference using our model is straightforward. Parsimony is achieved by adopting nonparametric factor structures and the use of shrinkage priors. In an application to US macroeconomic data, we illustrate the use of our model in understanding both the evolving nature of the Phillips Curve and how the effects of business cycle shocks on inflationary measures vary nonlinearly with movements in uncertainty.

C0205: **Deep dynamic factor models**
Presenter: **Giovanni Ricco**, University of Warwick, United Kingdom
Co-authors: Paolo Andreini, Cosimo Izzo

A novel deep neural network framework – that we refer to as Deep Dynamic Factor Model (D2FM) –, is able to encode the information available, from hundreds of macroeconomic and financial time series into a handful of unobserved latent states. While similar in spirit to traditional dynamic factor models (DFMs), differently from those, this new class of models allows for nonlinearities between factors and observables due to the autoencoder neural network structure. However, by design, the latent states of the model can still be interpreted as in a standard factor model. Both in a fully real-time out-of-sample nowcasting and forecasting exercise with US data and in a Monte Carlo experiment, the D^2FM improves over the performances of a state-of-the-art DFM.

C1283: **A large Bayesian VAR of the United States economy**
Presenter: **Domenico Giannone**, University of Washington, United States
Co-authors: Richard Crump, Argia Sbordone, Eric Qian, Stefano Eusepi

The United States' macroeconomic and financial sectors are modelled using a formal and unified econometric model. Through shrinkage, our Bayesian VAR provides a flexible framework for modeling the dynamics of thirty-one variables, many of which are tracked by the Federal Reserve. We show how the model can be used for understanding key features of the data, constructing counterfactual scenarios, and evaluating the macroeconomic environment both retrospectively and prospectively. Considering its breadth and versatility for policy applications, our modeling

approach gives a reliable, reduced-form alternative to structural models.

CO242 Room Virtual R01 NEW APPROACHES TO TIME SERIES ANALYSIS FOR MACRO AND FINANCE Chair: Mikkel Plagborg-Moller

C0189: **The real channel for nominal bond-stock puzzles**
Presenter: **Dongho Song**, Johns Hopkins University, United States
Co-authors: Mikhail Chernov, Lars Lochstoer

Evidence is presented showing that the mix of transitory and permanent shocks to consumption is changing over time. We identify three regimes: two highly persistent regimes where either permanent or transitory shocks are relatively more dominant, and a largely transitory disaster regime. We study the implications of this finding for asset prices. The transition from the second to the first regime in the mid-1990s makes the correlation between equities and bonds switch signs from positive to negative, as in the data. The real bond and equity yield curves are approximately flat. The nominal bond curve is upward-sloping. These results are achieved without relying on the nominal channel too much. That is, as in the data, the variation of inflation in the model is under 40% as a fraction of the variation in nominal yields.

C0190: **Finite-state Markov-chain approximations: A hidden Markov approach**
Presenter: **Eva Janssens**, University of Amsterdam, Netherlands
Co-authors: Sean McCrary

A novel finite-state Markov chain approximation method is proposed for Markov processes with continuous support. The method can be used for both uni- and multivariate processes, as well as non-stationary processes such as those with a life-cycle component. In contrast to existing methods, our discretization procedure provides both an optimal grid and transition probability matrix. We provide guidance on how to select the optimal number of grid points. The method is based on minimizing the information loss between a misspecified approximating model and the true data-generating process. The method outperforms existing discretization methods in several dimensions, including parsimoniousness. Furthermore, we demonstrate the performance of our discretization method compared to existing methods through the lens of an asset-pricing model and a life-cycle consumption-savings model. We find the choice of discretization method matters for the accuracy of the model solutions and for the welfare effects of risk.

C0934: **Uniform priors for impulse responses**
Presenter: **Jonas Arias**, Federal Reserve Bank of Philadelphia, United States
Co-authors: Juan Rubio-Ramirez, Daniel Waggoner

There has been a call for caution when using the conventional method for Bayesian inference in set-identified structural vector autoregressions on the grounds that the uniform prior over the set of orthogonal matrices could be nonuniform for key objects of interest. We challenge this call. Although the prior distributions of individual impulse responses induced by the conventional method may be nonuniform, they typically do not drive the posteriors if one does not condition on the reduced-form parameters. Importantly, when the focus is on joint inference, the uniform prior over the set of orthogonal matrices is not only sufficient but also necessary for inference based on a uniform joint prior distribution over the identified set for the vector of impulse responses. We also propose variants of the conventional method to conduct inference based on a uniform joint prior distribution for the vector of impulse responses. We generalize our results to vectors of objects of interest beyond impulse responses.

C1104: **Networking the yield curve surprises: Implications for monetary policy**
Presenter: **Tatevik Sekhposyan**, Texas A and M University, United States
Co-authors: Tatjana Dahlhaus, Julia Schaumburg

A flexible, time-varying network model is introduced to trace the propagation of interest rate surprises across different maturities. First, we develop a novel econometric framework that allows for unknown, potentially asymmetric contemporaneous spillovers across panel units, and establish the finite sample properties of the model via simulations. Second, we employ this innovative framework to model the dynamics of interest rate surprises jointly and to assess how various monetary policy actions, for example, short-term, long-term interest rate targeting and forward guidance, propagate across the yield curve. We find that the network of interest rate surprises is indeed asymmetric and defined by spillovers between adjacent maturities. Spillover intensity is high, on average, but shows strong time variation. Forward guidance is an important driver of spillover intensity. Pass-through from short-term interest rate surprises to longer maturities is muted, yet there are stronger spillovers associated with surprises at medium- and long-term maturities. We illustrate how our proposed framework helps our understanding of the ways various dimensions of monetary policy propagate through the yield curve and interact with each other.

C0932: **Standard errors for calibrated parameters**
Presenter: **Mikkel Plagborg-Moller**, Princeton University, United States
Co-authors: Matthew Cocci

Calibration, the practice of choosing the parameters of a structural model to match certain empirical moments, can be viewed as minimum distance estimation. Existing standard error formulas for such estimators require a consistent estimate of the correlation structure of the empirical moments, which is often unavailable in practice. Instead, the variances of the individual empirical moments are usually readily estimable. Using only these variances, we derive conservative standard errors and confidence intervals for the structural parameters that are valid even under the worst-case correlation structure. In the over-identified case, we show that the moment weighting scheme that minimizes the worst-case estimator variance amounts to a moment selection problem with a simple solution. Finally, we develop tests for over-identifying or parameter restrictions. We apply our methods empirically to a model of menu cost pricing for multi-product firms and a heterogeneous agent New Keynesian model.

CO292 Room BH (SE) 1.06 STATISTICAL IDENTIFICATION AND STRUCTURAL VARS Chair: Thorsten Drautzburg

C0266: **Incorporating economic theory into structural vector autoregressions: A non-invasive approach**
Presenter: **Sascha Keweloh**, TU Dortmund University, Germany

SVAR identification approaches are proposed to be based on information in higher moments with traditional restriction-based approaches. The proposed estimator allows the incorporation of prior economic knowledge using an adaptive ridge-type penalty. Therefore, the structure is incorporated in a non-invasive manner such that a correctly imposed structure improves the performance of the estimator, and the impact of an incorrectly imposed structure decreases with a sample size increase.

C0307: **Robust inference for non-Gaussian linear simultaneous equations models**
Presenter: **Adam Lee**, BI Norwegian Business School, Norway
Co-authors: Geert Mesters

All parameters in linear simultaneous equations models can be identified (up to permutation and scale) if the underlying structural shocks are independent and if, at most one of them is Gaussian. Unfortunately, existing inference methods that exploit such identifying assumptions suffer from size distortions when the true distributions of the shocks are close to Gaussian. To address this weak non-Gaussian problem, we develop a robust semi-parametric inference method that yields valid confidence intervals for the structural parameters of interest regardless of the distance to Gaussianity. We treat the densities of the structural shocks non-parametrically and construct identification robust tests based on the efficient score function. The finite sample properties of the methodology are illustrated in a large simulation study and an empirical study for production function estimation.

C0914: **Announcement-specific decompositions of unconventional monetary policy shocks & their macroeconomic effects**
Presenter: **Daniel Lewis**, University College London, United Kingdom

It is proposed to identify announcement-specific decompositions of asset price changes into monetary policy shocks exploiting heteroskedasticity in intraday data, accommodating both changes in the nature of shocks and the state of the economy across announcements. We compute decompositions with respect to Fed Funds, forward guidance, asset purchase, and Fed information shocks from 1996 to 2019. The decompositions illustrate which announcements of unconventional policy measures had significant effects during the Great Recession. Forward guidance and asset purchases have significant effects on yields, spreads, equities, and uncertainty. Positive shocks to all dimensions of monetary policy trigger macroeconomic contractions, while information shocks telegraph expansions.

C0884: **Identifying structural vector autoregression via leptokurtic economic shocks**
Presenter: **Markku Lanne**, University of Helsinki, Finland
Co-authors: Keyan Liu, Jani Luoto

The generalized method of moments (GMM) estimation of the non-Gaussian structural vector autoregressive (SVAR) model is revisited. It is shown that in the n-dimensional SVAR model, global and local identification of the contemporaneous impact matrix is achieved with as few as $n^2+n(n-1)/2$ suitably selected moment conditions, when at least $n-1$ of the structural errors are all leptokurtic (or platykurtic). We also relax the potentially problematic assumption of mutually independent structural errors in part of the previous literature to the requirement that the errors be mutually uncorrelated. Moreover, we assume the error term to be only serially uncorrelated, not independent in time, which allows for univariate conditional heteroskedasticity in its components. A small simulation experiment highlights the good properties of the estimator and the proposed moment selection procedure. The use of the methods is illustrated by means of an empirical application to the effect of a tax increase on U.S. gasoline consumption and carbon dioxide emissions.

C0879: **Refining set-identification in VARs through independence**
Presenter: **Thorsten Drautzburg**, Federal Reserve Bank of Philadelphia, United States
Co-authors: Jonathan Wright

Identification in VARs has traditionally mainly relied on second moments. Some researchers have considered using higher moments as well, but there are concerns about the strength of the identification obtained in this way. We propose refining existing identification schemes by augmenting sign restrictions with a requirement that rules out shocks whose higher moments significantly depart from independence. This approach does not assume that higher moments help with identification; it is robust to weak identification. In simulations, we show that it controls coverage well, in contrast to approaches that assume that the higher moments deliver point-identification. However, it requires large sample sizes and/or considerable non-normality to reduce the width of confidence intervals by much. We consider some empirical applications. We find that it can reject many possible rotations. The resulting confidence sets for impulse responses may be non-convex, corresponding to disjoint parts of the space of rotation matrices. We show that, in this case, augmenting sign and magnitude restrictions with an independence requirement can yield bigger gains.

CO032 Room BH (SE) 2.05 RECENT ADVANCES IN APPLIED MACROECONOMICS Chair: Alessia Paccagnini

C0315: **Relative prices and pure inflation since the mid-1990s**
Presenter: **Matteo Luciani**, Amazon.com, United States
Co-authors: Hie Joo Ahn

Consumer price inflation is decomposed into pure inflation, relative price inflation, and idiosyncratic inflation by estimating a dynamic factor model on a data set of 146 monthly disaggregated prices from 1995 to 2019. We find pure inflation is the trend around which PCE price inflation fluctuates. In contrast, relative price inflation and idiosyncratic inflation drive the fluctuation of PCE price inflation around the trend. Unlike Reis and Watson, we find that labor market slack is the main driver of pure inflation and that energy prices account for variation in relative price inflation.

C1479: **Investment response to monetary policy in a low interest rate environment: Evidence from the ECB's corporate QE**
Presenter: **Supriya Kapoor**, Trinity College Dublin, Ireland
Co-authors: Guillaume Horny

The purpose is to study how an easing in corporate bond funding conditions affects the asset structure of firms' fixed assets. ECBs Corporate Sector Purchase Program is employed as a quasi-natural experiment that reduces bond yields for firms eligible for ECB purchases. We identify eligible firms using the information on their bond ratings. Using consolidated balance sheet information on non-financial firms in France, we find that firms increase investment expenses but only to replace existing assets, whether tangible or intangible, instead of investing in new equipment to grow in scale. This replacement is, however, not homogeneous across asset classes, since intangible assets increase in importance relative to tangible ones. The shift towards intangible assets is stronger for firms with a BBB rating than for safer firms (AAA-A rating). This suggests that while BBB-rated firms were, to some extent, constrained in their funding, they do not use the proceeding to reinforce the collateral value of their assets. These effects are robust to the inclusion of several fixed effects. We conclude that easier access to market debt can have an effect on the mix of fixed assets used by firms to produce. This raises questions as to whether firms eligible for CSPP purchases increased their productivity since new equipment can be more efficient than the deprecated ones.

C1164: **Dealing with the statistical representation of DSGE models**
Presenter: **Alessia Paccagnini**, University College Dublin, Ireland

Novel empirical evidence about model validation in DSGE modeling is provided. First, we use several small - and medium-scale models as data-generation processes to create artificial pseudo-data. Second, using this pseudo-data, we identify shocks by using VAR and Local Projections to evaluate the effects of macroeconomic shocks such as monetary policy and fiscal policy shocks. As main findings, we document how both VAR and Local Projection help researchers recover the dynamics of a DSGE, even if they cannot always recover the dynamics of the true data generation processes. These results provide a guideline to investigate the misspecification in empirical DSGE modeling due to the statistical representation.

C1481: **How nonlinear is monetary policy**
Presenter: **Dilan Aydin Yakut**, University of Bologna, Central Bank of Ireland, Ireland
Co-authors: David Byrne, Robert Goodhead

The idea that monetary policy might have non-linear effects has a long history, dating back to Keynes' argument that monetary easing during deep recessions would be like pushing on a string. The position that the sources of monetary policy non-linearity may be highly multi-dimensional is taken. The focus is on potential state-dependent responses for yields and equities. Rather than a piecemeal investigation of a few potential sources of non-linearity, as is common in existing empirical work, we simultaneously examine the role of many variables. To do this, we combine an event study approach with large N methods to assess the role of potential interactors. We handle dimension reduction with both sparse and dense techniques, using LASSO and factor specifications to capture nonlinearity. We use monetary shocks identified by high-frequency information from Fed meeting days. Results indicate a role for both financial and macroeconomic variables in determining non-linearity.

C1134: **Macroeconomic effects of different tax policies: Narrative evidence from Canada**
Presenter: **Daniela Hauser**, Bank of Canada, Canada
Co-authors: Rodrigo Sekkel, Dmitry Matveev, Aaron Leonard

The pandemic has underlined the importance of understanding the effect of specific fiscal instruments, and the need to offer guidance for judging

the relative merits of different types of fiscal instruments. Estimating the effects of tax changes on economic activity poses a significant empirical challenge because observed tax changes are rare, discrete events whose motivations are often correlated with other drivers of short-term economic activity. While there are different approaches to identifying exogenous changes to federal taxes, one increasingly popular way to overcome the identification challenges is to use a narrative approach based on a wide variety of historical documents. A new narrative dataset is constructed for legislated changes to taxes on personal income, corporate income and consumption in Canada to analyze how these fiscal-policy changes affects the economy and public finances. Specifically, light is shed on which tax cuts stimulate the economy, and which tax increases harm economic recovery. Finally, we will discuss the implications of our empirical findings for theoretical models.

CO603 Room BH (SE) 2.10 FINANCIAL ECONOMETRICS AND APPLICATIONS Chair: Xiaohan Xue

C0860: **Improving the estimation and predictions of small time series models**
Presenter: **Gareth Liu-Evans**, University of Liverpool, United Kingdom

A new approach is developed for improving the point estimation and predictions of parametric time-series models. The method targets performance criteria such as estimation bias, root mean squared error, variance, or prediction error, and produces closed-form estimators focused towards these targets via a computational approximation method. This is done for an autoregression coefficient, for the mean reversion parameter in Vasicek and CIR diffusion models, for the Binomial thinning parameter in integer-valued autoregressive (INAR) models, and for predictions from a CIR model. The success of the prediction targeting approach is shown in Monte Carlo simulations and in out-of-sample forecasting of the US Federal Funds rate.

C1214: **Combining predictive distributions**
Presenter: **Xiaochun Meng**, University of Sussex, United Kingdom
Co-authors: James Taylor, Souhaib Ben Taieb

Combining distributional predictions is an important topic in the forecasting literature. Individual distributional predictions are aggregated to reach a consensus distribution that often has better forecasting accuracy. We propose a novel method for estimating the combining weights based on kernel scores. We show that the proposed methods have several appealing properties when compared to the traditional method based on the log score. We use simulation data and the ECB survey of professional forecasters data to support our theoretical results.

C0213: **Jump clustering, stock price efficiency and predictability of jumps in financial markets**
Presenter: **Jian Chen**, University of Reading, United Kingdom

The focus is on the clustering behaviours of assets' return jumps modelled by a self/cross-exciting process embedded in a stochastic volatility model. Based on the model, we carry out two exercises. Firstly, we relate the jump clustering behaviours to information flows and propose a new measurement of stock price efficiency. We show the capability of our new measurement to capture the speed at which stock prices possess new information, especially at the firm-specific level. Secondly, we propose a forecasting framework for asset return jumps and assess their predictability. We sample latent states with a particle filter in the out-of-sample and perform one-step-ahead probabilistic predictions on upcoming jumps. We further develop a statistic based on predicted probabilities of positive and negative jumps and show its usefulness in forecasting returns. We conduct empirical studies in the US stock market, commodity futures markets, and foreign exchange markets.

C1502: **A market-level tug of war: Investor heterogeneity and asset pricing**
Presenter: **Ran Tao**, University of Bristol, United Kingdom
Co-authors: Lei Zhao, Chardin Wese Simen

A daily tug of war between opposing investor clientele at the individual stock level has been documented in the asset pricing literature. We measure a market-level tug of war using the cross-sectional intensity of individual tug of war. The Capital Asset Pricing Model (CAPM) tends to perform better, and market betas are strongly and positively related to average returns on quiet days when the market-level tug of war is less intensive. We further show that the well-established finding that a robust risk-return trade-off exists on important information days (e.g. FOMC announcement days and influential firms' earnings announcement days) holds only when such information days coincide with quiet days. Overall, our findings suggest that investor heterogeneity has significant implications on asset pricing.

C0760: **When MIDAS meets LASSO: Forecasting tail risk using effective macroeconomic variables**
Presenter: **Xiaohan Xue**, University of East Anglia, United Kingdom
Co-authors: Marwan Izzeldin, Yi Luo

A new framework for the joint estimation and forecasting of Value at Risk (VaR) and Expected Shortfall (ES) is proposed, which incorporates low-frequency macroeconomic and financial indicators into the quantile-based MIDAS model. Using an innovative machine-learning approach that maximizes the penalized Asymmetric Laplace (AL) likelihood function with an Adaptive-Lasso penalty, the most informative variables are selected in a "big data" setting. A dynamic selection process enables the visualizing of the variable-selection evolution. In the empirical analysis, three variables (namely, realized volatility, term spread and housing starts) are consistently selected for most of the rolling windows and serve as the strongest predictors of future tail risk. More information may be required to predict more extreme VaR and ES. The out-of-sample backtesting results show that our method passes most backtests with relatively higher p-values and achieves the minimum loss in the joint forecasting of VaR and ES.

CO657 Room BH (SE) 2.12 RECENT METHODS FOR ANALYZING INFLATION Chair: Carlos Montes-Galdon

C0972: **Trend inflation and monetary policy**
Presenter: **Luis Uzeda**, Bank of Canada, Canada

A new empirical framework is developed to investigate the relationship between monetary policy and trend inflation in the U.S. economy. We combine two modeling approaches: measuring trend inflation using an unobserved components model and estimation of monetary policy rules that allow for their reaction parameters to change in a state-dependent fashion. State dependency is determined by linking trend inflation dynamics to plausible changes in the conduct of monetary policy. We document two main results: (1) the relationship between monetary policy and trend inflation has strengthened during the Great Moderation relative to the Great Inflation period; and (2) the stabilization of trend inflation around the mid-1980s is consistent with a regime change akin to the adoption of an inflation targeting strategy.

C0368: **Modeling tail risks of inflation using unobserved component quantile regressions**
Presenter: **Michael Pfarrhofer**, University of Vienna, Austria

Methods are proposed for Bayesian inference in time-varying parameter (TVP) quantile regressions (QRs) featuring conditional heteroskedasticity. We use data augmentation schemes to render the model conditionally Gaussian and develop an efficient sampling algorithm. Regularization of the high-dimensional parameter space is achieved via dynamic shrinkage priors. The merits of the proposed approach are illustrated in a simulation study. A simple version of TVP-QR based on an unobserved components model is applied to dynamically trace the quantiles of inflation in the United States, the United Kingdom and the euro area. In an out-of-sample forecast exercise, we find the proposed model to be competitive and perform particularly well for higher-order and tail forecasts. A detailed analysis of the resulting predictive distributions reveals that they are sometimes skewed and occasionally feature heavy tails.

C1110: **Skewed SVARs: Tracking the structural sources of macroeconomic tail risks**
Presenter: **Carlos Montes-Galdon**, European Central Bank, Germany
Co-authors: Eva Ortega

A VAR model is proposed with structural shocks (SVAR) that are identified using sign restrictions, and whose distribution is subject to time-varying skewness. We also present an efficient Bayesian algorithm to estimate the model. The model allows tracking joint asymmetric risks to macroeconomic variables included in the SVAR, and provides a structural narrative to the evolution of those risks. When faced with euro area data, our estimation suggests that there has been a significant variation in the skewness of demand, supply and monetary policy shocks. Such variation can explain a significant proportion of the joint dynamics of real GDP growth and inflation, and also generates important asymmetric tail risks in those macroeconomic variables. Finally, compared to the literature on growth- and inflation-at-risk, we find that financial stress indicators are not enough to explain all the macroeconomic tail risks.

C0614: **Anchoring long-run inflation expectations in a panel of professional forecasters**
Presenter: **Sebastian Rast**, De Nederlandsche Bank, Netherlands
Co-authors: Leonardo Melosi, Jonas Fisher

Panel data from the U.S. Survey of Professional Forecasters are used to estimate a model of individual forecaster behavior in an environment where inflation follows a trend-cycle time series process. Our model allows us to estimate the sensitivity of forecasters' long-run expectations to incoming inflation and news about future inflation, and measure the coordination of beliefs about future inflation. We use our model of individual forecasters to study average long-run inflation expectations. Short-term changes in inflation have small effects on average expectations; the sensitivity to news is over twice as large but is still relatively small. These findings provide a partial explanation for why the anchoring and subsequent de-anchoring of average inflation expectations from 1991 to 2020 were such long-lasting episodes. Our model suggests coordination of beliefs also played a role, slowing down but not preventing the pull on average expectations from inflation from running persistently below target. We apply our model to the case of a U.S. central banker setting policy in September 2021. Our results suggest the high inflation readings of mid-2021 would have to be followed by overshooting of the Fed's target generally at the high end of the Fed's Summary of Economic Projections to re-anchor long-term expectations at their pre-Great Recession level.

C1500: **Modelling Euro area inflation expectations: The value of mixing sources and frequencies**
Presenter: **Eva Ortega**, Banco de Espana, Spain
Co-authors: Ricardo Gimeno

A daily model is proposed for obtaining inflation expectations in the euro area. It is a modified version of a standard dynamic term structure model with (i) a time-varying drift in inflation expectations and (ii) mixed frequencies to allow for various sources of information: daily inflation compensation data from euro area Inflation Linked Swaps at 15 different maturities, survey data from the ECBs quarterly Survey of Professional Forecasters, and observed monthly HICP inflation.

CC753 Room BH (SE) 1.01 TIME SERIES I Chair: Philipp Otto

C0478: **Joint semiparametric INAR bootstrap inference for model coefficients and innovation parameters**
Presenter: **Maxime Faymonville**, TU Dortmund University, Germany
Co-authors: Carsten Jentsch, Christian Weiss, Boris Aleksandrov

For modeling the serial dependence in time series of counts, various approaches have been proposed in the literature. In particular, models based on an autoregressive-type structure, such as the well-known integer-valued autoregressive (INAR) models, are very popular in practice. Besides the binomial thinning, these models are determined by autoregressive coefficients and a discrete innovation distribution. The literature mainly deals with the parametric estimation of INAR models, which restricts the flexibility of the considered model class in applications. Using semiparametric estimation, it is possible to jointly estimate the autoregressive coefficients and the innovation distribution, where the estimation of the innovation distribution works fully non-parametric. Using empirical process theory, the resulting semiparametric estimator is known to be consistent with some complicated limiting distribution which enables asymptotic inference and model diagnostics on the innovations. We consider a corresponding semiparametric INAR bootstrap procedure. We show that the bootstrap estimator for the autoregressive model coefficients and the innovation distribution provides the same limiting distribution such that the semiparametric bootstrap becomes asymptotically valid for the estimation of the innovation distribution. Simulations are used to illustrate the finite sample performance of the semiparametric INAR bootstrap using several common and uncommon innovation distributions.

C1162: **Mixed effects quantile autoregressive modeling for point-referenced daily maximum temperatures in Aragon, Spain**
Presenter: **Jorge Castillo-Mateo**, University of Zaragoza, Spain
Co-authors: Ana C Cebrian, Jesus Asin, Alan Gelfand

Different spatial patterns of climate change are analyzed across quantiles associated with point-referenced daily maximum temperatures in Aragon, northeastern Spain. For that purpose, regression through asymmetric Laplace errors is considered in the context of a very flexible mixed effects autoregressive model, introducing two temporal scales and four spatial processes. Moreover, while the autoregressive model yields conditional quantiles, it is demonstrated how to extract marginal quantiles from the conditional quantiles with the asymmetric Laplace specification. Marginal quantiles enjoy direct interpretation as well as the benefit of spatial interpolation, i.e., they do not require the previous day's temperature.

C1769: **Forecasting natural gas prices with spatio-temporal copula-based time series models**
Presenter: **Sven Pappert**, TU Dortmund University, Germany
Co-authors: Antonia Arsova

Commodity price time series possesses interesting features, such as heavy-tailedness, skewness, heteroskedasticity, and non-linear dependence structures. These features pose challenges for modeling and forecasting. We explore how spatio-temporal copula-based time series models can be effectively employed for these purposes. We focus on price series for fossil fuels and carbon emissions. Further, we illustrate how the t-copula may be used in conditional heteroskedasticity modeling. The possible emergence of non-elliptical probabilistic forecasts in this context is examined and visualized. The problem of finding an appropriate point forecast given a non-elliptical probabilistic forecast is discussed. We propose a solution where the forecast is augmented with an artificial neural network (ANN). The ANN predicts the best (in MSE sense) quantile to use as a point forecast. In a forecasting study, we find that the copula-based models are competitive.

C1893: **Long run effects of high energy prices on energy intensity: A sparse filter approach**
Presenter: **Jan Bruha**, CNB, Czech Republic

A model is presented linking energy prices to energy intensity. We use a sparse filter to show on a panel of advanced economies that periods of high energy prices are associated with a strong decline in energy intensity. Based on this finding, we derive conclusions for the long-run effects of high energy prices on external positions of energy importers and exporters.

C0331: **Monte Carlo likelihood ratio tests for Markov switching models**
Presenter: **Gabriel Rodriguez Rondon**, McGill University, Canada
Co-authors: Jean-Marie Dufour

Markov switching models have wide applications in economics, finance, and other fields. Many studies focusing on identifying the number of

regimes in a Markov switching model have been limited to hypothesis tests with a null of one regime and an alternative hypothesis of two regimes. We use Monte Carlo procedures to deal with nuisance parameters and circumvent the issues plaguing conventional hypothesis testing procedures by working with the sample distribution of the test statistic. The tests proposed here can deal with non-stationary processes, non-Gaussian errors, and multivariate settings. They are also applicable to the general setting where we are interested in testing a null hypothesis with M_0 regimes against an alternative with $M_0 + m$ regimes where both $M_0, m \geq 1$. Further, the maximized Monte Carlo likelihood proposed ratio test (MMC-LRT) is an identifications-robust, valid test procedure both in finite samples and asymptotically. Simulation results are provided for both univariate and multivariate settings and suggest the proposed tests can control the level of the test and have good power. Finally, we present an empirical application using U.S. GNP growth data to showcase the usefulness of our proposed test procedures.

CC784 Room BH (SE) 1.02 EMPIRICAL FINANCE Chair: Mohammad Jahan-Parvar

C0382: **Drivers of hedge fund failures**

Presenter: **Huda Aldhahi**, University of East Anglia, United Kingdom

Various survival methods are implemented to investigate the hedge funds' survival and precisely address the determinants of hedge fund failure using the HFR database from 1994 to 2020. The primary analysis uses the non-parametric survival analysis, the Semi-parametric Cox, and Weibull accelerated failure time (AFT). Also, we investigate the elements that influence hedge funds' survival using a wide range of variables, including funds' age, size, performance, management fee, incentive fee, volatility, leverage, and cancellation policy periods. The main conclusion of our findings can be summarized as follows: hedge fund age, size, and performance are the primary determinants that support hedge fund survival. In contrast, volatility and leverage negatively influence hedge fund survival. Finally, Management fee, lockup period, advance notice period, and incentive fee variables have mixed results depending on measures, style, and the evaluation period. Our findings maintained their empirical significance and were consistent in pre- and post-financial crisis periods. Our findings will allow new hedge fund managers and investors to evaluate the predicted hedge fund's lifetime, thus diversifying their portfolios by selecting an ideal hedge fund classification based on its highest survival time and alerting existing hedge fund managers and investors to any possible fund liquidation.

C0795: **Conditional leverage and the term structure of option-implied equity risk premia**

Presenter: **Hugues Langlois**, HEC Paris, France

Co-authors: Fousseni Chabi-Yo

In a one-period economy, bounds for the equity risk premium were derived, which use options of the same maturity as the horizon at which the premium is measured. In contrast, we provide an expression and an empirical methodology to measure the premium at a given horizon in a multi-period economy using options of multiple maturities. The premium depends on risk-neutral leverage effects and the expected future risk-neutral market variance and skewness, which contribute to increasing the premium at short horizons. Our measure outperforms in terms of prediction accuracy and portfolio allocation performance. The term structure of expected excess holding period returns is flatter on average and dramatically more negative during market turmoil than those implied by previous measures.

C1410: **Market effects of central bank credit markets support programs in Europe**

Presenter: **Oleg Sokolinskiy**, Federal Reserve Board, United States

Co-authors: Yuriy Kitsul, Jonathan Wright

Using responses of credit default swap indexes to ECB monetary policy announcements, we isolate a novel credit policy component of monetary policy surprises. We examine how such unconventional monetary policy surprises affect investor perceptions of credit risk and the functioning of primary corporate debt markets. Favorable credit surprises cause declines in uncertainty about credit risk and suggest a more stable outlook on its dynamics over the following months. Both net and gross corporate bond issuance increase as a result of favorable credit surprises, with the largest response in investment grade issuance. We argue that this provides evidence for the efficacy of a local channel of unconventional monetary policy.

C1599: **Projecting long-run compound returns: The limits of data-driven inference**

Presenter: **Tamas Kiss**, Orebro University, School of Business, Sweden

Co-authors: Erik Hjalmarsson, Andreas Dzemski, Adam Farago

What is the unconditional forecast, or projection, for the payoff of a stock investment over very long horizons, such as 30 years? Empirically, the question is inherently difficult to answer, since we, at most, observe a handful of actual 30-year returns. Some possible solutions have recently been proposed to deal with this problem (in particular, using bootstrap-resampling schemes to construct a large number of long-run returns), but little is known about the statistical properties of these methods. In particular, whereas some recent work has dealt with obtaining empirical point estimates of long-run stock return distributions, the central question of sampling uncertainty has been left mostly unanswered. We analyze the properties of different empirical methods used for projecting long-run return distributions. The aim is to provide an understanding of the limits of what we might feasibly be able to say, with some meaningful precision, regarding the distribution of long-run returns. We provide formal theoretical results on the properties of recently proposed bootstrap methods and contrast these with alternative parametric methods. Initial findings highlight that confidence bands around long-run distributions are very wide and often dwarf potential differences between empirical distributions under different distributional assumptions and inference methods.

C1715: **Do investors compensate for unsustainable consumption with sustainable assets?**

Presenter: **Emily Kormanyos**, Goethe University Frankfurt and Leibniz Institute SAFE, Germany

Categorized bank transactions are combined with Multiregional Input-Output data (MRIO) to estimate carbon footprints of consumption for over 6,000 investors. Based on checking and trading account transactions, we provide novel evidence of significant and negative spillovers between sustainable consumption and investments, especially for assets with favorable GHG-emission ratings. Several tests show that unsustainable consumption explains cross-sectional variation in sustainable investor portfolios better than alternative explanations based on heterogeneous sustainability preferences or financial motives. These results suggest that unsustainable consumers aim to offset their consumption-based emissions through sustainable investments.

CC801 Room BH (SE) 1.05 ASSET ALLOCATION Chair: Massimiliano Caporin

C1372: **Moving forward from predictive regressions: Boosting asset allocation decisions**

Presenter: **Henri Nyberg**, University of Turku, Finland

Co-authors: Lauri Nevasalmi

A flexible utility-based empirical approach is introduced to determine asset allocation decisions between risky and risk-free assets directly. This is in contrast to the commonly used two-step approach where least squares optimal statistical equity premium predictions are first constructed to form portfolio weights before economic criteria are used to evaluate resulting portfolio performance. Our single-step customized gradient boosting method is specifically designed to find optimal portfolio weights in a direct utility maximization. Empirical results of the monthly U.S. data show the superiority of boosted portfolio weights over several benchmarks, generating interpretable results and profitable asset allocation decisions.

C1887: **Bet on a bubble asset: An optimal portfolio allocation strategy**

Presenter: **Arthur Thomas**, Paris Dauphine University - PSL, France

Co-authors: Gilles De Truchis, Elena Dumitrescu, Sebastien Fries

Portfolio allocation is discussed when one asset exhibits phases of locally explosive behavior. We model the conditional distribution of such an asset through mixed causal-non-causal models, which mimic well the speculative bubble behaviour. Relying on a Taylor-series-expansion of a CRRA utility function approach, the optimal portfolio(s) is(are) located on the mean-variance-skewness-kurtosis efficient surface. We analytically derive these four conditional–moments and show in a Monte-Carlo simulations exercise that incorporating them into a two-assets portfolio optimization problem leads to substantial improvement in the asset allocation strategy. All performance evaluation metrics support the higher out-of-sample performance of our investment strategies over standard benchmarks such as the mean-variance and equally-weighted portfolio. An empirical illustration using the Brent oil price as a speculative asset confirms these findings.

C0221: **A time-varying parameter model with Bayesian shrinkage for global minimum variance portfolio prediction**

Presenter: **Roman Liesenfeld**, University of Cologne, Germany
Co-authors: Laura Reh, Guilherme Moura, Tore Selland Kleppe

A novel dynamic approach is proposed to high-dimensional portfolio selection based on predictions for the Global Minimum Variance Portfolio (GMVP). Using Bayesian regularization techniques, we aim to robustify the portfolios against estimation risk. Exploiting that the GMVP weights can be obtained as the population coefficients of a linear regression of one benchmark return on a vector of return differences, we set up a linear state space model with time-varying parameters and stochastic volatility. This specification allows addressing both the time variation in the assets' conditional covariance structure and the heteroscedasticity in the market. Bayesian inference techniques with LASSO-type priors provide data-driven shrinkage to alleviate overfitting and to identify time-invariant coefficients automatically. Our approach allows for scalability to high dimensional applications and performs well in applications in which the number of observations per asset is low. The applicability and robustness of our approach are demonstrated through comprehensive simulation and empirical analysis. In particular, a simulation study shows that the proposed approach can perform better than the true model when the number of observations is not much larger than the number of assets. An application to daily financial returns also shows that the model performs better than a wide range of existing approaches in terms of out-of-sample forecasting accuracy.

C1949: **Principles of Bayesian portfolio choice**

Presenter: **Jan Vecer**, Charles University, MFF, Ke Karlovu 3, 121 16 Praha 2, Czech Republic

Utility maximization depends on the choice of the underlying riskless asset as a numeraire. We show that the only numeraire invariant utility is a logarithmic function. We also note that the prices can be expressed as the likelihood ratio of the respective state price densities. On the other side, each state price density generates an asset that corresponds to the log utility optimal portfolio with respect to all assets. This is important in portfolio diversification; more opinions about the state price density generate more assets to invest in. We show that the expected log return of the price is a relative entropy between the state price densities. When the market agent that maximizes log utility uses a mixture distribution of the state price densities of the market assets, the resulting optimal portfolio is static. When the market agent uses a prior distribution for her market opinion distribution, the resulting wealth of each parameter updates in a Bayesian fashion.

C1920: **Persistence-based portfolio choice along the FOMC cycle**

Presenter: **Federico Severino**, Universite' Laval, Canada
Co-authors: Fulvio Ortu, Pietro Reggiani

The Federal Reserve holds two main sets of monetary policy meetings, the 'Federal Open Market Committee' (FOMC) and the 'Board Meetings', which gather with a 6-week and 2-week cadence, respectively. The cadence of these meetings has been shown to be associated with cycles of corresponding frequency in stock markets. These can be fruitfully exploited through a portfolio strategy that invests in the whole market at alternate weeks (even-week strategy). This simple investment rule is based on the cycles identified empirically but, so far, lacks a theoretical foundation. We provide a rigorous framework to detect cycles in the stock market, and to determine optimal portfolio choices that profit from such cycles. We use a filtering approach for stationary time series to isolate uncorrelated components of stock returns that are precisely associated with two- and six-week cycles. Then, we replicate these components using tradeable assets from the U.S. market, and design an optimal portfolio strategy that maximizes the investors' wealth and outperforms the even-week strategy.

Monday 19.12.2022 08:40 - 09:55 Parallel Session L – CFE-CMStatistics

EO178 Room S-1.01 STATISTICAL METHODS FOR COMPLEX DATA Chair: Maria Brigida Ferraro

E1050: **Individuality-based fuzzy cluster-scaled principal component analysis for high-dimension low-sample data**
Presenter: **Mika Sato-Ilic**, University of Tsukuba, Japan

A fuzzy clustering-based Principal Component Analysis (F-PCA) is presented by considering the individuality of subjects for high-dimension, low-sample size (HDLSS) data. Analyzing HDLSS data, including the difference of subjects, is useful as an implementation of a custom-made system for healthcare considering the individual history of daily activities. For example, if we observe data over subjects by sensors worn on the body during activities, analyzing high dimensional times with a low number of sensor measurements over the subjects is useful for implementing the custom-made system for healthcare. A simultaneous analysis is necessary for obtaining the well-classified result among different subjects and the efficient reduction of high dimensional data with sufficient explainability of the original data. Without external information on the difference between the subjects, we need to capture the difference of subjects from the single target original data. Since the fundamental idea of F-PCA is the inclusion of the weights measured by the fuzzy cluster scale commonly obtained over the subjects to the covariance of variables of the original data, the proposed F-PCA is adaptable to this analysis. The proposed F-PCA is shown to have a better performance with a comparison of the results of ordinary PCA by several numerical examples.

E1251: **Nonparametric moment-based estimation of simulated models via regularized regression**
Presenter: **Raffaello Seri**, University of Insubria, Italy
Co-authors: Mario Martinoli

A new method for the estimation of simulated models is presented. It exploits nonparametric least absolute shrinkage and selection operator (Lasso) to find the parameters of a simulation model producing statistics that are close to the ones obtained in real-world data. The simulation model is run for several values of the parameters, statistics are computed on each run, and the function linking the generated statistics and the associated parameters is estimated nonparametrically. Estimates of the parameters are then obtained through the previous nonparametric estimate using real-world statistics as explanatory variables. At odds with simulated minimum-distance techniques (e.g., indirect inference and simulated method of moments), our framework does not involve any explicit objective function, and no optimization algorithm is required. The asymptotic rate of convergence of the estimator to the true value and the error in the estimation of the coefficients and the prediction are explicitly and rigorously characterized. The approach is evaluated through a small simulation study.

E1382: **Clustering of star-shaped sets with a fuzzy approach**
Presenter: **Maria Brigida Ferraro**, Sapienza University of Rome, Italy
Co-authors: Elena Fernandez Iglesias, Ana Belen Ramos-Guajardo, Gil Gonzalez-Rodriguez

Star-shaped sets represent a large class of compact and convex sets as a particular case. The original shape of these sets is identified by means of the center-radial characterization. These data can be defined as complex because multiple aspects of them are enclosed in a single representation. Since, in some situations, it may be very useful to partition these sets into groups, a fuzzy clustering method is addressed. In detail, the proposal is a generalization of the well-known fuzzy k-means that takes into account the nature of the data through an appropriate distance measure. A real case study is reported to check the proposal's adequacy.

EO416 Room S-1.22 NEW PERSPECTIVES ON OLD QUESTIONS IN SURVIVAL ANALYSIS Chair: Liming Xiang

E1598: **Robust prediction of failure time through unified Bayesian analysis of nonparametric transformation models**
Presenter: **Chong Zhong**, The Hong Kong Polytechnic University, Hong Kong
Co-authors: Junshan Shen, Jin Yang, Catherine Liu

Nonparametric transformation models (NTMs) have sparked much interest in survival prediction owing to their flexibility with both transformations and error distributions unspecified. However, fitting these models has been hampered because they are unidentified. Existing approaches typically constrain the parameter space to ensure identifiability, but they incur intractable computation and cannot scale up to complex data; other approaches address the identifiability issue by making strong a priori assumptions on either of the nonparametric components and thus are subject to misspecifications. Utilizing a Bayesian workflow, we address the challenge by constructing new weakly informative nonparametric priors for infinite-dimensional parameters so as to remedy flat likelihoods associated with unidentified models. To facilitate the applicability of these new priors, we subtly impose an exponential transformation on top of NTMs, which compresses the space of infinite-dimensional parameters to positive quadrants while maintaining interpretability. We further develop a cutting-edge posterior modification technique for estimating the fully identified parametric component. Simulations reveal that our method is robust and outperforms the competing methods, and an application to a Veterans lung cancer dataset suggests that our method can predict survival time well and help develop clinically meaningful risk scores, based on patients' demographic and clinical predictors.

E1993: **Disease progression-based feature screening for ultrahigh- dimensional survival-associated biomarkers**
Presenter: **Liming Xiang**, Nanyang Technological University, Singapore
Co-authors: Mengjiao Peng

The increased availability of ultrahigh-dimensional biomarker data and the high demand for identifying biomarkers importantly related to survival outcomes made feature screening methods commonplace in the analysis of cancer genome data. In the presence of progression-free survival (PFS), a surrogate endpoint for overall survival (OS), the correlation between OS and PFS has suggested a high concordance in both survival endpoints; namely, patients with higher PFS would most likely have longer OS. We propose a novel feature screening method by incorporating surrogate information of PFS into the selection of important biomarker predictors for more accurate inference of OS after disease progression. The proposal is based on the rank of correlation between individual features and the conditional distribution of OS given observations of PFS. It is advantageous for its flexible model nature, which requires no marginal model assumption for OS or PFS, and the minimal computational cost for implementation. Theoretical results show its ranking consistency, sure screening, and false rate control properties. Simulation results demonstrate that the proposed screener leads to a more accurate feature selection than the method without considering the prior information about PFS. An application to breast cancer genome data illustrates its practical utility and facilitates disease classification using selected biomarker predictors.

E1568: **Hierarchical multi-parameter regression survival models**
Presenter: **Fatima-Zahra Jaouimaa**, University of Limerick, Ireland
Co-authors: Il Do Ha, Kevin Burke

Standard survival models introduce covariates through a single (scale) parameter, and we refer to this standard practice as Single-Parameter Regression (SPR). In contrast, Multi-Parameter Regression (MPR) allows covariates to enter the model through multiple distributional parameters, i.e., scale and shape. This approach to modelling has been shown to produce flexible and robust models with a relatively low model complexity cost. However, it is very common to have clustered data arising from survival analysis studies, and this is something that is underdeveloped in the MPR context. Therefore, we extend MPR models to handle multivariate survival data by introducing random effects in both the scale and the shape regression components. We consider a variety of possible dependence structures for these random effects (independent, shared, and correlated), and estimation proceeds using a h-likelihood approach. As the shape parameter may be viewed as a dispersion parameter for log-time, our proposal

bears similarities to Double Hierarchical Generalized Linear Modelling (DHGLM). We investigate the performance of our estimation procedure using simulated data, and also consider a real data example.

EO370 Room K0.20 NONCLASSICAL EXTREME VALUE ANALYSIS Chair: Yi He

E0548: **Extreme value inference for general heterogeneous data**
Presenter: **Yi He**, University of Amsterdam, Netherlands
Co-authors: John Einmahl
Extreme value statistics are extended to independent data with possibly very different distributions. In particular, we present novel asymptotic normality results for the Hill estimator, which now estimates the extreme value index of the average distribution. Due to the heterogeneity, the asymptotic variance can be substantially smaller than that in the i.i.d. case. As a special case, we consider a heterogeneous scales model where the asymptotic variance can be calculated explicitly. The primary tool for the proofs is the functional central limit theorem for a weighted tail empirical process. A simulation study shows the good finite-sample behavior of our limit theorems. We also present applications to assess the tail heaviness of earthquake energies and cross-sectional stock market losses.

E0684: **High conditional quantiles for panel data**
Presenter: **Xuan Leng**, Xiamen University, China
Panel quantile regression models play an essential role in real finance, econometrics, insurance, and risk management applications. However, direct estimates of the extreme conditional quantiles may lead to unstable results due to data sparsity on the far tail. Moreover, the presence of individual effects in panel quantile regressions complicates the inference for high quantiles. A two-stage method is proposed to estimate/predict the high conditional quantiles. The intermediate quantiles are first predicted according to panel quantile regressions, and the extreme quantiles are obtained by extrapolating the intermediate ones in the second stage. The asymptotic properties of the prediction method rely on a set of second-order conditions for heteroscedastic extremes. We use a metric called Average Absolute Deviation Error to evaluate the prediction performance of high conditional quantiles over different cross-sections. The asymptotic distributions of the metric for both intermediate and extreme quantiles are studied. We demonstrate the finite sample performance of the two-stage prediction, which is compared to the direct prediction for extreme conditional quantiles. Finally, we apply the two-stage method to the macroeconomic and housing price data and find strong evidence of housing bubbles and common economic factors.

E0717: **Random networks with heterogeneous reciprocity**
Presenter: **Tiandong Wang**, Fudan University, China
Co-authors: Sid Resnick
Users of social networks display diversified behavior and online habits. For instance, a user's tendency to reply to a post can depend on the user and the person posting. For convenience, we group users into aggregated behavioral patterns, focusing here on the tendency to reply to or reciprocate messages. The reciprocity feature in social networks reflects the information exchange among users. We study the properties of a preferential attachment model with heterogeneous reciprocity levels, give the growth rate of model edge counts, and prove the convergence of empirical degree frequencies to a limiting distribution. This limiting distribution is not only multivariate, and regularly varying, but also has the property of hidden regular variation.

EO602 Room S0.03 ADVANCES IN TIME SERIES AND SPATIO-TEMPORAL DATA Chair: Soudeep Deb

E0470: **A Bayesian approach to identify changepoints in spatio-temporal ordered categorical data: An application to COVID-19**
Presenter: **Siddharth Rawat**, Indian Institute of Management Bangalore, India
Co-authors: Soudeep Deb, Candace Berrett
Although there is substantial literature on identifying structural changes for continuous spatio-temporal processes, that is not true for categorical spatio-temporal data. The purpose is to bridge that gap and propose a novel spatio-temporal model to identify changepoints in ordered categorical data. The model leverages an additive mean structure with separable gaussian space-time processes. Our proposed technique is defined in such a way that it can detect a shift in the mean structure as well as in the covariance structures in both the spatial and temporal associations. Our approach's capability to handle ordinal categorical data provides an added advantage from an application perspective. We implement the model through a Bayesian framework, which gives a computational edge over a classical method. For application, we use county-wise COVID-19 data from New York by categorizing the daily cases according to CDC guidelines. Our model is able to identify changepoints in the data and helps in providing interesting insights about the "waves" encountered during the pandemic.

E0925: **High-dimensional time series segmentation via factor-adjusted vector autoregressive modelling**
Presenter: **Hyeyoung Maeng**, Durham University, United Kingdom
Co-authors: Haeran Cho, Idris Eckley, Paul Fearnhead
Vector autoregressive (VAR) models are popularly adopted for modelling high-dimensional time series, and their piecewise extensions allow for structural changes in the data. In VAR modelling, the number of parameters grows quadratically with the dimensionality, which necessitates the assumption of sparsity in high dimensions. However, it is debatable whether sparse VAR models are adequate for handling datasets exhibiting strong serial and cross-sectional correlations. We propose a piecewise stationary time series model that simultaneously allows for strong correlations, as well as structural changes, where pervasive serial and cross-sectional correlations are accounted for by a (possibly) time-varying factor structure, and any remaining idiosyncratic dependence between the variables, is handled by a piecewise stationary, sparse VAR model. We propose an accompanying two-stage change point detection methodology which fully addresses the challenges arising from not observing either the factor-driven or the VAR processes directly. Its consistency in estimating both the total number and the locations of the change points in the latent components, is established under conditions considerably more general than those in the existing literature. We demonstrate the competitive performance of the proposed methodology on simulated datasets and an application to US blue chip stocks data.

E1714: **Effect of media attention on crude oil price volatility using a non-parametric time series regression**
Presenter: **Rishideep Roy**, Indian Institute of Management Bangalore, India
Co-authors: Soudeep Deb, Sayar Karmakar, Jyotishka Ray Choudhury
Many factors may affect the prices of crude oil sold across the globe. Media attention is a well-known factor that often plays a significant role in influencing crude oil prices during major global events. We are interested in estimating the volatility of prices of different kinds of light crude oil during the recent Russia-Ukraine War, using internet search data. We use the non-parametric stochastic regression model to estimate the mean and volatility functions, using the Nadaraya-Watson estimator. We also obtain simultaneous confidence bands for the same.

EO320 Room S0.11 TEXT MINING FOR SOCIAL IMPACT Chair: Laura Vana

E0852: **Measuring gender differences in personalities through natural language in the labor force**
Presenter: **Dania Eugenidis**, Justus-Liebig-University Giessen, Germany
Co-authors: David Lenz
Gender stereotypes still play a major role in the perception and representation of people in the workplace. Traditional measures, such as question-

naires, often lack objectivity and thus struggle to provide the full picture. However, evidence-based policymaking requires accurate indicators of gender inequalities to promote equality. This framework depicts the first-ever study examining the external portrayal of gender stereotypes on a company level using publicly available big data. Specifically, nearly one million company websites are called in using natural language processing. Shortcomings of traditional quantitative measures are to be overcome regarding timeliness, granularity and cost efficiency. That way, it is possible for the first time to conduct a fully automated, objective and almost comprehensive analysis of the linguistic portrayal of gender in a corporate context. A subsequent comparison to the literature takes place by contextualizing the gender stereotype measures following the personality traits of the Big Five-factor model and their sublevels. The results of the statistical analysis indicate significant stereotypes within personality traits for large portions of the sample. These differences in gender presentation are mostly consistent with those found in the literature, which serve as a validation for the presented framework.

E1106: **Detecting gender bias in children's textual literature**
Presenter: **Camilla Damian**, TU Wien, Austria
Co-authors: Laura Vana

Gender stereotypes form early in the child's development and are carried over throughout adolescence into adulthood, leaving long-lasting effects which may impact activity and career choices, as well as academic performance. Books, in particular, can have considerable influence, as their characters serve to shape role models of femininity and masculinity for young children. Thus, gender under- and misrepresentation in children's textual literature can contribute to the internalization and reinforcement of negative stereotypes. To address this issue, we aim to identify and measure relevant dimensions of gender bias in children's books with the aid of both qualitative and quantitative techniques: systematic literature review across disciplines, synthesis and (expert) validation on the one hand and state-of-the-art NLP methods on the other. By exploiting such an integrated research framework, we believe that we can automate the detection of potentially biased text while enhancing the interpretability and transparency of the results.

E1397: **Gender bias in text: Automated detection and mitigation system**
Presenter: **Jad Doughman**, American University of Beirut, Lebanon

Given that language is the primary tool used to convey our perceptions, then any form of biased misrepresentation has the potential to change how an entity is portrayed in our minds. The source of bias in language can be traced to an androcentric worldview that was prevalent among 18th-century grammarians and was centered around the belief that: "human beings were to be considered male unless proven otherwise". Given that there is clear evidence of gender bias in most languages and its direct contribution to reinforcing and socializing sexist thinking, then there is a need to detect and highlight these manifestations in the ever-growing repertoire of textual content on the internet alongside printed writings such as educational textbooks. Previously, most proposed solutions to detect gender bias in texts were based on the frequency of gendered words and pronouns, in contrast, our feature-based approach would focus on capturing contextual and semantic queues in its classification process. The underlying motivation is to enable the technical community to combat gender bias in text and halt its propagation using ML and NLP techniques.

EO726 Room S0.12 STATISTICAL MODELING AND MACHINE LEARNING WITH APPLICATIONS IN DATA SCIENCE Chair: Yousri Slaoui

E1787: **Clustering in attributed weighted nodes network using a stochastic block model with application to EEG data**
Presenter: **Yousri Slaoui**, University of Poitiers, France
Co-authors: Abir El Haj, Pierre-Yves Louis, Cyril Perret

The aim is to cluster networks with attributed weighted nodes. This question is motivated by the need to specify different electrophysiology stable periods performed by the brain during a psycho-linguistic experience, preparation of handwriting from the electrical activity produced by neurons in the brain and recorded by the electroencephalogram. The aim is to explore the evolution of the average intensity of the obtained clusters over time by classifying the 128 electrodes obtained by the electroencephalographic (EEG) recordings. We develop a stochastic block model (SBM) with several attributes to estimate the parameters of the model and to classify the nodes. Finally, we perform a numerical application using the electroencephalographic data to validate the proposed approach.

E1794: **Non-parametric recursive regression for Q estimation in actor-critic reinforcement learning**
Presenter: **Leo Grill**, Universite de Poitiers, France
Co-authors: Yousri Slaoui, Stephane Le Masson, David Nortershauser

An algorithm is presented to estimate the Q value for reinforcement learning in an actor-critic context. The non-parametric estimator learns recursively while the interactions between the actor and the environment generate the data. The estimator helps the convergence of the model with bias and variance control.

E1516: **Feedforward neural networks as statistical models**
Presenter: **Andrew McInerney**, University of Limerick, Ireland
Co-authors: Kevin Burke

Feedforward neural networks (FNNs) are typically viewed as pure prediction algorithms, and their strong predictive performance has led to their use in many machine-learning applications. However, quite simply, FNNs are non-linear regression models, where the covariates are mapped to the response through a series of weighted summations and non-linear functions. Their success in predictivity can be attributed, at least in part, to their ability to capture complex relationships through the modelling of higher-order interactions. However, their flexibility comes with an interpretability trade-off; thus, FNNs have been historically less popular among statisticians, who tend to use more interpretable additive models. Nevertheless, classical statistical theory, such as significance testing and uncertainty quantification, is still relevant for FNNs. Supplementing FNNs with methods of statistical inference, model selection and covariate-effect visualisations, can shift the focus away from black-box prediction and make FNNs more akin to traditional statistical models. This can pave the way towards more inferential analysis, and, hence, increase the utility of the FNN within the statistician's toolbox.

EO286 Room S0.13 TOPICS IN MULTIVARIATE MODELLING AND HIGH DIMENSION Chair: Benjamin Poignard

E0507: **Sparse factor model of high dimension**
Presenter: **Benjamin Poignard**, Osaka University, Japan
Co-authors: Yoshikazu Terada

The problem of estimating a factor model-based variance-covariance matrix is considered when the factor loading matrix is assumed sparse. We develop a penalized estimating function framework to account for the identifiability issue of the factor loading matrix while fostering sparsity in potentially all its entries. We prove the oracle property of the penalized estimator; that is, the penalization procedure can recover the true sparse support, and the estimator is asymptotically normally distributed. Consistency and support recovery are established when the number of parameters is diverging. These theoretical results are supported by empirical studies.

E1197: **Haar-weave-metropolis kernel**
Presenter: **Kengo Kamatani**, ISM, Japan
Co-authors: Xiaolin Song

Recently, many Markov chain Monte Carlo methods have been developed with deterministic reversible transform proposals inspired by the Hamil-

tonian Monte Carlo method. The deterministic transform is relatively easy to reconcile with the local information (gradient etc.) of the target distribution. However, as the ergodic theory suggests, these deterministic proposal methods seem to be incompatible with robustness and lead to poor convergence, especially in the case of target distributions with heavy tails. On the other hand, the Markov kernel using the Haar measure is relatively robust since it learns global information about the target distribution introducing global parameters. However, it requires a density-preserving condition, and many deterministic proposals break this condition. We carefully select deterministic transforms that preserve the structure and create a Markov kernel, the Weave-Metropolis kernel, using the deterministic transforms. By combining with the Haar measure, we also introduce the Haar-Weave-Metropolis kernel. In this way, the Markov kernel can employ the local information of the target distribution using the deterministic proposal, and thanks to the Haar measure, it can employ the global information of the target distribution. Finally, we show through numerical experiments that the performance of the proposed method is superior to other methods in terms of effective sample size and mean square jump distance per second.

EO482 Room Safra Lecture Theatre ADVANCES IN BAYESIAN COMPUTATION TECHNIQUES II Chair: Siew Li Linda Tan

E0193: **Weakly informative priors and prior-data conflict checking for likelihood-free inference**
Presenter: **David Nott**, National University of Singapore, Singapore
Co-authors: Atlanta Chakraborty, Michael Evans

Bayesian likelihood-free inference, which is used to perform Bayesian inference when the likelihood is intractable, enjoys an increasing number of important scientific applications. However, many aspects of a Bayesian analysis become more challenging in the likelihood-free setting. One example of this is prior-data conflict checking, where the goal is to assess whether the information in the data and the prior are inconsistent. We consider methods for prior-data conflict checking, which are applicable regardless of whether the likelihood is tractable or not. In constructing our checks, we consider checking statistics based on prior-to-posterior Kullback-Leibler divergences. The checks are implemented using mixture approximations to the posterior distribution and closed-form approximations to Kullback-Leibler divergences for mixtures, which make Monte Carlo approximation of reference distributions for calibration computationally feasible. When prior-data conflicts occur, it is useful to consider weakly informative prior specifications in alternative analyses as part of a sensitivity analysis. As the main application of our methodology, we develop a technique for searching for weakly informative priors in likelihood-free inference, where the notion of a weakly informative prior is formalized using prior-data conflict checks.

E1540: **Bayesian inference using synthetic likelihood: Asymptotics and adjustments**
Presenter: **Robert Kohn**, University of New South Wales, Australia
Co-authors: David Frazier, Christopher Drovandi, David Nott

Implementing Bayesian inference is often computationally challenging in complex models, especially when calculating the likelihood is difficult. Synthetic likelihood is one approach for carrying out inference when the likelihood is intractable, but it is straightforward to simulate from the model. The method constructs an approximate likelihood by taking a vector summary statistic as being multivariate normal, with the unknown mean and covariance estimated by simulation. Previous research demonstrates that the Bayesian implementation of synthetic likelihood can be more computationally efficient than approximate Bayesian computation, a popular likelihood-free method, in the presence of a high-dimensional summary statistic. Three contributions are made. The first shows that if the summary statistics are well-behaved, then the synthetic likelihood posterior is asymptotically normal and yields credible sets with the correct level of coverage. The second compares the computational efficiency of Bayesian synthetic likelihood and approximate Bayesian computation. We show that Bayesian synthetic likelihood is computationally more efficient than approximate Bayesian computation. Based on the asymptotic results, the third proposes using adjusted inference methods when a possibly misspecified form is assumed for the covariance matrix of the synthetic likelihood, such as a diagonal or a factor model, to speed up computation.

E1623: **Bayesian inference for the Conway-Maxwell distribution**
Presenter: **Luiza Piancastelli**, University College Dublin, Ireland
Co-authors: Nial Friel

The Conway-Maxwell-Poisson (COM-Poisson) distribution is a two-parameter generalisation of the Poisson distribution, which accommodates under- and over-dispersion. A main difficulty with its use is that it involves an intractable normalising constant. We provide some techniques to overcome this issue and illustrate how this applies in the context of GLM regression with a COM-Poisson response. We also highlight a multivariate extension.

EO543 Room Virtual R02 NEW CHALLENGES IN DESIGN OF EXPERIMENTS II Chair: Victor Casero-Alonso

E1057: **Approximate Laplace importance sampling for Bayesian design of experiments in nonlinear models**
Presenter: **Tim Waite**, University of Manchester, United Kingdom
Co-authors: David Woods, Yiolanda Englezou

One of the major challenges in Bayesian optimal design is to approximate the expected utility function in an accurate and computationally efficient manner. We explore the performance of nested Monte Carlo methods for approximating the expected Shannon information gain. We emphasise Laplace Importance Sampling (LIS) and a new cheaper counterpart, Approximate Laplace Importance Sampling (ALIS). Both methods are thoroughly compared with existing approximations, including Double Loop Monte Carlo, nested importance sampling, and Laplace approximation, on a range of examples common in the Statistics literature. It is found that LIS and ALIS give an efficient trade-off between mean squared error and computational cost for utility estimation. We also show that LIS and ALIS give improved designs compared to existing methods in problems with large numbers of model parameters when combined with the approximate co-ordinate exchange algorithm for design optimization.

E1291: **Multi-objective optimisation of split-plot designs**
Presenter: **Kalliopi Mylona**, King's College London, United Kingdom
Co-authors: Matteo Borrotti, Francesco Sambo

Scientists can now address scientific issues of increasing complexity thanks to modern experiments. Often, factors in experiments have levels that are more difficult to set than others and in these cases, using a split-plot design offers a solution. Numerous approaches to finding optimal designs focus on maximising a specific criterion. To tackle the drawbacks of one-objective optimisation, multi-criteria techniques have been developed; however, they mostly concentrate on measuring the precision of fixed factor effects, ignoring the estimation of the variance components in split-plot experiments. To achieve pure-error estimates of the variance components, the Multi-Stratum Two-Phase Local Search (MS-TPLS) algorithm for multi-objective optimisation of experimental designs is expanded. The best Pareto front and associated designs, for motivating examples, are evaluated against other designs from the literature. According to the findings, the designs derived from the Pareto front are strong candidates for solutions based on the various objectives.

E1266: **Optimal designs for fractional polynomial models**
Presenter: **Victor Casero-Alonso**, University of Castilla-La Mancha, Spain
Co-authors: Jesus Lopez-Fidalgo, Chiara Tommasi, WengKee Wong

The aim is to show design issues for Fractional Polynomials (FP) models. We have constructed D- and I-optimal designs for estimating model parameters and prediction in FP models with a single factor, and we have developed an applet to facilitate users in finding various types of

tailor-made optimal designs for their problems. We have considered three studies that used experimental design to model risk assessments fitted by FP models, to illustrate the usefulness of the optimal experimental design. Furthermore, there is interest in obtaining KL-optimal designs to discriminate between two FP models, even among several FP models. Algorithms are needed to achieve that goal. We have adapted the nature-inspired Particle Swarm Optimization algorithm to obtain such designs

EO258 Room K2.31 (Nash Lec. Theatre) IDENTIFICATION AND EFFICIENT ESTIMATION IN CAUSAL INFERENCE Chair: Tetiana Gorbach

E0508: **Clustering and structural robustness in causal diagrams**
Presenter: **Santtu Tikka**, University of Jyvaskyla, Finland

Clustering of vertices is considered in directed acyclic graphs that represent structural causal models. Clustering can often clarify the visual representation of the causal model, but arbitrary clustering might break important causal connections. We define a specific type of cluster, called transit cluster, and show that under the corresponding clustering, the graph retains important properties related to causal effect identifiability if specific assumptions hold. We further show that a subset of transit clusters, called transit components, can be found efficiently, and that any transit cluster can be represented as a union of such components. We also consider the inverse problem, where one begins with a clustered graph and looks for larger graphs from which the original graph may have been obtained as a result of clustering.

E1577: **Demystifying statistical learning based on efficient influence functions**
Presenter: **Oliver Hines**, London School of Hygiene and Tropical Medicine, United Kingdom
Co-authors: Oliver Dukes, Karla DiazOrdaz, Stijn Vansteelandt

Evaluation of treatment effects and more general estimands is typically achieved via parametric modelling, which is unsatisfactory since model misspecification is likely. Data-adaptive model building (e.g. statistical/machine learning) is commonly employed to reduce the risk of misspecification. However, naive use of such methods delivers estimators whose bias may shrink too slowly with sample size for inferential methods to perform well, including those based on the bootstrap. Bias arises because standard data-adaptive methods are tuned towards minimal prediction error as opposed to, e.g. minimal MSE in the estimator. This may cause excess variability that is difficult to acknowledge, due to the complexity of such strategies. Building on results from non-parametric statistics, targeted learning and debiased machine learning overcome these problems by constructing estimators using the estimand's efficient influence function under the non-parametric model. These increasingly popular methodologies typically assume that the efficient influence function is given, or that the reader is familiar with its derivation. We focus on the derivation of the efficient influence function and explain how it may be used to construct statistical/machine-learning-based estimators. We discuss the requisite conditions for these estimators to perform well and use diverse examples to convey the broad applicability of the theory.

E1596: **Discussion**
Presenter: **Xavier de Luna**, Umea University, Sweden

Research presented in this session will be discussed, focusing on identification and efficiency issues in causal inference.

EO486 Room K2.40 FUNCTIONAL TIME SERIES: THEORY AND APPLICATIONS Chair: Yanrong Yang

E1593: **Does climate sensitivity differ across regions? A varying-coefficient approach**
Presenter: **Yang Yang**, University of Newcastle, Australia
Co-authors: Heather Anderson, Jiti Gao, Farshid Vahid, Wei Wei

The global mean surface temperature has been increasing in the last six decades in response to growing greenhouse gas concentrations. While Earth is getting warmer globally, local regions are observed to experience unequal increases in temperature. We measure climate sensitivity in various land regions around the world with a dynamic varying-coefficient panel data model and spatial-temporal climate data. The proposed inference method can accommodate heterogeneous co-integration relationships between global and local variables, allowing co-moving climate time series to possess stochastic and deterministic trending components and spatial-temporal dependence. Applied to observational data of mean surface temperatures, solar radiation, and carbon dioxide concentrations between 1959-2017, our model reveals an estimate of a 3.7-degree increase in global land temperature after a doubling of CO2 concentrations. Moreover, our empirical estimates indicate that high-latitude regions in the Northern Hemisphere are most vulnerable to climate warming.

E1641: **Identifying features from practical FPCA on functional time series**
Presenter: **Yuan Gao**, The Australian National University, Australia
Co-authors: Yanrong Yang, Han Lin Shang, Yang Yang

As a typical dimension-reduction tool, functional principal component analysis (FPCA) extracts features for functional data in terms of the sample covariance operator. What kind of features does FPCA produce? Under a general separable covariance structure, we show that this set of FPCA features may include principal components of the population covariance structure (i.e., cross-sectional common variation), basis functions of the nonstationary subspace (i.e. temporal common movement), and their mixture. We provide asymptotic results for the sub-space expanded by these features. We also construct an alternative algorithm to differentiate the two kinds of features and demonstrate this by applying it to the U.S. mortality rates and the global temperature data.

E1754: **A simulation and estimation algorithm for residual correlation analysis of long-range dependence (LRD) FANOVA models**
Presenter: **Diana Paola Ovalle-Munoz**, University of Granada, Spain
Co-authors: Maria Dolores Ruiz-Medina

A simulation algorithm is derived for the generation of multifractional time series models in the context of infinite-dimensional processes. In particular, multifractionally integrated functional autoregressive moving average processes are generated, displaying different orders of long-range dependence (LRD) in time at each spatial resolution level. Applying previous results by Ruiz-Medina on weak-consistent minimum contrast parameter estimation based on the periodogram operator, an estimation algorithm is implemented to approximate the long-memory operator, characterizing the strong dependence structure of these models. The application of these algorithms in the implementation of residual correlation analysis in the context of FANOVA models with multifractionally integrated functional autoregressive moving average noise is illustrated as well.

EC828 Room S-1.04 MACHINE LEARNING II Chair: Bettina Gruen

E0375: **Unsupervised topic identification in large short text corpora using mixture models**
Presenter: **Jocelyn Mazarura**, University of Pretoria, South Africa
Co-authors: Alta De waal, Pieter De Villiers

Topic modelling is a subfield of natural language processing whose objective is to discover latent topics in large unlabelled corpora. Over the years, short texts, such as tweets and reviews, have become increasingly relevant due to the growing popularity of social media and online shopping. Traditional topic models assume that a document is generated from multiple topics. Whilst this assumption may be acceptable for long texts, such as e-books and news articles, many studies have shown that the one-topic-per-document assumption imposed by mixture models, such as the Dirichlet-multinomial mixture (DMM) model, fits short texts better. Most topic models are constructed under the assumption that documents follow a multinomial distribution. The Poisson distribution is an alternative distribution to describe the probability of count data. It has been successfully applied in text classification, but its application to topic modelling is not well documented, specifically in the context of a generative probabilistic

model. The main contributions are a new Gamma-Poisson mixture (GPM) model and a collapsed Gibbs sampler, which enables the model to learn the number of topics contained in the corpus automatically. The results show that the GPM performs better than the DMM at selecting the number of topics in labelled corpora. Furthermore, the GPM produces better topic coherence scores, thus making it a viable option for the challenging task of topic modelling of short text.

E0626: **Merged linear Gaussian cluster-weighted models: An interpretable machine learning model**
Presenter: **Sangkon Oh**, Sungkyunkwan University, Korea, South
Co-authors: Byungtae Seo

Cluster-weighted models (CWMs) are useful tools for identifying latent functional relationships between response variables and covariates. However, owing to excess distributional assumptions made on the covariates, these models can suffer misspecifications of component distributions, which could also undermine the estimation accuracy and render the model structure complicated for interpretation. To address this issue, we consider CWMs with univariate responses and propose a novel CWM by modelling each regression cluster as a finite mixture to enhance flexibility while retaining parsimony. We prove that the proposed method can provide more meaningful regression clusters in the data than those of existing methods and not only has good prediction accuracy when predicting new data but is also interpretable. Additionally, we present a procedure to construct such a proposed CWM and a feasible expectation-maximization algorithm to estimate the model parameters. Numerical demonstrations, including simulations and real data analysis, are also provided.

E1751: **From Gutenberg to BERT: How transformers can change information extraction from text**
Presenter: **Daniela Ushizima**, Lawrence Berkeley National Laboratory / UC San Francisco, United States
Co-authors: Eric Chagnon

Around 1440, Gutenberg revolutionized knowledge dissemination with the advent of the printing press using efficient mechanical devices. Over a half millennium later, access to text has changed from scarcity to overly abundant: the main challenge became how to distill information from huge amounts of textual data. Extracting knowledge from text has undergone a technological upheaval with text mining and natural language processing, but how have these innovations affected scientific activities, such as literature review? Our efforts toward designing algorithms for topic modeling and content recommendation, are describedd given large sets of scientific articles. By using deep learning models, such as the Bidirectional Encoder Representations from Transformers (BERT), our python-based code has been turning text data into information that helps us to identify key topics within different science domains, for example, the most relevant technologies for materials analysis given a set of laboratories. The main advantages are: high-level mechanisms for I/O, semantic similarity among articles that enable recommendations, and topic word scores and evolution of topics over time for quick feedback using visualization.

EC820 Room K0.16 GRAPHICAL MODELS Chair: Natalia Bochkina

E1638: **Bayesian structure learning in undirected graphical models: Review and empirical comparisons**
Presenter: **Lucas Vogels**, University of Amsterdam, Netherlands
Co-authors: Reza Mohammadi, Ilker Birbil, Marit Schoonhoven

Graphical models are an elegant way to depict the conditional dependencies among variables using a graph. Bayesian structure learning is the area occupied with revealing the structure of this graph using Bayesian methods. Although multiple solution methods have been proposed in this field over the last decade, no comprehensive review or empirical comparison is available. This review is presented. We will list and classify all methods, compare their performance in a simulation study, and give suggestions for future research.

E1891: **Consistent model selection and elicited prior information: The posterior equivalence principle for chain event graphs**
Presenter: **Peter Strong**, University of Warwick, United Kingdom
Co-authors: Jim Smith

When using elicited prior parameter conditional distributions, current Bayesian model selection techniques can lead to a lack of consistency between how the elicited information and the data are treated. We propose a solution: the posterior equivalence principle. This satisfies the condition where, when performing model selection with the same set of models and the same parameter posterior distribution for each model, the distribution over the set of models should be the same. We demonstrate how we can satisfy this condition for the Bayesian Dirichlet score on Chain Event Graphs by setting a structural prior.

E1901: **A data-driven Bayesian graphical ridge estimator**
Presenter: **Jarod Smith**, University of Pretoria, South Africa
Co-authors: Mohammad Arashi, Andriette Bekker

Bayesian methodologies prioritising accurate associations above sparsity in Gaussian graphical model (GGM) estimation remain relatively scarce in scientific literature. It is well accepted that the l2 penalty enjoys a smaller computational footprint in GGM estimation, whilst the l_1 penalty encourages sparsity in the estimand. The Bayesian adaptive graphical lasso prior is used as a departure point in the formulation of a computationally efficient graphical ridge-type prior for events where accurate associations are prioritised over sparse representations. A novel block Gibbs sampler for simulating precision matrices is constructed using a ridge-type penalisation. The Bayesian graphical ridge-type prior is extended to a Bayesian adaptive graphical ridge-type prior. Synthetic experiments indicate that the graphical ridge-type estimators enjoy computational efficiency, in moderate dimensions, and numerical performance, for relatively non-sparse precision matrices, when compared to their lasso counterparts. The adaptive graphical ridge-type estimator is applied to cell signalling data to infer key associations between phosphorylated proteins in human T-cell signalling. All computational workloads are carried out using the baygel R package.

EC679 Room BH (S) 2.05 BAYESIAN STATISTICS II Chair: Asaf Weinstein

E1809: **Predictions for the gamma distribution model and information geometry of Levy measures**
Presenter: **Fumiyasu Komaki**, The University of Tokyo, Japan

Some properties of predictive densities for the Gamma distribution and the inverse Gaussian distribution models are discussed. The performance of predictive densities is evaluated by the Kullback-Leibler divergence. For the normal and Poisson models, the correspondence between prediction and parameter estimation has played an essential role in prediction theory. On the other hand, such a simple relationship does not hold for the Gamma model and the inverse Gaussian model, and estimation of the Levy measures of the corresponding subordinators plays a role corresponding to the parameter estimation in the normal model and the Poisson model. The relationship between prediction for the Gamma and inverse Gaussian models and information geometry of the space of Levy measures is also discussed.

E1846: **Simultaneous analysis in Bayesian multidimensional unfolding with application to party expert surveys**
Presenter: **Kodai Tachibana**, University of Tokyo, Japan
Co-authors: Junko Kato, Kensuke Okada

Multidimensional unfolding has been applied as a spatial model of choice and judgment in social sciences. However, when estimating the coordinate parameters in an ordinary multidimensional unfolding model with the Markov chain Monte Carlo algorithm, the problem of indeterminacy arises, preventing naive implementation from working properly. This problem occurs because rotation, reflection, and translation of the configuration matrix do not change the likelihood. One of the well-known methods to manage this issue is fixing the required number of coordinates. This

approach, however, causes a problem in which it is no longer possible to estimate the uncertainty of the fixed coordinates and a problem in which arbitrary constraints must be made. We propose a different approach using equality constraints in a case in which two datasets are simultaneously analyzed. The proposed constraint allows us to estimate all the relevant parameters under realistic model assumptions. The proposed method is illustrated with expert survey datasets on political parties. Concretely, the spatial representation of the parties and survey items are simultaneously estimated by constraining the party positions to be equal between the two analyzed datasets. The utility and applicability of the proposal are discussed.

E1832: **Bayesian hierarchical functional regression model for ordinal responses with flexible link functions**
Presenter: **Jangwon Lee**, Korea University, Korea, South
Co-authors: Taeryon Choi

In the field of functional regression research, the study of ordinal functional response has many limitations compared to other types of functional responses. We propose a Bayesian hierarchical functional regression model with flexible link functions. To derive a posterior sampling scheme, we use a latent variable that is categorized by the cut-off points. The latent variable is modeled by a Bayesian functional mixed effect model based on the spectral representation of Gaussian processes. We assume a hierarchical structure on the spectral coefficient to deal with a global mean curve, group curves, and subject-specific curves. In ordinal regression, probit or logit are commonly used in a link function. But these link functions correct the skewness of response probability. To overcome this limitation, we adopt various types of link functions, such as complementary-log-log, generalized extreme value, and symmetric power link functions. Also, we consider a shape restriction on the nonparametric terms, such as monotone increasing and monotone decreasing, to better predict the probability of a response. We illustrate simulation examples and real applications to air quality index (AQI) data.

EC826 Room K2.41 STATISTIC FOR COVID Chair: Irene Garcia-Camacha Gutierrez

E0350: **Multiscale decomposition of spatial lattice data for detecting hotspots of COVID-19 cases in South Africa**
Presenter: **Rene Stander**, University of Pretoria, South Africa
Co-authors: Inger Fabris-Rotelli, Din Chen

During a pandemic such as COVID-19, it is important to know where positive cases are clustered for local governments to implement measures to control the spread of the disease. The detection of such hotspot areas is an important part of spatial analysis. Several methods have been used, such as measures for local spatial association and spatial scan statistics. We propose the use of the Discrete Pulse Transform (DPT) on spatial lattice data along with the multiscale Ht-index as a measure of saliency on the extracted pulses to detect significant hotspots.

E1121: **The value of the New Hampshire birth cohort: Impact of SARS Covid 2 on children's respiratory infections and symptoms**
Presenter: **Susana Diaz Coto**, Dartmouth College, United States
Co-authors: Janet Peacock, Vicki Sayarath, Juliette Madan, Margaret Karagas

Systematic data collection is one of the keys of scientific progress. Nowadays, these data frequently involve diverse sources and types of information, including microbiome, genetic, epigenetic and/or phenotypes, among others. The storage and harmonization of all these variables represent a non-trivial task, which deserves critical thinking and adequate systematic processes. The New Hampshire Birth Cohort Study (NHBCS) systematically collects a huge quantity of information related to mothers and newborns in New Hampshire (northeast of US) with the goal of examining associations between environmental exposures, maternal/child characteristics, and maternal/child outcomes. The NHBCS is part of the Environmental influences on Child Health Outcomes (ECHO) program that integrates 71 different but similar cohorts in US, and with more than 50,000 children enrolled is the largest US study of its kind. On the one hand, we present the technical structure and processes developed in the NHBCS to have a high-quality dataset; and, on the other hand, we illustrate the practical application for providing valuable knowledge on real-world issues. Particularly, we have used the NHBCS to study the impact of the virus SARS-Cov-2 on upper and lower tract respiratory infections and symptoms.

E2005: **COVID-19 hospitalization and mortality after conditioning on osmolality and the effect of CA and Vitamin D**
Presenter: **Ayse Ulgen**, Nottingham Trent University, United Kingdom
Co-authors: Hakan Sivgin, Sirin Cetin, Meryem Cetin, Wentian Li

Osmolality, the concentration of solute particles, can be estimated from the measurement of three other blood test variables (sodium, glucose, and urea), and was rarely used for a prognosis for COVID-19. As a result of the analysis of the data obtained from COVID-19 patients, we found osmolality to be an excellent prognostic biomarker for both mortality and hospitalization. On the other hand, both hypocalcemia and vitamin-D deficiency are also significantly associated with a higher mortality rate in our data, while only calcium and not vitamin D is associated with a higher hospitalization rate. Different types of tests, including logistic regression, t-test, Wilcoxon test, lead to the same conclusion. After conditioning on osmolality in multiple logistic regression, calcium level remains to be significantly associated with both mortality and hospitalization (p-value <0.001). However, vitamin D loses its association with mortality when conditioning on osmolality (significant only at 0.05 level).

CV779 Room Virtual R01 APPLIED ECONOMETRICS Chair: Radu Craiu

C0337: **Predictive performance of Bayesian VEC-SV-GARCH models before and during the Covid-19 pandemic**
Presenter: **Lukasz Kwiatkowski**, Krakow University of Economics, Poland
Co-authors: Justyna Wroblewska, Anna Pajor

The main aim is to check whether taking into account long-term relations in heteroscedastic VAR models affects their predictive performance over the period of the pandemic. Additionally, we check whether updating the posterior upon the arrival of new observations affects the predictive performance of the models before and during the pandemic. In the empirical analysis, the so-called small model of monetary policy is considered separately for five economies: the United States, the United Kingdom, the Euro Area, Poland and Hungary. The heteroscedasticity is captured by means of hybrid specifications combining stochastic volatility and GARCH (SV-GARCH), or some of the special cases thereof. Estimation and prediction are performed within the Bayesian approach, with a focus on the evaluation and comparison of the models' predictive performance by means of predictive likelihood. The results indicate that allowing for conditional heteroskedasticity enhances the VEC models' predictive performance both before and during the pandemic. Also, in most cases, not updating the posterior decreases the predictive ability of models before as well as during the pandemic. For most developed economies, including long-term relations does not improve forecasts for long horizons over the pandemic. However, in most cases incorporating long-term relations improves forecasts for shorter horizons over the pandemic.

C1591: **Asymmetry and interdependence when evaluating U.S. energy information agency forecasts**
Presenter: **Yunyi Zhang**, Xiamen University, China

The purpose is to evaluate US Energy Information Agencies (EIA) forecasts of the world petroleum market, emphasising the importance of taking a multivariate perspective, considering asymmetric loss and allowing for time-variation. Forecasts for total demand, total supply, total stock withdrawals and oil prices are biased, with biases that change over time and differ across variables. A loss function that takes into account asymmetry and interdependence can rationalise these biases. The implied asymmetric loss gives less weight to the under-prediction of both demand and supply, while for oil prices, we document significant regime changes in the implied loss due to asymmetry. The EIA forecasts dominate a simple random walk benchmark when evaluated using symmetric and independent loss in the form of MSE statistical criteria. Yet, when allowing for asymmetry and interdependence that rationalize the EIA forecasts, the performance of the EIA forecasts worsens and is comparable to the random walk benchmark.

C1820: **Forecasting sovereign CDS prices**
Presenter: **Vineet Upreti**, Swansea University, United Kingdom
Co-authors: Marco Realdon
A key question in forecasting term structures of interest rates and credit spreads is the need and merit of using pricing models that impose the absence of arbitrage across different maturities. This question is addressed in forecasting the CDS prices of nineteen major and diverse sovereigns. Arbitrage-free affine pricing models forecast future term structures of sovereign CDS prices better than arbitrage-prone popular dynamic Nelson-Siegel (DNS) forecasting models. This is the case for both one-week and one-month forecast horizons. Affine pricing models outperform because they are arbitrage-free and often have more flexibility to match the present term structure of CDS prices closely. Simple autoregressions best forecast CDS prices at the one-week forecast horizon, because they perfectly match the present term structure of CDS prices but produce the worst forecasts at the one-month horizon.

CO692 Room Virtual R03 RECENT ADVANCES IN FINANCIAL ECONOMETRICS Chair: Jiajing Sun

C0888: **Stock market volatility forecasting: Can interval data improve it?**
Presenter: **Meiting Zhu**, University of Chinese Academy of Sciences, China
Co-authors: Yongmiao Hong, Shouyang Wang, Zishu Cheng, Jiani Heng
Estimating and forecasting stock volatility is critical to portfolio allocation, asset pricing, and risk management for market participants to make investment decisions and for policymakers to make economic policy. We study the interval-valued models to check whether these models can improve the forecast accuracy in stock volatility, especially the threshold autoregressive interval (TARI) model that utilizes the nonlinear information of the interval-valued stock indexes. Our data sample consists of 19 developed and emerging stock market indices. We find that the TARI model is superior in stock volatility forecasts to other point-based models GARCH and TGARCH models, as well as the conditional autoregressive range (CARR) model proposed based on the range data. As the forecast horizon increases, the difference in forecast accuracy of most market indices becomes statistically indistinguishable. To evaluate the practical implications of our findings, we study a portfolio problem, which reveals that asset allocation based on the interval model forecasts outperforms asset allocation based on other competing models.

C1311: **Normal mixture quasi-maximum likelihood estimation of double autoregressive models**
Presenter: **Christina Dan Wang**, NYU Shanghai, China
The estimation of a double autoregressive model (DAR) with skewed and heavy-tailed innovation is investigated. A new estimation method, the normal mixture quasi-maximum likelihood estimation (NM-QMLE), is proposed to estimate the DAR model with the non-Gaussian behavior. Under regularity conditions, consistency and asymptotic normality are established for NMQMLE. The numerical simulation for the DAR model with heavy-tailed and skewed innovation indicates that the NMQMLE outperforms several commonly adopted QMLE's. Finally, An empirical example on the S&P 500 index illustrates the application of the new estimation method.

C1665: **Adjusted-range-based Kolmogorov-Smirnov type statistics for structural breaks and parameter constancy**
Presenter: **Jiajing Sun**, University of Chinese Academy of Sciences, China
Co-authors: Yongmiao Hong, Brendan McCabe, Shouyang Wang
A self-normalization approach is proposed based on the adjusted range of a partial sum, which is robust to those irregularities and helps to rectify the better size but less power phenomenon for existing self-normalized statistics. We also introduce adjusted-range-based Kolmogorov-Smirnov (KS) type statistics to test for structural breaks in the mean, approximately linear statistics in general, and correlation coefficients/matrix. Furthermore, our proposed test statistics are portmanteau and can cater for general alternatives. Under suitable conditions, it can also be applied to detect the constancy of parameters. Monte Carlo simulations and empirical studies demonstrate the adequacy of our proposed method.

CO224 Room Virtual R04 MACHINE LEARNING: NEW DEVELOPMENTS Chair: Qingliang Fan

C0201: **L2-relaxation: With applications to forecast combination and portfolio analysis**
Presenter: **Zhentao Shi**, CUHK, Hong Kong
Forecast combination with many forecasts or minimum variance portfolio selection with many assets is tackled. A novel convex problem called L2-relaxation is proposed. In contrast to standard formulations, L2-relaxation minimizes the squared Euclidean norm of the weight vector subject to a set of relaxed linear inequality constraints. The magnitude of relaxation, controlled by a tuning parameter, balances the bias and variance. When the variance-covariance (VC) matrix of the individual forecast errors or financial assets exhibits latent group structures a block equicorrelation matrix plus a VC for idiosyncratic noises, the solution to L2-relaxation delivers roughly equal within-group weights. Optimality of the new method is established under the asymptotic framework when the number of the cross-sectional units N potentially grows much faster than the time dimension T. Excellent finite sample performance of our method is demonstrated in Monte Carlo simulations. Its wide applicability is highlighted in three real data examples concerning empirical applications of microeconomics, macroeconomics, and finance.

C0207: **Managers versus machines: Do algorithms replicate human intuition in credit ratings?**
Presenter: **Gabriel Vasconcelos**, Pontifical Catholic University of Rio de Janeiro and BOCOM BBM, Brazil
Co-authors: Matthew Harding
Machine learning techniques are used to investigate whether it is possible to replicate the behavior of bank managers who assess the risk of commercial loans made by a large commercial US bank. Even though a typical bank already relies on an algorithmic scorecard process to evaluate risk, bank managers are given significant latitude in adjusting the risk score in order to account for other holistic factors based on their intuition and experience. We show that it is possible to find machine learning algorithms that can replicate the behavior of bank managers. The input to the algorithms consists of a combination of standard financials and "soft" information available to bank managers as part of the typical loan review process. We also document the presence of significant heterogeneity in the adjustment process that can be traced to differences across managers and industries. Our results highlight the effectiveness of machine learning-based analytic approaches to banking and the potential challenges to high-skill jobs in the financial sector.

C0491: **Time-varying minimum variance portfolio**
Presenter: **Qingliang Fan**, The Chinese University of Hong Kong, Hong Kong
A new time-varying minimum variance portfolio (TV-MVP) is proposed in a large investment universe of assets. Our method extends the existing literature on minimum variance portfolios by allowing for time-varying factor loadings, which facilitates the capture of the dynamics of the covariance structure of asset returns (and hence, the optimal investment strategy in a dynamic setting). We also use a shrinkage estimation method based on a quasi-likelihood function to regularize the residual covariances further. We establish the desired theoretical properties of the proposed time-varying covariance and the optimal portfolio estimators under a more realistic heavy-tailed distribution. Specifically, we provide consistency of the optimal Sharpe ratio of the TV-MVP and the sharp risk consistency. Moreover, we offer a test of constant covariance structure and show the asymptotic distribution of the test statistic. Simulation and empirical studies suggest that the performance of the proposed TV-MVP is superior, in terms of estimation accuracy and out-of-sample Sharpe ratio, compared with that of other popular contemporary methods.

C2021: **Deep learning with non-linear factor models: Adaptability and avoidance of curse of dimensionality**
Presenter: **Maurizio Daniele**, ETH Zurich, KOF Swiss Economic Institute, Switzerland

Co-authors: Mehmet Caner

Deep learning literature is connected with non-linear factor models. We show that deep learning estimation leads to a substantial improvement in the non-linear factor model literature. We provide bounds on the expected risk and prove that these upper bounds are uniform over a set of multiple response variables. We extend our results to an additive model setting and show its connection to non-linear factor models for financial applications. Compared to traditional factor models, which assume rigid linear relations between the factors and the observed variables, our deep neural network factor model (DNN-FM) offers major improvements in modeling flexibility. We develop a novel data-dependent estimator of the error covariance matrix in deep neural networks and prove that the estimator is consistent in spectral norm. Moreover, we show the consistency and provide the rates of convergence of the covariance matrix and precision matrix estimators for asset returns. The rates of convergence do not depend on the number of factors. Various Monte Carlo simulations confirm our large sample findings and reveal superior accuracies of the DNN-FM in estimating the true underlying functional form, as well as the covariance and precision matrix compared to competing approaches. Moreover, in an out-of-sample portfolio forecasting application, it outperforms, in most cases, alternative portfolio strategies in terms of out-of-sample portfolio standard deviation and Sharpe ratio.

CO102 Room Virtual R05 THE ECONOMETRICS OF BANKING AND FINANCE Chair: Leone Leonida

C1890: **Political replacement effect and financial development: Evidence across countries**

Presenter: **Alfonsina Iona**, Queen Mary University of London, United Kingdom

Co-authors: Leone Leonida, Dawit Zerihun Assefa

The Politics and Finance literature argues that political factors are responsible for shaping a country's financial development. In line with this view, we analyse the impact of political competition on financial development across 124 countries over the period 1970-2015. The results show that political competition Granger causes financial development, and there is a political replacement effect of political competition on financial development. However, the relationship between political competition and financial development is U-shaped. The results are robust to the origin of a country's legal system, to different subsamples and alternative measures of financial development and political competition.

C1889: **Market-driven securitization**

Presenter: **Eleonora Muzzupappa**, Kingś College London, United Kingdom

How, and how much, does the stock market's performance affect banks' securitization activity? The analysis of a panel of EU and US banks shows that the former shapes the latter both directly and by interacting with some balance-sheet items. We find that the impact of the stock market performance upon the banks' securitization, the channels with which it interacts with the balance sheet items and the sign that these impacts take depend upon the market discipline, that shapes both the banks' business model of securitization, and the condition of the financial market.

C1869: **Is monotonicity of the investment-cash flow sensitivity satisfied? Evidence on a joint hypothesis**

Presenter: **Leone Leonida**, King's College London, United Kingdom

The considered question is whether the monotonicity of the investment-cash flow sensitivity is empirically satisfied. We propose an analysis of the so-called monotonicity condition that sidesteps the major uncertainties that must be faced when it comes to sorting firms according to the degree of financing constraints. We show that, because the true degree of financing constraints is unobservable, imposing the sorting scheme at the outset to the sample puts at risk the conclusion about whether the condition holds. This leads to considering this uncertainty and testing a joint null hypothesis. We show that, if the sample is appropriately sorted and monotonicity holds, then the point of sample separation does not affect the monotonic relationship between the observable average sensitivities of any two complementary classes of observations. We test this property by building on the most common metrics of the degree of financing constraints. We show that (1) there exists a monotonic relation among the sorting metrics; (2) the monotonicity condition is not empirically met as there exists a non-monotonic relation between ICFS and the true degree of financing constraints; (3) the ICFS is inverse basin-shaped and, finally, (5) this basin shape encompasses all the shapes documented by previous studies.

CO561 Room BH (SE) 1.02 USING LARGE DATASETS TO ANALYSE HOUSEHOLD FINANCE Chair: Jonathan Crook

C1561: **Identifying current account risk profiles to detect suspicious accounts**

Presenter: **Rui Ying Goh**, University of Edinburgh, United Kingdom

Co-authors: Galina Andreeva, Yi Cao, Johannes de Smedt

Anomalous financial behaviour signals a high exposure risk to suspicious activities, e.g. money laundering. Financial institutions build Know Your Client (KYC) profiles as the due diligence to first investigate typical account behaviour and then flag anomalous accounts which highly deviate from the typical ones. A two-stage approach is developed to assess anomaly risk from current account cash flow transactions. The first stage focuses on cluster analysis to explore normal current account profiles from the RFMP (Recency, Frequency, Monetary, Persistence) dimensions of transactional activities. The P dimension extends the popular RFM customer lifetime value marketing framework, to spot irregularities and detect anomalies. In the second stage, we evaluate the accounts with anomaly scores and examine the characteristics of high-risk accounts from two-dimensional visualisation plots and bag-of-word analysis of the transaction descriptions. The results reveal that the typical account profiles portray personal financial behaviour under different financial circumstances for day-to-day spending or saving purposes. We highlight potential suspicious account behaviour i.e., unusually large number of irregular or over-persistent transactions and disproportionate spending on transfer payments. These findings enhance KYC profiling, where financial institutions can spot accounts with high anomaly risk, before escalating the account owner's profile for further financial checks.

C1772: **Statistical properties of measurement error in earnings in labor market survey data**

Presenter: **Stella Martin**, University of Muenster, Germany

Co-authors: Kevin Stabenow, Mark Trede

Large-scale surveys on income play an important role in empirical economic research while being subject to a measurement error little is known about. Little literature, almost entirely based on one small linkage of employee and employer information on earnings, reveals that there might, in fact, be a systematic bias in self-reports of earnings. We use a novel linkage of survey and administrative data. The German Socioeconomic Panel spanning 37 waves (1984 to 2020) and data from the German Pension Insurance on individuals' entire earnings records provide a comprehensive panel on employment biographies and earnings information. We use these data to investigate the statistical properties of the measurement error in earnings on an individual and household level, where, unlike previous work, the rich panel structure in our data allows us to focus not only on the distribution of measurement error in the cross-section, but also on its autocorrelation and time series properties.

C0247: **Who is overindebted in the UK**

Presenter: **Jonathan Crook**, University of Edinburgh, United Kingdom

Successive waves of the UK's Wealth and Assets Survey are used to answer two questions: what determines whether a household and a person are over-indebted? Which households/people would be over-indebted if they were given credit? We consider different definitions of being overindebted and investigate the sensitivity of being overindebted to various household and personal characteristics, including levels and changes in employment status, income, state of health, degree of risk aversion and subjective discount rate. We do this by converting successive waves of the WAS into a panel structure at both household and personal levels. We estimate sample random effects and selection models. We analyse changes in the determinants of over-indebtedness over time.

CO126 Room BH (SE) 1.05 BAYESIAN TIME SERIES ANALYSIS Chair: Gary Koop

C1281: **A dynamic degree and strength corrected stochastic block model with infinite communities**
Presenter: **Ovielt Antonio Baltodano Lopez**, Ca' Foscari University, Italy
Co-authors: Roberto Casarin, Mauro Costantini

The high heterogeneity in real network data affects the performance of community detection methods from a modeling and computational perspective. We propose a dynamic stochastic block model with infinite communities that allows making inferences on the number of communities using Bayesian nonparametric techniques after controlling for degree and strength heterogeneity produced by observable and unobservable factors. We cope with the poor mixing of the number of communities by using an MCMC that combines the forward filtering backward sampling and the merge-split approaches within an adaptive framework. The application of this model to the effect of COVID on the global international trade network shows the complexity of trends experienced during the pandemic, and it identifies cases that resulted in negative consequences and others in an increase of trade opportunities.

C1326: **Dynamic identity-link latent space infinite-mixture: An application on DAX components**
Presenter: **Antonio Peruzzi**, Ca' Foscari University of Venice, Italy
Co-authors: Roberto Casarin

Finance literature suggests that cross-correlations among assets increase during periods of financial distress, and that cross-correlation's very own clustering structure varies over time. An Identity-Link Latent-Space Infinite-Mixture model with random-walk intercept is proposed to analyze the clustering structure of cross-correlation over time. The model allows for the representation of stocks on a d-dimensional Euclidean space and the clustering of assets into groups. Model estimation is carried out within a Bayesian framework, which allows including prior extra-sample information in the inference and accounting for parameter uncertainty. We apply the model to time-varying correlations among the DAX components. We find evidence of clustering effects and positive dependence between the number of clusters and both annualized volatility and average cross-correlation.

C1804: **Unusual weather in unusual economic times**
Presenter: **Leopold Ringwald**, International Institute for Applied System Analysis (IIASA), Austria
Co-authors: Florian Huber, Tamas Krisztin, M. Marcellino

The impacts of extreme weather anomalies on the US macroeconomy are studied by allowing for non-linearities and quantile-specific impacts. The Actuaries Climate Index identifies extreme outliers in weather patterns, which tracks changes in extreme temperatures, heavy rainfall, drought, high wind, and sea level. The approach allows us to estimate these impacts conditional on the *pth* quantile in the response. This way, we can distinguish the effect of weather shocks on different states of the macroeconomy. We use Bayesian Markov chain Monte Carlo (MCMC) methods for estimation and structural inference with our quantile factor VAR model (QF-BART). Our results indicate a significant negative response over time of weather shocks on industrial production after 1990. Moreover, the initial impact on the lowest and highest percentile differs in magnitude and sign.

CO316 Room BH (SE) 2.10 ADVANCES IN MACROECONOMETRIC MODELLING Chair: Anthoulla Phella

C0288: **Spike and slab priors on variable orderings in VARs**
Presenter: **Ping Wu**, University of Strathclyde, United Kingdom
Co-authors: Gary Koop

It is increasingly common to estimate Bayesian Vector Autoregressions (VARs) in a structural form involving the Cholesky decomposition of the reduced form error covariance matrix. The resulting structural form has an error covariance matrix which is diagonal, allowing for equation-by-equation estimation of the VAR, leading to a huge reduction in the computational burden. However, this leads to order dependence. Posterior and predictive results differ depending on the way the variables are ordered in the VAR. We propose the use of spike and slab priors over different variable orderings and allow the data to select the optimal ordering. We develop two models and Markov Chain Monte Carlo (MCMC) methods for posterior sampling over orderings based on the Plackett-Luce and Bradley-Terry models. In a macroeconomic exercise involving VARs with 20 variables, we demonstrate the effectiveness of our two approaches in choosing the optimal ordering and find substantive forecasting improvements relative to a strategy of subjectively selecting a single ordering.

C0408: **The time-varying evolution of inflation risks**
Presenter: **Anthoulla Phella**, University of Glasgow, United Kingdom
Co-authors: Dimitris Korobilis, Alberto Musso, Bettina Landau

A Bayesian quantile regression model with time-varying parameters (TVPs) is developed for forecasting inflation risks. The proposed parametric methodology bridges the empirically established benefits of TVP regressions for forecasting inflation with the ability of quantile regression to model the whole distribution of inflation flexibly. In order to make our approach accessible and empirically relevant for forecasting, we derive an efficient Gibbs sampler by transforming the state-space form of the TVP quantile regression into an equivalent high-dimensional regression form. An application of this methodology points to a good forecasting performance of quantile regressions with TVPs augmented with specific credit and money-based indicators for the prediction of the conditional distribution of inflation in the euro area, both in the short and longer run, and specifically for tail risks.

C0416: **Carbon tax impact on investments: Still a story of economic growth?**
Presenter: **Yiqiao Sun**, European Central Bank, Germany

The aim is to study the impact of carbon policies on investments across sectors in Europe and to contribute to the understanding of the propagation of climate policy impact through the European economy. Business investment is key in complementing public investment to enable the transition to carbon neutrality. Aggregate investment is often found to decline in response to the strengthening of climate policies, but at the same time, more stringent climate policies can incentivize investments in green innovation. As sectors are affected differently by carbon taxes, we focus on sector-level evidence, assuming that challenges to investments and investment opportunities are sector-specific. We investigate the dynamics of different investment components to infer about the growth potential of the sectors faced with the climate transition. We identify carbon shocks using local projection methods. To study the impact of carbon policies on investments, we combine carbon tax rates data with investment data from EU Klems capital accounts for broad sectors, and in particular, the investment in research and development, which indicates the sectors' potential for innovation over the medium term. Our preliminary findings point to a moderate but persistent rise in investments in response to a carbon tax shock, but large differences across sectors and investment components emerge in contrast to results obtained from a standard SVAR.

CC797 Room S-1.06 MACHINE LEARNING IN FINANCE Chair: Sandra Paterlini

C0446: **The impact of TCFD reporting: A new application of zero-shot analysis to climate-related financial disclosures**
Presenter: **Elena Toenjes**, Justus-Liebig-University Giessen, Germany
Co-authors: Alix Auzepy, Christoph Funk

Climate-related disclosures in 3,335 reports are examined based on a sample of 188 banks that officially endorsed the recommendations of the Task Force for Climate-related Financial Disclosures (TCFD). In doing so, we introduce a new application for zero-shot text classification based on the

BART model and an MNLI task. By developing a set of robust and fine-grained labels, we show that zero-shot analysis provides high accuracy in classifying companies' climate-related disclosures without further model training. Overall, our findings show that TCFD-supporting banks increase their level of disclosure after the launch of the TCFD recommendations and following their individual declaration of support. However, we also find significant variation in the extent of reporting by topic, suggesting that some recommendations have not yet been fully met. Our findings yield important conclusions for the design of climate-related disclosure frameworks.

C1292: **Machine learning applications to valuation of options on non-liquid markets**
Presenter: **Jiri Witzany**, University of Economics in Prague, Czech Republic
Co-authors: Milan Ficura

Recently, there has been considerable interest in machine learning (ML) applications for the valuation of options. However, it is usually assumed that there is a relatively liquid market with plain vanilla option quotations that can be used to calibrate (using an ML approach such as a neural network - NN) the volatility surface, or to estimate parameters of an advanced stochastic model. In the second stage, the calibrated volatility surface (or the model parameters) are used to value given exotic options, again using a trained NN (or another ML model). The two NNs are typically trained offline by sampling many model and market parameter combinations and calculating the options market values. We focus on the quite common situation of a non-liquid option market where we lack sufficiently many plain vanilla option quotations to calibrate the volatility surface, but we still need to value an exotic option or a just plain vanilla option subject to a more advanced stochastic model as it is typical on energy markets. We show that it is possible to use selected moments of the underlying historical price return series complemented with a volatility risk premium estimate to value such options using the ML approach.

C1755: **Machine learning techniques in joint default assessment**
Presenter: **Patrizia Semeraro**, Politecnico di Torino, Italy
Co-authors: Elisa Luciano, Margherita Doria

The aim is to study the consequences of capturing non-linear dependence among the covariates that drive the default of different obligors in the overall riskiness of their credit portfolio. Joint default modeling is, without loss of generality, the classical Bernoulli mixture model. Marginal and joint defaults depend on a set of covariates, common to all obligors. Linear and nonlinear dependence among covariates is captured by ML methods, while LR captures linear dependence only. We show through an application to credit card data that the ability of machine learning methods to capture nonlinear dependence among the covariates produces higher default correlation and, therefore, more conservative risk measures of the quantile type.

CC796 Room BH (SE) 1.01 VALUE-AT-RISK Chair: Vincenzo Candila

C0767: **A novel methodology to enriching the Archimedean family of copulas application to electricity peak demand estimation**
Presenter: **Moshe Kelner**, University of Haifa and Noga - Israel System Operator, Israel
Co-authors: Zinoviy Landsman, Udi Makov

A copula is an effective and elegant valuable tool for modeling dependence between random variables. Among the many families of this function, one of the most prominent is the Archimedean family, which has its unique structure and features. Most copula functions in this family have only a single dependence parameter, limiting the scope of the dependence structure. A modification of the Archimedean inverse generator is presented as a way to maintain membership in the family while increasing the number of dependence parameters. This is achieved by compounding the inverse generator with a density function of the dependence parameter. The method is demonstrated using the generalized gamma as a compounding density function of the dependence parameter of the bi-variate Clayton copula inverse generator. This enriches the Clayton copula from a single parameter to a three-parameter function and generates a new Archimedean family, the Clayton generalized Gamma (CGG), that comprises several members. In addition, the conditional VaR is established and used to obtain a confidence interval of one variable given the other. Using the CGG, we propose a probability model for electricity peak demand as a function of wet. Two new measures of fit, an economic measure and a conditional coverage measure based on the conditional VaR, were introduced to select the most appropriate family member based on empirical data on daily peak demand and minimum temperature in the winter.

C1711: **Forecast calibration, backtests, and loss decompositions for Value-at-Risk forecasts**
Presenter: **Marius Puke**, University of Hohenheim, Germany
Co-authors: Timo Dimitriadis

The evaluation of Value-at-Risk (VaR) forecasts is an own strand of literature rooted in the importance of banking and insurance regulation. Usually, one distinguishes between absolute and relative forecast evaluation, where the former refers to backtesting procedures and the latter to the use of loss functions. We make use of recent contributions to forecast calibration assessment and illustrate that absolute and relative forecast evaluation are similar tasks. To establish the connection between absolute and relative forecast evaluation, we revisit a decomposition of loss functions into measures of miscalibration (MCB), discrimination (DSC), and uncertainty (UNC) and show empirically that common backtests entirely overlook the (DSC) component. This corresponds to ignoring the forecasting models ability to discriminate between low and higher-risk periods. As a consequence, that might lead to inconclusive backtest results. Instead, the loss decomposition reveals this additional, hitherto unexploited information, provoking more informative insights.

C1840: **High-frequency volatility measurement and forecasting: Parametric vs non-parametric approaches**
Presenter: **Bjoern Schulte-Tillmann**, University of Munster, Germany
Co-authors: Mawuli Segnon, Timo Wiedemann

Recently, the plethora of existing (non-parametric) realized variance (RV) estimators was extended by a parametric estimator based on price durations that are modeled by an autoregressive conditional duration (ACD) specification. Encouraged by its superior performance, we study the impact of different specifications on estimation accuracy and forecasting performance. To this end, we introduce the factorial hidden Markov duration (FHMD) process to model the dynamics governing financial durations and derive its statistical properties. Utilizing high-frequency data of ten actively traded stocks, we consider further parametric price duration-based RV estimators and the most common non-parametric approaches. We find that the duration-based approaches achieve substantial accuracy gains in an application to forecasting (i) the integrated variance and (ii) the Value-at-Risk.

CC794 Room BH (SE) 1.06 MACROECONOMETRICS II Chair: Niklas Ahlgren

C1636: **Macroeconomics with a thick pen**
Presenter: **Marc Gronwald**, International Business School Suzhou, China

The co-movement of macroeconomic time series is measured using two Thick Pen methods: the Thick Pen Measure of Association (TPMA) as well as Multi-Thickness Thick Pen Measure of Association (MTTPMA). Analysed are, first, US long- and short-term interest rates, and, second, growth rates of per capita GDP and consumption. The key findings are: first, while the co-movement of the long-term features of the two interest rates is found to be very high, their short-term features are not only related to a much lower extent; in certain periods, these are even found to be out of sync. What is more, the degree of co-movement of the long-term features considerably declined in recent years. Second, the co-movement of both the long-term and short-term features of GDP and consumption growth rates is not only generally higher, but it also fluctuates considerably less over time.

C0529: **Understanding the fiscal price puzzle: Evidence from a nonlinear VAR approach**
Presenter: **Ngoc Trang Nguyen**, Monash University, Australia

Theoretical models predict that inflation increases in response to a positive government spending shock. However, many empirical studies provide evidence that a government spending shock is deflationary. Considering this gap, the purpose is to examine the impact of government spending shocks on prices in the US and how the effects differ at and away from the zero lower bound period using a parsimonious Interacted-VAR model. The decline in prices is persistent and significant in both states of the economy. During periods of active monetary policy, a fiscal spending innovation raises output on impact, whose effect dies down after about three years. Private consumption and output drop quickly in response to government spending shock and remain negative for most of the horizons before recovering slowly to pre-shock levels in presence of the zero constraint. Government spending multipliers under both regimes are high on impact and decrease to below one after one year. The multiplier in the normal state is more persistent and remains positive in the long run, while the multiplier in the ZLB state displays a quick drop and turns negative after two years, although the results indicate no statistical difference between the multipliers.

C0320: **Quantile local projections: Identification, smooth estimation, and inference**
Presenter: **Josef Ruzicka**, Nazarbayev University, Kazakhstan

Standard impulse response functions measure the average effect of a shock on a response variable. However, different parts of the distribution of the response variable may react to the shock differently. A popular method to capture this heterogeneity is quantile local projections. We show how to identify them by short-run restrictions, long-run restrictions or external instruments, and establish their asymptotics. To overcome their excessive volatility, we introduce two novel smoothing estimators. We propose information criteria for optimal smoothing and apply the estimators to shocks in financial conditions and monetary policy. We demonstrate that financial conditions affect the entire distribution of future GDP growth and not just its lower part.

CC768 Room BH (S) 2.03 TIME SERIES AND DYNAMIC MODELS Chair: Jose Olmo

C2025: **SEASCAPE.PAS: A program for seasonal adjustment**
Presenter: **Stephen Pollock**, University of Leicester, United Kingdom

The SEASACAPE program provides two alternative sets of procedures for effecting the seasonal adjustment of economic data. In the first set of procedures, which operate in the time domain, the seasonal adjustment filter is augmented by additional filters that are targeted at the adjacent frequencies. In the second set of procedures, a Fourier transform is deployed to reveal the elements of the data at all the frequencies. This allows the elements that are in the vicinities of the seasonal frequencies to be attenuated or eliminated at will. The SEASCAPE program has been coded in an object-orientated version of Pascal associated with the Delphi compiler. The program has been compiled by the Lazarus compiler, which is a freely available clone of Delphi. The program and its code are available online.

C1572: **Nowcasting with signature methods**
Presenter: **Giulia Mantoan**, The Alan Turing Institute, United Kingdom
Co-authors: Lingyi Yang, Aureo de Paula, Lars Nesheim, Samuel Cohen, Giulia Mantoan, Silvia Lui, Emma Small, Craig Scott, Will Malpass

Nowcasting refers to the "forecast" of the current ("now") state of the economy. This is necessary as key economic variables are often published with a significant delay of over a month. The nowcasting literature has arisen to address the need to have fast, reliable estimates of delayed economic indicators. The path signature is a mathematical object which captures geometric properties of sequential data; it naturally handles missing data from mixed frequency and/or irregular sampling - issues often encountered when merging multiple data sources - by embedding the observed data in continuous time. Calculating path signatures and using them as features in models have achieved state-of-the-art results in other fields such as finance, medicine, and cyber security. We look at the nowcasting problem by applying regression on signatures, a simple linear model on these nonlinear objects that we show subsumes the popular Kalman filter. We quantify the performance via a simulation exercise and application to US GDP growth, where in the latter, we compare performance with the dynamic factor model. By embedding discrete information in continuous time, this approach allows greater flexibility for future applications on data with complex sampling patterns.

C2030: **Determinants of market volatility: A latent threshold dynamic model**
Presenter: **Azam Shamsi Zamenjani**, University of New Brunswick, Canada
Co-authors: John Maheu

Measuring, modeling, and forecasting volatility are of great importance in financial applications such as asset pricing, portfolio management, and risk management. We investigate the predictability of stock market volatility by macro-finance variables in a dynamic regression framework using latent thresholding. The latent threshold models allow data-driven shrinkage of regression coefficients, by collapsing them to zero for irrelevant predictor variables and allowing for time-varying nonzero coefficients when supported by the data. This is a parsimonious framework that selects what potential predictor variables should be included in the regressions and when. We discuss the Bayesian model specification of dynamic regressions using latent thresholding. We incorporate a large number of potential predictors and let the data determine the relevant predictors over time. We applied the models to monthly S&P 500 volatility and nd that using macro-finance variables in volatility forecasts enhances model performance statically and economically, particularly when we allow for dynamic inclusion/exclusion of these variables.

CC802 Room BH (SE) 2.05 CRYPTOCURRENCY MARKETS Chair: Massimiliano Caporin

E0299: **Dynamic spillover in the cryptocurrency market**
Presenter: **Chih-Chiang Wu**, Yuzn Ze University, Taiwan
Co-authors: Wei-Peng Chen, Withz Aimable

Liquidity spillovers are analyzed among the eight largest cryptocurrencies. Empirical results show that Bitcoin is mainly a dominator in the liquidity connectedness in the cryptocurrency market. Furthermore, we show that the total liquidity spillover index is positively related to uncertainties in the US economies (i.e., VIX and EPU). However, it is negatively related to uncertainties of commodity markets (i.e., EVZ and GVZ) and the volatility of VIX (VVIX). These results suggest that the US economy uncertainties and the commodity market volatilities are critical determinants and play essential roles in cryptocurrency connectedness. Overall, these determinants are helpful for the decision-making of individual investors, financial institutions, and policymakers to understand the liquidity risk in the crypto market.

C1803: **Will Bitcoin ever become less volatile?**
Presenter: **Ladislav Kristoufek**, Czech Academy of Sciences, Czech Republic

The drivers of Bitcoin volatility are examined, and possible future developments are discussed, specifically what conditions need to be met for the volatility to decrease. Our instrumental variables analysis implies there needs to be a considerable inflow of small users into the system who are ideally not exchange traders and they do perform small transfers. Increasing exchange volume, on-chain transfers value, and Bitcoin price by themselves increase volatility of the cryptoasset.

C1802: **Fundamental and speculative components of the cryptocurrency pricing dynamics**
Presenter: **Jiri Kukacka**, Czech Academy of Sciences, Czech Republic
Co-authors: Ladislav Kristoufek

The driving forces behind cryptoassets price dynamics are often perceived as being dominated by speculative factors and inherent bubble-bust

episodes. The fundamental components are believed to have a weak, if any, role in the price formation process. Five cryptoassets with different backgrounds are studied, including Bitcoin, Ethereum, Litecoin, XRP, and Dogecoin, between 2016 and 2022. It utilizes the cusp catastrophe model to connect the fundamental and speculative drivers with possible price bifurcation characteristics of events of a market collapse. The findings show that all studied assets except Dogecoin demonstrate their price and returns dynamics emerge from complex interactions among both fundamental and speculative components, including episodes of price bifurcations. Bitcoin shows the strongest fundamentals, with the on-chain activity and economic factors driving the fundamental part of the dynamics. Investor attention and off-chain activity mainly drive the speculative component for all studied assets. Within the fundamental drivers, the analyzed cryptoassets present their coin-specific factors, which can be tracked to their protocol specifics and are economically sound.

CC804 Room BH (SE) 2.09 YIELD CURVE Chair: Pierangelo De Pace

C0482: **Similarity-based recession prediction in different interest rate environments**

Presenter: **Visa Kuntze**, University of Turku, Finland

Co-authors: Henri Nyberg, Samuel Rauhala

A flexible nonparametric similarity-based approach is developed to predict the state of the business cycle in different interest rate environments. Our approach provides methodological advantages over parametric logit and probit models and new empirical perspectives on the usefulness of the term spread as the main leading indicator and its connection to the prevailing interest rate level. Empirical results on the U.S., euro area and Japan show that the predictability of business cycle regimes depends on not just the term spread but also monetary policy conditions measured by the level of the short-term interest rate.

C1273: **A new term structure model for pricing bonds**

Presenter: **Ioannis Paraskevopoulos**, Universidad Pontificia Comillas, Spain

A novel stochastic term structure interest rate model is presented to describe the evolutions of level, slope and curvature of the yield curve. We use it to price all green and non-green bonds within an arbitrage-free dynamic term structure model and we link it with a previous yield curve while we extend certain specifications. We test our model using large bond data set, including green and nongreen, vanilla bonds. A stochastic Kalman Filter model will be employed to test the robustness of the calibration of this model to the market data.

C0641: **Yield curve estimation and liquidity risk in corporate bond market**

Presenter: **Takeshi Kobayashi**, NUCB Business School, Japan

The aim is to estimate the zero coupon yield curve of Japanese corporate bonds at firms' level. We compare the different types of models and identify their characteristics and propose a different type of weighting method considering liquidity risk in the Japanese corporate bond market. We examine the model performance by maturity bucket and time series. We show that the Steeley (B-spline) model fits best among the models. The results also show the yield discrepancy provides an improvement in pricing errors. It contributes to the literature by considering liquidity in the estimation of the corporate bond yield curve. For academics and practitioners, clear evidence of the practical importance of estimation methods and the basic data for credit risk modeling is provided.

Monday 19.12.2022 10:25 - 12:05 Parallel Session M – CFE-CMStatistics

EO522 Room S-1.04 STATISTICAL METHODS FOR DEPENDENCE Chair: Elisa Perrone

E1753: **Copula-based clustering of time series based on multivariate comonotonicity**
Presenter: **Fabrizio Durante**, University of Salento, Italy
Co-authors: Sebastian Fuchs, Roberta Pappada

In recent years, copula-based measures of association have been exploited to develop clustering methods that can take into account the dependence among different (one-dimensional) time series. In spatial statistics, such methods are particularly helpful in identifying hidden spatial patterns that define sub-regions characterized by a similar stochastic behaviour (i.e. regionalization). However, the majority of regionalization techniques focus on the spatial clustering of a single variable of interest, thus ignoring the role of compound events for extremes. Motivated by these problems, we propose a dissimilarity-based clustering procedure to group geographic sites characterized by multiple time series. In particular, the procedure tends to clustersites that exhibit a weak form of comonotonic behavior, which is more tailored for some applications. Different strategies to create such dissimilarity indices are hence illustrated and compared in a simulation study.

E0891: **Fast estimation of Kendall's tau and conditional Kendall's tau matrices under structural assumptions**
Presenter: **Alexis Derumigny**, Delft University of Technology, Netherlands
Co-authors: Rutger van der Spek

Kendall's tau and conditional Kendall's tau matrices are multivariate (conditional) dependence measures between the components of a random vector. For large dimensions, available estimators are computationally expensive and can be improved by averaging. Under structural assumptions on the underlying Kendall's tau and conditional Kendall's tau matrices, we introduce new estimators that have a significantly reduced computational cost while keeping a similar error level. In the unconditional setting, we assume that, up to reordering, the underlying Kendall's tau matrix is block-structured with constant values in each of the off-diagonal blocks. The estimators take advantage of this block structure by averaging over (part of) the pairwise estimates in each of the off-diagonal blocks. Conditional Kendall's tau matrix estimators are constructed similarly as in the unconditional case by averaging over (part of) the pairwise conditional Kendall's tau estimators. We establish their joint asymptotic normality, and show that the asymptotic variance is reduced compared to the naive estimators. Then, we perform a simulation study which displays the improved performance of both the unconditional and conditional estimators. Finally, the estimators are used for estimating the value at risk of a large stock portfolio; backtesting illustrates the obtained improvements compared to the previous estimators.

E1761: **Revisiting the Williamson transform in the context of multivariate Archimedean copulas**
Presenter: **Nicolas Dietrich**, Universitat Salzburg, Austria
Co-authors: Wolfgang Trutschnig, Thimo Kasper

A very recent result states that within the family of all d-dimensional Archimedean copulas, standard pointwise convergence implies $d-1$ weak conditional convergence (that is, weak convergence of almost all $(d-1)$-Markov kernels), and it is well-known from the literature that pointwise convergence within the family of multivariate Archimedean copulas is equivalent to the convergence of the corresponding normalized generators. We view multivariate Archimedean copulas via the Williamson transform, i.e., we study probability measures on $(0,\infty)$ whose corresponding Williamson transform coincides with the considered generators. Using this handy interrelation, it is not only possible to derive alternative handy formulas for the mass of level sets of the copulas but also to prove that both afore-mentioned notions of convergence may fully be characterized in terms of weak convergence of the probability measures on $(0,\infty)$. Furthermore, even singularity properties of the Archimedian copulas may directly be derived from the probability measures on $(0,\infty)$.

E1792: **Multivariate Bernoulli copulas: properties and limitations**
Presenter: **Elisa Perrone**, Eindhoven University of Technology, Netherlands
Co-authors: Roberto Fontana

Possible extensions to higher dimensions of the notion of Bernoulli copula are analyzed. We show that there are various ways to define Bernoulli copulas in dimension three (and higher), and that the choice is generally arbitrary. We investigate the pros and cons of these extensions and inquire into the uniqueness of the associated statistical models in relation to additional stochastic constraints, such as exchangeability.

EO304 Room S-1.06 MULTIVARIATE ANALYSIS OF COMPLEX DATA Chair: Thomas Verdebout

E1536: **Testing for auto-calibration**
Presenter: **Julien Trufin**, Universita Libre de Bruxelles, Belgium
Co-authors: Michel Denuit, Julie Huyghe, Julien Trufin, Thomas Verdebout

Dominance relations and diagnostic tools based on Lorenz and Concentration curves have been previously proposed to compare competing regression function estimators. This approach turns out to be equivalent to forecast dominance when the estimators under consideration are auto-calibrated. A new characterization of auto-calibration is established, based on the graphs of Lorenz and Concentration curves. This result is exploited to propose an effective testing procedure for auto-calibration. A simulation study is conducted to evaluate its performance, and its relevance for practice is demonstrated on a real data set.

E1606: **Inference based on measure transportation for directional data**
Presenter: **Thomas Verdebout**, Universite Libre de Bruxelles, Belgium
Co-authors: Marc Hallin, Hang Liu

A concept of directional distribution function, ranks and signs based on measure transportation are proposed. We provide some theoretical properties of the corresponding empirical versions and show how goodness-of-fit tests can be obtained. Our results are illustrated with Monte-Carlo simulations.

E1656: **Regression trees for extreme events, applications to natural disasters and cyber insurance pricing**
Presenter: **Olivier Lopez**, Sorbonne Universite, France

A regression tree procedure adapted to the analysis of extreme events is introduced. We show theoretical results assessing the performance of the procedure for finite sample size. We then show how this tool can be used to build priors for bayesian credibility pricing in insurance. Two applications are considered, one in a natural disaster, and the other in cyber insurance. In each case, the regression tree approach allows linking a claim to a risk class, in order to improve the information conveyed by historical data from the victim.

E1661: **Power enhancement for dimension detection of Gaussian signals**
Presenter: **Gaspard Bernard**, Universite Libre de Bruxelles (ULB), Belgium
Co-authors: Thomas Verdebout

The focus is on the classical problem of testing $\mathcal{H}_{0q}^{(n)}: \lambda_q^{(n)} > \lambda_{q+1}^{(n)} = \ldots = \lambda_p^{(n)}$, where $\lambda_1^{(n)},\ldots,\lambda_p^{(n)}$ are the ordered latent roots of covariance matrices $\Sigma^{(n)}$. We show that the usual Gaussian procedure $\phi^{(n)}$ for this problem essentially shows no power against alternatives of weaker signals of the form $\mathcal{H}_{1q}^{(n)}: \lambda_q^{(n)} = \lambda_{q+1}^{(n)} = \ldots = \lambda_p^{(n)}$. This is very problematic if the latter procedure is used to perform inference on the true dimension of

the signal. We show that the same test $\phi^{(n)}$ enjoys some local and asymptotic optimality properties to detect alternatives to the equality of the $p-q$ smallest roots of $\Sigma^{(n)}$ provided that $\lambda_q^{(n)}$ and $\lambda_{q+1}^{(n)}$ are such that $n^{1/2}(\lambda_q^{(n)}-\lambda_{q+1}^{(n)})$ diverges to ∞ as $n\to\infty$. We obtain tests $\phi_{\text{new}}^{(n)}$ for the problem that keeps the local and asymptotic optimality properties of $\phi^{(n)}$ when $n^{1/2}(\lambda_q^{(n)}-\lambda_{q+1}^{(n)})\to\infty$ and properly detect alternatives of the form $\mathcal{H}_{1q}^{(n)}$. Our results are illustrated via simulations and on a gene expression dataset from which we also discuss the problem of estimating the dimension of the signal.

EO048 Room S-1.27 INNOVATIONS IN LATENT VARIABLE MODELLING Chair: Giorgia Zaccaria

E0724: **Covariate-modulated rectangular latent Markov models with an unknown number of regime profiles**
Presenter: **Alfonso Russo**, University of Rome Tor Vergata, Italy
Co-authors: Alessio Farcomeni, Maria Grazia Pittau, Roberto Zelli

A multivariate latent Markov model with a number of latent states is derived that can possibly change at each time point. There are several applications in Macroeconomics, Microeconomics, Ecology or Epidemiology, where it might be desirable to allow for time-varying latent structures since new patterns might emerge or disappear over time, especially with multivariate outcomes. We give two main methodological contributions. First of all, we specify a completely general rectangular latent Markov model, where outcomes are a mix of continuous and categorical measurements and both the manifest and latent distributions are conditioned on covariates, which can be particularly useful to explain transitions across latent stages. Secondly, we derive an efficient transdimensional Markov Chain Monte Carlo sampler, to obtain the posterior distribution of all the parameters and for the sequence of a number of latent states. Bayesian inference is based on a Reversible Jump approach that is separately performed for each time occasion. In a simulation study, we show that our approach has a lower bias than competitors and that it can recover the true underlying sequence of latent states with high probability, even when covariates are omitted and independent of whether the true latent sequence varies with time or not. We conclude with an analysis of the well-being of 100 nations, as expressed by the dimensions of the Human Development Index, for six-time points spanning a period of 22 years.

E0811: **A finite mixture approach for biclustering bipartite networks**
Presenter: **Dalila Failli**, University of Florence, Italy
Co-authors: Maria Francesca Marino, Francesca Martella

Bipartite networks are particularly useful for representing relationships between disjoint sets of nodes, called sending and receiving nodes. The mixtures of latent trait analyzers are modified to achieve a twofold objective with regard to bipartite networks: i) performing joint clustering of sending and receiving nodes; ii) using a latent trait to model the dependence between receiving nodes. Therefore, the suggested model cannot only partition the data matrix into homogeneous blocks, as in the biclustering approach, but also capture the latent variability of network connections within each block, as in the latent trait framework. The proposal also admits the inclusion of nodal attributes on the latent layer of the model, in order to understand how these influence cluster formation. An EM algorithm with a variational approximation is proposed to estimate the model parameters. The performance of the model is evaluated through a simulation study with a different number of nodes and partitions.

E1306: **Extending latent class models for dealing with multilevel cross-classified data structures**
Presenter: **Silvia Columbu**, University of Cagliari, Italy
Co-authors: Nicola Piras, Jeroen Vermunt

Latent class models, and mixture models in general, are a well-established statistical approach for clustering. When the data have different levels of units, the mixture model must be formulated in order to accommodate the nesting of the data. One possibility is to extend the model by considering separate mixtures for each level of the structure. Such an extension considers a clustering of lower-level units by also allowing a clustering of higher-levels. The simplest multilevel structure is that of observations hierarchically nested within higher-level units; however, it is also interesting the situation in which the same unit belongs simultaneously to two or more groups, i.e. units are cross-classified. A version of finite mixtures is presented that handles the latter situation, with mixture components for first-level units, and mixture components for the multiple higher-levels. Maximum likelihood estimation through standard EM algorithms is not feasible; therefore, it is proposed the use of a stochastic version which includes a Gibbs sampler in the implementation. Numerical experiments with simulated data of different nature will be presented in order to show the performances of the estimation procedure and the interest in the approach for data clustering.

E1399: **Ultrametric models for dimensionality reduction**
Presenter: **Giorgia Zaccaria**, University of Milano-Bicocca, Italy
Co-authors: Carlo Cavicchia, Maurizio Vichi

Many relevant multidimensional phenomena are characterized by nested latent concepts having different levels of abstraction, from the most specific to the most general. They can be represented by a tree-shape structure by supposing hierarchical relationships among observed variables. In literature, several methodologies have been proposed to both model the hierarchical relationships among observed variables that reflect unobserved ones, and assess the existence of latent variables of "higher-order". Nonetheless, these methodologies are usually developed with sequential procedures that do not optimize a unique objective function, and/or a confirmatory approach. We propose a new class of parsimonious, simultaneous, exploratory models which are based on the ultrametricity notion for matrices. The latter was introduced in mathematics and became widespread in statistics in relation to distances in hierarchical clustering. However, the definition of an ultrametric matrix differs from that of an ultrametric distance matrix and has interesting properties that make it useful for studying hierarchical relationships among variables, as to be one-to-one associated with a hierarchy of latent concepts. The proposal aims at identifying a parsimonious hierarchy by firstly partitioning the variables into groups, each one associated with a latent concept, and then inspecting their relationships via an ultrametric structure.

EO402 Room K0.16 RECENT ADVANCEMENTS IN STATISTICAL NETWORK ANALYSIS Chair: Jonathan Stewart

E0517: **Spike-and-slab priors for dimension selection in static and dynamic network eigenmodels**
Presenter: **Joshua Loyal**, Florida State University, United States
Co-authors: Yuguo Chen

Latent space models (LSMs) are frequently used to model network data by embedding a network's nodes into a low-dimensional latent space; however, correctly choosing the dimension of this space remains a challenge. The contribution is two-fold. First, we propose a new Bayesian LSM for dynamic networks that not only fixes parameter identifiability issues that have previously impeded dimension selection but also models a larger class of dynamic networks than previous approaches. Second, we propose a Bayesian approach to dimension selection for static and dynamic LSMs based on an ordered spike-and-slab prior that provides improved dimension estimation and satisfies several appealing theoretical properties. In particular, we show that the static model's posterior concentrates on low-dimensional models near the truth. These models are accompanied by a novel parameter expansion scheme that allows for efficient MCMC estimation using a Metropolis-within-Gibbs sampler with Hamiltonian Monte Carlo proposals. We demonstrate our approach's versatility and consistent dimension selection on simulated networks. Lastly, we use the static and dynamic models to study a static protein interaction network and the global arms trades dynamics during the Cold War.

E1255: **Supervised centrality via sparse spatial autoregression**
Presenter: **Chenlei Leng**, University of Warwick, United Kingdom

The social opinions, behaviors and sentiments of the players in a social network are closely associated with their network positions. Identifying the influential players in a network is of importance as it helps to understand how ties are formed, how information is propagated, and in turn, can guide the dissemination of new information by focusing on important players. Motivated by a Weibo social network on 2021 Henan Floods, where response variables on each node are available, we propose a novel notion of supervised centrality to account for the fact that the centrality of a node is task-specific. To estimate the supervised centrality and identify important players, we develop a novel sparse spatial autoregression model by introducing individual heterogeneity to each user. To overcome the computational difficulties with fitting the model for large social networks, we further develop a forward-addition algorithm and show that it can consistently identify a superset of the influential nodes. We apply our model to analyze three responses in the Henan Floods data: the number of comments, reposts and likes, and obtain interesting results. A simulation study further corroborates the developed theory.

E1343: **Model selection for network data based on spectral information**
Presenter: **Jairo Pena**, Florida State University, United States
Co-authors: Jonathan Stewart

A methodology is presented for model selection in the context of modeling network data. Network data, often represented as a graph, consists of a set of pairwise observations between elements of a population of interests. The statistical network analysis literature has developed many different classes of network data models, with notable model classes including stochastic block models, latent node position models, and exponential families of random graph models. We develop a novel methodology that exploits the information in the Laplacian matrix's spectrum to provide a measure of goodness-of-fit of a defined set of network data models to the observed network. We explore the performance of our proposed methodology to popular classes of network data models through simulation studies, and demonstrate the utility in practice by applying our methodology to a collaboration network of network science researchers and the Sampson monk network.

E1406: **A Bayesian approach to space- and time-indexed Markov processes, with application to the Italian premier football league**
Presenter: **Michael Schweinberger**, Pennsylvania State University, United States
Co-authors: Guanyu Hu

Technological advances have paved the way for collecting a wealth of data on interactions among team players in football, baseball, and other team-based sports. The resulting data involve networks of interactions within and between opposing teams and are indexed by space and time. Such space- and time-indexed network data are vital to understanding and predicting the performance of teams, because a team's performance is more than the sum of the strengths of its players. We pursue a Bayesian approach to modeling entire games of opposing sport teams as space- and time-indexed continuous-time Markov processes. We present an application to data recorded during the 2020/2021 season of the Italian premier football league (Serie A), which includes some of the best-known teams in European football.

EO609 Room K0.18 MODERN CAUSAL METHODS FOR CLINICAL AND HEALTH POLICY RESEARCH Chair: Nandita Mitra

E0468: **Hierarchical Bayesian agent based models and promise for causal inference**
Presenter: **Samrachana Adhikari**, NYU School of Medicine, United States

While agent-based models are popular simulation models to assess hypothetical interventions and have been widely used in infectious disease modeling, HIV prevention studies and urban planning, among others, parameter inference and validation of such models remain a challenge. We will explore hidden Markov model representation of agent-based models, with a particular focus on infectious disease, that allows us to utilize Bayesian modeling and estimation tools such as particle filters for estimation and inference. Current methodological challenges and the potential of such a framework in assessing causality by simulating counterfactual will be discussed.

E0737: **Causal influence, causal effects, and path analysis in the presence of intermediate confounding**
Presenter: **Ivan Diaz**, NYU Langone Health, United States

Recent approaches to causal inference have focused on the identification and estimation of causal effects, defined as (properties of) the distribution of counterfactual outcomes under hypothetical actions that alter the nodes of a graphical model. We explore an alternative approach using the concept of causal influence, defined through operations that alter the information propagated through the edges of a directed acyclic graph. Causal influence may be more useful than causal effects in settings in which interventions on the causal agents are infeasible or of no substantive interest, for example, when considering gender, race, or genetics as a causal agent. Furthermore, the proposed "information transfer" interventions allow us to solve a long-standing problem in causal mediation analysis, namely the non-parametric identification of path-specific effects in the presence of treatment-induced mediator-outcome confounding. We propose efficient non-parametric estimators for a covariance version of the proposed causal influence measures, using data-adaptive regression coupled with semi-parametric efficiency theory to address model misspecification bias while retaining root-n-consistency and asymptotic normality. We illustrate the use of our methods in two examples using publicly available data.

E0856: **Estimation of heterogeneous policy-relevant causal effects under the difference-in-differences framework with spillover**
Presenter: **Gary Hettinger**, University of Pennsylvania, United States
Co-authors: Youjin Lee, Nandita Mitra

Public policy interventions are commonly evaluated using the difference-in-differences (DiD) approach. However, this approach does not directly account for the effect of the policy spilling over to neighboring regions such as nearby cities. For example, the implementation of an excise tax on sweetened beverages in Philadelphia was shown to be associated with a substantial decrease in volume sales of taxed beverages in Philadelphia but also showed an increase in beverage sales in bordering counties which were not subject to the tax. The latter association could potentially be explained by cross-border shopping behaviors of Philadelphia residents. Because spillover effects can offset the total effect of such interventions, particularly for specific sub-populations, understanding the dynamics of such effects is essential to holistically evaluate public policies. To address these concerns, we extend difference-in-differences methods to identify the causal effects of policy interventions under various spillover conditions. We propose doubly robust estimators for the average treatment effect on the treated and on the neighboring control that relax standard assumptions on interference and model specification. We apply these methods to estimate the causal effects of the Philadelphia beverage tax. Additionally, we use our methods to analyze the heterogeneity of such effects across spatial and demographic characteristics of Philadelphia and its bordering counties.

E1103: **Semi-parametric g-computation to understand the effect of antiretroviral therapy on subsequent weight gain**
Presenter: **Andrew Spieker**, Vanderbilt University Medical Center, United States

G-computation is a longitudinal generalization of standardization suitable for settings in which there is time-dependent confounding. While highly useful as a tool for estimating longitudinal causal effects, its reliance on parametric models is sometimes criticized. We discuss the utility of cumulative probability models for use in g-computation as a way to relax parametric assumptions. Simulations suggest this approach to be robust and feasible to implement in the real world. We illustrate the utility of this methodology through a study of core and ancillary agents comprising antiretroviral therapies and their effects on weight gain in a large cohort of persons living with HIV. Specifically, we hypothesize that modern integrase strand transfer inhibitors and tenofovir alafenamide are associated with greater mean weight gain as compared to other core and ancillary agents.

EO152 Room K0.19 DIGITAL HEALTH AND INDIVIDUALIZED TREATMENT REGIMENS. Chair: Ashkan Ertefaie

E1325: **Improving the efficiency of time-varying causal effect moderation analysis in mobile health**
Presenter: **Walter Dempsey**, University of Michigan, United States

Twin revolutions in wearable technologies and smartphone-delivered digital health interventions have significantly expanded the accessibility and uptake of mobile health (mHealth) interventions. Sequentially randomized experiments called micro-randomized trials (MRTs) have grown in popularity as a means to empirically evaluate the effectiveness of mHealth intervention components. MRTs have motivated a new class of causal estimands, termed causal excursion effects. We revisit the estimation of causal excursion effects and present two new tools for improving efficiency. First, we will present a method to improve efficiency by including auxiliary variables. This method extends the covariate-adjustment RCT literature to the time-varying setting. Second, we will consider a meta-learner perspective, where any supervised learning algorithm can be used to assist in the estimation of the causal excursion effect. Theoretical comparisons accompanied by extensive simulation experiments demonstrate the relative efficiency gains. The practical utility of the proposed methods is demonstrated by analyzing data from a multi-institution cohort of first-year medical residents in the United States.

E1405: **Analyzing Event-triggered Adaptive Interventions using Data from Sequentially Randomized Trials**
Presenter: **Mason Ferlic**, University of Michigan, United States

A dynamic treatment regime consists of a protocolized sequence of decision rules used to guide an intervention across multiple stages of treatments contingent on the evolving status of the individual. Technological advances in mobile and digital health have made it possible to monitor dynamic treatment response in near real-time and adapt future treatment to individual needs. Technology-assisted adaptive interventions with a digital tailoring variable are becoming more commonplace. In such mobile monitoring environments, the set of decision rules is also allowed to vary with time, enabling researchers to answer more complex questions regarding the time-varying effect of treatment and the dynamic trajectory of response status. We introduce a new approach to analyzing technology-assisted adaptive interventions embedded in a sequential, multiple-assignment, randomized trial (SMART) on a continuous, longitudinal outcome. We propose a simple two-stage regression algorithm that adjusts for time-varying transitions to second-stage treatment. Through simulation studies, we illustrate the validity of estimated treatment effects and examine operating characteristics under different levels of model misspecification. We show that unadjusted standard errors are anti-conservative. Using data from a SMART, we illustrate our methodology in a case study involving digitally monitored weight loss treatment.

E1477: **Learning robust treatment regimes with imperfect data**
Presenter: **Tao Shen**, National University of Singapore, China
Co-authors: Yifan Cui

Robust methods have recently gained popularity for dynamic treatment regime estimation. Building on this, new methods are proposed which can be used to learn dynamic treatment regimes in a confounded setting with imperfect data. In our experiments, we find our approach to perform well relative to a number of existing methods.

E1473: **Optimal dynamic treatment regimes and partial welfare ordering**
Presenter: **Sukjin Han**, University of Bristol, United Kingdom

Dynamic treatment regimes are treatment allocations tailored to heterogeneous individuals (e.g., via previous outcomes and covariates). The optimal dynamic treatment regime is a regime that maximizes counterfactual welfare. We introduce a framework in which we can partially learn the optimal dynamic regime from observational data, relaxing the sequential randomization assumption commonly employed in the literature but instead using (binary) instrumental variables. We propose the notion of sharp partial ordering of counterfactual welfares with respect to dynamic regimes and establish a mapping from data to partial ordering via a set of linear programs. We then characterize the identified set of the optimal regime as the set of maximal elements associated with the partial ordering. We relate the notion of partial ordering with a more conventional notion of partial identification using topological sorts. Practically, topological sorts can be served as a policy benchmark for a policymaker. We apply our method to understand returns to schooling and post-school training as a sequence of treatments by combining data from multiple sources. The framework can be used beyond the current context, e.g., in establishing rankings of multiple treatments or policies across different counterfactual scenarios.

EO640 Room K0.20 SCALABLE INFERENCE METHODS FOR COMPLEX PROBLEMS Chair: Daniel Paulin

E1868: **Unbiased estimation for discretized models and its extension to underdamped langevin dynamics**
Presenter: **Ajay Jasra**, KAUST, Saudi Arabia

The focus is on computing expectations w.r.t. probability measures which are subject to discretization error. Examples include partially observed diffusion processes or inverse problems, where one may have to discretize time and/or space, in order to work with the probability of interest practically. Given access only to these discretizations, we consider constructing the construction of unbiased Monte Carlo estimators of expectations w.r.t. such target probability distributions. It is shown how to obtain such estimators using a novel adaptation of randomization schemes and Markov simulation methods. Under appropriate assumptions, these estimators possess finite variance and finite expected cost. This approach has two important consequences:(i) unbiased inference is achieved at the canonical complexity rate, and (ii) the resulting estimators can be generated independently, thereby allowing strong scaling to arbitrarily many parallel processors. Several algorithms are presented and applied to Bayesian inverse problems. We also show how this framework can be extended to unbiased MCMC associated with underdamped Langevin dynamics.

E1873: **Speeding up inference for high dimensional time series models**
Presenter: **Daniel Paulin**, University of Edinburgh, United Kingdom

Techniques for fitting multivariate VARMA models will be overviewed. Despite the extensive work in this area, the computation cost of existing methods becomes very high for multivariate time series with more than ten variables. To address this challenge, we propose a new technique that offers improvements both in computational efficiency and predictive accuracy.

E1928: **On the infinite depth limit of finite width neural networks**
Presenter: **Soufiane Hayou**, National University of Singapore, Singapore

The infinite depth limit of finite-width residual neural networks is discussed. The infinite-width limit of deep neural networks has been extensively studied. The converse (infinite depth) remains, however, poorly understood. With proper scaling, we show that by fixing the width and taking the depth to infinity, the vector of pre-activations converges in distribution to a zero-drift diffusion process that is essentially governed by the activation function. Unlike the infinite-width limit where the neurons exhibit a Gaussian behavior, we show that the infinite-depth limit (with finite width) yields different distributions depending on the choice of the activation function. We further discuss the sequential limit infinite-depth-then-infinite-width and show some key differences with the converse infinite-width-then-infinite-depth limit.

E1931: **Multilevel Bayesian deep neural networks**
Presenter: **Neil Chada**, King Abdullah University of Science and Technology, Saudi Arabia
Co-authors: Ajay Jasra, Kody Law, Sumeetpal Singh

The application of multilevel Monte Carlo for Bayesian computation tasks in machine learning is considered. There has recently been a synergy of statistics and machine learning, promoting the application and development of new methodologies. Based on this, we promote the use of

multilevel Monte Carlo, which is a technique used to reduce the cost to attain a particular order of MSE with trace-class neural network priors. We provide some theoretical insights, and demonstrate the performance of our methodology on different model problems, such as classification and reinforcement learning.

EO456 Room K0.50 HIGH-DIMENSIONAL LEARNING INFERENCE FOR DATA SCIENCE Chair: Rajen D Shah

E0748: **FACT: High-dimensional random forests inference**
Presenter: **Chien-Ming Chi**, Academia Sinica, Taiwan
Co-authors: Yingying Fan, Jinchi Lv

Random forests are one of the most widely used machine learning methods over the past decade. Yet, because of its black-box nature, the results by random forests can be hard to interpret in many big data applications. Quantifying the usefulness of individual features in random forests learning can greatly enhance its interpretability. Existing studies have shown that some popular feature importance measures for random forests suffer from the bias issue. In addition, there is a lack of comprehensive size and power analyses for most of these existing methods. We approach the problem via hypothesis testing, and suggest a framework of the self-normalized feature-residual correlation test (FACT) for evaluating the significance of a given feature in the random forests model with bias-resistance property, where our null hypothesis concerns whether the feature is conditionally independent of the response given all other features. The vanilla version of our FACT test can suffer from the bias issue in the presence of feature dependency. We exploit the techniques of imbalancing and conditioning for bias correction, and further incorporate the ensemble idea into the FACT statistic through feature transformations for enhanced power. Under a general high-dimensional nonparametric model setting with dependent features, we formally establish that FACT can provide theoretically justified random forests feature p-values and enjoy appealing power through nonasymptotic analyses.

E0980: **Tracy-Widom law of ridge-regularized F-matrix and applications**
Presenter: **Haoran Li**, Auburn University, United States

In Multivariate Data Analysis, many central problems can be formulated as a double Wishart problem where two Wishart matrices, W_1 and W_2, are involved. Important cases include MANOVA, CCA, and tests for linear hypotheses in multivariate linear regression. The traditional Roy's largest root test relies on the largest eigenvalue of the F-matrix $F = W_1 W_2^{-1}$. In a high-dimensional setting, the test is infeasible due to the singularity of W_2. To fix the singularity, we propose a ridge-regularized test where a ridge term is added to W_2. We derive the asymptotic Tracy-Widom distribution of the largest eigenvalue of the regularized F-matrix. Efficient methods for estimating the asymptotic mean and variance are designed through the Marchenko-Pastur equation. The power characteristics are studied under a class of local alternatives. A simulation study is carried out to examine the numerical performance of the proposed tests.

E1653: **High-dimensional asymptotics for single-index models via approximate message passing**
Presenter: **Yoshimasa Uematsu**, Hitotsubashi University, Japan
Co-authors: Kazuma Sawaya, Masaaki Imaizumi

The purpose is to investigate the precise asymptotic behavior of some estimators for single-index models with unknown links in a high-dimensional setting. Unlike the conventional non-asymptotic scheme, our setting allows a non-sparse coefficient vector while the dimension diverges proportionally with the sample size. Extending the generalized approximate message passing (GAMP) framework, we first uncover the bias in the asymptotic distribution caused by the non-sparse high-dimensionality and then propose a bias-corrected estimator. Finally, numerical experiments confirm the validity of our method.

E1659: **Linear discriminant analysis with label noise**
Presenter: **Timothy Cannings**, University of Edinburgh, United Kingdom

The effect of label noise in linear discriminant analysis is investigated. The main goal is to derive the minimax rate of convergence in this problem, where our results capture the explicit dependence of the minimax rate on all of the key parameters in the model. Our theory reveals a delicate interplay between the level of separation between the class conditional distributions (as measured by the Mahalanobis distance), the proportion of the training data points that are labelled incorrectly, as well as the dimension of the feature space and training sample size. Somewhat surprisingly, applying the vanilla approach to linear discriminant analysis is suboptimal in several regimes under label noise. We, therefore, introduce a new approach to classification in this setting, based on higher-order moment estimators and the midhinge estimator, which is rate optimal in all regimes.

EO663 Room S0.03 TIME SERIES AND SPATIAL STATISTICS: METHODOLOGY AND APPLICATIONS Chair: Adam Sykulski

E0633: **Efficient and accurate estimation from dependent data: The debiased spatial Whittle likelihood**
Presenter: **Sofia Olhede**, EPFL, Switzerland
Co-authors: Adam Sykulski, Arthur Guillaumin, Frederik Simons

A computationally and statistically efficient method is proposed for estimating the parameters of a stochastic covariance model observed on a regular spatial grid in any number of dimensions. Our proposed method, which we call the Debiased Spatial Whittle likelihood, makes important corrections to the well-known Whittle likelihood to account for large sources of bias caused by boundary effects and aliasing. We generalize the approach to flexibly allow for significant volumes of missing data, including those with lower-dimensional substructure, and for irregular sampling boundaries. We build a theoretical framework under relatively weak assumptions, which ensures consistency and asymptotic normality in numerous practical settings, including missing data and non-Gaussian processes. We also extend our consistency results to multivariate processes. We provide detailed implementation guidelines which ensure the estimation procedure can be conducted in O(n log n) operations, where n is the number of points of the encapsulating rectangular grid, thus keeping the computational scalability of Fourier and Whittle-based methods for large data sets. We validate our procedure over a range of simulated and real-world settings, and compare it with state-of-the-art alternatives, demonstrating the enduring practical appeal of Fourier-based methods, provided they are corrected by the developed procedures.

E0999: **Comparing populations of high-dimensional spectra**
Presenter: **Robert Krafty**, Emory University, United States
Co-authors: Marie Tuft, Fabio Ferrarelli, Ori Rosen, Zeda Li

Technological advances have led to an increase in the breadth and number of studies that collect high-dimensional time series signals, such as EEG, from multiple groups and whose scientific goal is to understand differences in time series spectra between the groups. Although methods have been proposed for comparing populations of power spectra that are univariate functions of frequency, often referred to as analysis of power (ANOPOW), none exist when time series are high-dimensional and spectra are complex Hermitian matrix-valued functions. We discuss a nonparametric Bayesian approach for ANOPOW with high-dimensional time series. The method models the collection of time series through a novel functional mixed effects factor model that can capture spectral differences between groups while accounting for within-group spectral variability. The approach is motivated by and used to analyze resting-state high-dimensional EEG in patients hospitalized for a first psychotic episode to understand how their electrophysiology differs from that of healthy controls.

E1058: **Wavelet spectra for multivariate point processes**
Presenter: **Ed Cohen**, Imperial College London, United Kingdom

Co-authors: Alex Gibberd

Humans are harvesting vast event datasets that manifest themselves as a list of times at which particular events of interest occur. Often these are multivariate in nature, with events being of different types or arriving on multiple channels. A key question is to what extent the data-generating point processes are correlated and to track non-stationary correlation structure. Wavelets provide the flexibility to analyse stochastic processes at different scales in a time-localised manner and have had a profound impact in statistics, particularly in time series analysis. We apply them to multivariate point processes as a means of detecting and analysing unknown non-stationarity, both within and across component processes. To provide statistical tractability, a temporally smoothed wavelet periodogram is developed and distributional results are extended to wavelet coherence; a time-scale measure of inter-process correlation. This statistical framework is further used to construct a test for stationarity in multivariate point-processes. The methodology is applied to neural spike train data, where it is shown to detect and characterise time-varying dependency patterns.

E1065: **Modeling and detecting changes in spatio-temporal processes**
Presenter: **Gaurav Agarwal**, Lancaster University, United Kingdom
Co-authors: Idris Eckley, Paul Fearnhead

Changepoints have been extensively studied for time series data, but there is limited literature on detecting changes in spatial processes over time. A likelihood-based methodology is developed for the simultaneous estimation of both changepoints and model parameters of spatio-temporal processes. Contrasting to existing spatial changepoint methods, which fit a piecewise stationary model assuming independence across segments, we fit a nonstationary model without any independence assumption. To deal with the complexity of the full likelihood model, we propose a computationally efficient Markov approximation. We study the effect of such an approximation and compare our method with existing methodologies through a comprehensive set of simulation studies. The method is employed for changepoint detection and missing data prediction in daily wind speeds across different synoptic weather stations in Ireland over a period of two years.

EO448 Room S0.11 STATISTICS AND COMPUTING FOR STOCHASTIC PROCESSES Chair: Kengo Kamatani

E0566: **Model comparison for ergodic SDEs in YUIMA**
Presenter: **Shoichi Eguchi**, Osaka Institute of Technology, Japan

There are several studies of model selection for stochastic differential equations (SDEs), which include the contrast-based information criterion for ergodic diffusion processes, AIC-type information criterion for ergodic Levy-driven SDEs, and the Schwarz-type information criterion for locally asymptotically quadratic models. Based on these studies, in R package yuima, the function for model selection for diffusion processes has been implemented. However, this function is not compatible with the model selection for Levy-driven SDEs. We will overview the model selection methods for Levy-driven SDEs and explain the improvements of the model selection function.

E0784: **Hypothesis testing for a parabolic linear SPDE with a small perturbation**
Presenter: **Yusuke Kaino**, Kobe University, Japan
Co-authors: Masayuki Uchida

A statistical hypothesis test is considered for a parabolic linear second-order stochastic partial differential equation (SPDE) in one space dimension with a small perturbation from high-frequency data which are observed in time and space. We aim to test whether a parameter of the SPDE model is zero. We propose three kinds of test statistics based on the high-frequency data: likelihood ratio type test statistic, Wald type test statistic and Rao's score type test statistic. It is shown that under some regularity conditions, these test statistics converge in distribution to a chi-squared random variable with one degree of freedom under the null hypothesis, and the tests are consistent. Moreover, we give some simulation results of the test statistics and examine the asymptotic behavior of the test statistics under the null and alternative hypotheses.

E0910: **Estimation of diffusion processes by online gradient descent**
Presenter: **Shogo Nakakita**, The University of Tokyo, Japan

An online parametric estimation method is proposed for discretely observed diffusion processes via an online gradient descent algorithm. Online estimation is a classical topic in time series analysis; however, few studies discuss it in parametric estimation of stochastic differential equations under general settings. The aim is to estimate parameters in an online manner by constructing convex quasi-log-likelihood functions and optimise them via online gradient descent, whose computational complexity for each refreshment is linear with respect to the dimension of the parameter of interest. It is shown that the proposed estimator has non-asymptotic uniform risk bounds with respect to the class of stochastic differential equations, even when the model is misspecified. Our result is based on three results: convergence guarantee for stochastic mirror descent methods with bias and dependence; simultaneous exponential ergodicity of multidimensional diffusion processes; and proposal of loss functions and their good approximations for parametric estimation.

E1215: **Information criterion for jump diffusion models**
Presenter: **Yuma Uehara**, Kansai University, Japan

A model selection problem is considered for jump-diffusion models based on high-frequency samples. The terminal time is supposed to diverge (ergodic setting), and our interest is to select drift and diffusion coefficients and jump distribution among candidates. Unlike the diffusion case, we cannot directly use the stochastic flow approach in order to evaluate the transition density when the jump term is parametrized in some way. Hence we introduce the approximated transition density function, which corresponds to the case where at most, one jump occurs within one observation interval. To validate such an approximation, we present new transition density estimates. From the estimates, we propose an explicit AIC-type information criterion constructed by the threshold-based quasi-likelihood function.

EO577 Room S0.12 BERNSTEIN-VON MISES THEOREM: RECENT RESULTS Chair: Natalia Bochkina

E0243: **Bayesian fixed-domain asymptotics for covariance parameters in a Gaussian process model**
Presenter: **Cheng Li**, National University of Singapore, Singapore

Gaussian process models typically contain finite dimensional parameters in the covariance function that need to be estimated from the data. We study the Bayesian fixed-domain asymptotics for the covariance parameters in a universal kriging model with an isotropic Matern covariance function, which has many applications in spatial statistics. We show that when the dimension of the domain is less than or equal to three, the joint posterior distribution of the microergodic parameter and the range parameter can be factored independently into the product of their marginal posteriors under fixed-domain asymptotics. The posterior of the microergodic parameter is asymptotically close in total variation distance to a normal distribution with shrinking variance, while the posterior distribution of the range parameter does not converge to any point mass distribution in general. Our theory allows unbounded prior support for the range parameter and flexible designs of sampling points. We further study the asymptotic efficiency and convergence rates in posterior prediction for the Bayesian kriging predictor with covariance parameters randomly drawn from their posterior distribution. In the special case of the one-dimensional Ornstein-Uhlenbeck process, we derive explicitly the limiting posterior of the range parameter and the posterior convergence rate for asymptotic efficiency in posterior prediction. We verify these asymptotic results in numerical experiments.

E1099: **Nonparametric Bernstein–von Mises theorems for discretely observed compound Poisson processes**
Presenter: **Jakob Soehl**, Delft University of Technology, Netherlands

Co-authors: Richard Nickl

Nonparametric Bayesian statistical inference is studied for the parameters governing a pure jump process of the form $Y_t = \sum_{k=1}^{N(t)} Z_k, \quad t \geq 0$, where $N(t)$ is a standard Poisson process of intensity λ, and Z_k are drawn i.i.d. from jump measure μ. A high-dimensional wavelet series prior for the Lévy measure $\nu = \lambda\mu$ is devised and the posterior distribution arises from observing discrete samples $Y_\Delta, Y_{2\Delta}, \ldots, Y_{n\Delta}$ at fixed observation distance Δ, giving rise to a nonlinear inverse inference problem. We derive contraction rates in uniform norm for the posterior distribution around the true Lévy density that are optimal up to logarithmic factors over Hölder classes, as sample size n increases. We prove a functional Bernstein–von Mises theorem for the distribution functions of both μ and ν, as well as for the intensity λ, establishing the fact that the posterior distribution is approximated by an infinite-dimensional Gaussian measure whose covariance structure is shown to attain the information lower bound for this inverse problem. As a consequence, posterior-based inferences, such as nonparametric credible sets, are asymptotically valid and optimal from a frequentist point of view.

E1409: **Semi-parametric Bernstein-von Mises theorem for linear models with one-sided error**

Presenter: **Natalia Bochkina**, University of Edinburgh, United Kingdom

Co-authors: Judith Rousseau, Jean-Bernard Salomond, Johan van der Molen Moris

The problem of linear regression with one-sided errors whose density is unknown is considered from a Bayesian perspective. We state general sufficient conditions for the local concentration of the marginal posterior of the parameters in the linear regression model (Bernstein - von Mises type theorem), which have a faster $1/n$ contraction rate and a constrained multivariate exponential distribution with random constraint, under an adaptive estimation of the unknown density. Consistent estimation of the unknown density of errors at zero is important, as this value is the scale parameter of the limiting constrained exponential distribution. In particular, to ensure that the error density is asymptotically consistent pointwise in a neighbourhood of zero, instead of usual Dirichlet mixture weights, we consider a non-homogeneous Completely Random Measure mixture. We illustrate the performance of this approach on simulated data, and apply it to model the distribution of bids in procurement auctions.

EO540 Room S0.13 STATISTICAL SUMMITS: METHODOLOGY AND COMPUTING II Chair: JT Ferreira

E0413: **Small area estimation with partially linear mixed measurement error models**

Presenter: **Mohammad Arashi**, Ferdowsi University of Mashhad, Iran

Co-authors: Elahe Hosseini, Davood Shahsavani, Mohammad Reza Rabiei

In small area estimation, the use of direct conventional methods will not lead to reliable estimates because the sample size is small compared to the population. Fay-Herriot model is commonly used in small area estimation in which borrowing strength from the related sites and other sources and uses auxiliary information to improve estimation. However, the assumption of normality is a limiting assumption for heavy-tailed data and outlying observations. Also, it is usually assumed that the predictors are measured without errors, which can be easily violated in small-area estimation. We provide a more flexible model beyond these limitations, which is more accurate than the existing models. Specifically, we study small area estimation in the partially linear mixed-effects model where measurement error is present for the predictors. We consider a large class of distributions for the error disturbances. Numerical studies are carried out to illustrate the superior performance of the proposed model in the prediction accuracy sense.

E0723: **Asymmetric Laplace scale mixtures**

Presenter: **Luca Bagnato**, Catholic University of the Sacred Heart, Italy

Co-authors: Antonio Punzo

Recent studies about cryptocurrency returns show that its distribution can be highly-peaked, skewed, and heavy-tailed, with a large excess kurtosis. To accommodate all these peculiarities, we propose the asymmetric Laplace scale mixture (ALSM) family of distributions. Each member of the family is obtained by dividing the scale parameter of the conditional asymmetric Laplace (AL) distribution by a convenient mixing random variable taking values on all or part of the positive real line and whose distribution depends on a parameter vector m providing greater flexibility to the resulting ALSM. Advantageously with respect to the AL distribution, the members of our family allow for a wider range of values for skewness and kurtosis. For illustrative purposes, we consider different mixing distributions; they give rise to ALSMs having a closed-form probability density function where the AL distribution is obtained as a special case under a convenient choice of m. We examine some properties of our ALSMs, such as hierarchical and stochastic representations and moments of practical interest. We describe an EM algorithm to obtain maximum likelihood estimates of the parameters for all the considered ALSMs. We fit these models to the returns of two cryptocurrencies, considering several classical distributions for comparison. The analysis shows how our models represent a valid alternative to the considered competitors in terms of AIC, BIC and likelihood-ratio tests.

E0931: **A personal journey into mixture modelling**

Presenter: **Sollie Millard**, University of Pretoria, South Africa

The focus is on a personal journey into mixture modelling. We start by considering the initial industry-based problems that ultimately resulted in my research endeavors into mixture modeling. We continue by highlighting my involvement in selected research papers, conference presentations and post-graduate student projects that address some of the challenges faced in model-based clustering and classification applications. Both mixture of distributions and mixture of regressions are considered. To conclude, an overview of current research projects focusing on non-parametric mixture modeling and feature selection is given. Finally, some thoughts on future research are presented.

E0370: **A journey of the Dirichlet distribution in the analysis of compositional data sets**

Presenter: **Seitebaleng Makgai**, University of Pretoria, South Africa

Data sets that consist of proportions (and thus subject to unit-sum constraints) are known as compositional data sets. These types of data sets naturally arise in a variety of disciplines, such as the medical sciences, biology, as well as in psychology. The Dirichlet distribution is a well-known candidate in modelling compositional data sets. However, in the presence of some extreme points or outliers, the Dirichlet distribution fails to model such data sets, making other model extensions necessary. As a solution to this shortfall, a technique called the beta-generating technique is applied in developing Dirichlet-type distributions that present greater flexibility in modelling various compositional data sets in the medical and biological sciences. These developments result in the proposal of the Dirichlet-Gamma distribution. As part of the study, the performance of the Dirichlet-Gamma distribution is investigated in a Bayesian context. The usefulness of this model is demonstrated through the application of a real data set in the medical and biological sciences.

EO118 Room Safra Lecture Theatre NOVEL PERSPECTIVES IN BAYESIAN STATISTICS Chair: Pier Giovanni Bissiri

E1483: **A novel approach to (strong) posterior contraction rates via Wasserstein dynamics**

Presenter: **Emanuele Dolera**, University of Pavia, Italy

Posterior contractions rates (PCRs) for Bayesian consistency are considered. The statistical model is a family $M = f_{t\,t\in T}$ of densities with T included in a separable Hilbert space, possibly infinite-dimensional. Two main approaches have been developed: the former considers neighborhoods of the true density f_{t_0} in the space of densities; the latter neighborhoods of the true value of the parameter t_0 in T. We follow the latter when the Hilbertian metric on T is stronger than the one induced by restricting to M usual metrics on densities (L_p, Hellinger, Kullback-Leibler, chi-square). We present two main statements, valid when: 1) we dispose of a Banach space-valued sufficient statistics which is classically consistent; 2) only the empirical

distribution is at disposal as sufficient statistics. Critical to our approach is an assumption of Lipschitz-continuity for the posterior with respect to observed data, ensuing from the dynamic formulation of the Wasserstein distance. This sets forth a connection between PCRs and other problems: Laplace methods for integrals, Sanovs principle, rates of mean Glivenko-Cantelli theorems, and estimates of weighted Poincar-Wirtinger constants. As a complement, we present novel improvements in evaluating both Laplace integrals and Poincar-Wirtinger constants in infinite dimensions. Finally, we illustrate our method by explicitly evaluating the PCRs for logistic-Gaussian model and linear regression.

E1441: **Statistical modeling within the generalized Bayes paradigm**
Presenter: **Tommaso Rigon**, University of Milano-Bicocca, Italy
Co-authors: Amy Herring, David Dunson

Loss-based clustering methods, such as $k-$means and its variants, are standard tools for finding groups in data. However, the lack of quantification of uncertainty in the estimated clusters is a disadvantage. Model-based clustering based on mixture models provides an alternative, but such methods face computational problems and large sensitivity to the choice of kernel. A generalized Bayes framework is proposed that bridges these paradigms through the use of Gibbs posteriors. In conducting Bayesian updating, the log-likelihood is replaced by a loss function for clustering, leading to a rich family of clustering methods. The Gibbs posterior represents a coherent updating of Bayesian beliefs without needing to specify a likelihood for the data, and can be used for characterizing uncertainty in clustering. We consider losses based on Bregman divergence and pairwise similarities, and develop efficient deterministic algorithms for point estimation along with sampling algorithms for uncertainty quantification. Several existing clustering algorithms, including k-means, can be interpreted as generalized Bayes estimators under our framework, and hence we provide a method of uncertainty quantification for these approaches; for example, allowing calculation of the probability a data point is well clustered.

E0663: **Bayesian estimates from loss functions**
Presenter: **Yu Luo**, Imperial College London, United Kingdom

In the usual Bayesian setting, a full probabilistic model is required to link the data and parameters, but in general, such a model is not robust to model misspecification. An alternative that has gained attention in the frequentist domain is to utilize decision theory, and draw inference via loss functions without direct reference to a probability model for the observable quantities. Recently, there has been much research on Bayesian inference via loss functions, with a predominant focus on Gibbs posteriors. We will introduce another perspective to generate Bayesian estimates from loss functions via the Bayesian decision theory and non-parametric Bayesian inference. In particular, this updating framework generalizes the Bayesian bootstrap approach through Bayesian predictive inference instead of standard prior-to-posterior inference. Examples are drawn from causal inference.

E1345: **General Bayesian inference**
Presenter: **Pier Giovanni Bissiri**, -, Italy
Co-authors: Chris Holmes, Stephen Walker

A coherent procedure is considered for general Bayesian inference based on updating a prior belief distribution to a posterior when the parameter of interest is connected to observations via a loss function. If such loss is the negative log-likelihood, then the Bayesian approach is recovered as a natural special case. This general updating process follows from a decision-theoretic approach involving cumulative loss functions where the Kullback-Leibler divergence plays a central role. Moreover, it is the only updating mechanism satisfying a coherence property together with some other natural assumptions.

EO246 Room Virtual R02 STATISTICAL ANALYSIS FOR STOCHASTIC DIFFERENTIAL EQUATIONS Chair: Masayuki Uchida

E1234: **Partial mixing and asymptotic expansion for batched bandits**
Presenter: **Nakahiro Yoshida**, University of Tokyo, Japan

The asymptotic expansion method based on partial mixing was proposed previously and applied to the asymptotic expansion of the additive functional of a partially mixing epsilon-Markov process, such as a jump-diffusion process satisfying a stochastic differential equation in the random environment. We discuss an application of this scheme to the asymptotic expansion of an estimator appearing in the batched bandits. In the batched bandit, the environment of each stage is randomly set according to the random outcomes of the previous stage. We introduce a backwards asymptotic expansion formula and assess the backward propagation of errors.

E0642: **Threshold estimation for jump-diffusions under small noise asymptotics**
Presenter: **Yasutaka Shimizu**, Waseda University, Japan
Co-authors: Mitsuki Kobayashi

The focus is on parameter estimation of stochastic differential equations driven by a Wiener process and a compound Poisson process as small noises. The goal is to give a threshold-type quasi-likelihood estimator and show its consistency and asymptotic normality under new asymptotics. One of the novelties is that we give a new localization argument, which enables us to avoid truncation in the contrast function that has been used in earlier works and to deal with a wider class of jumps in threshold estimation than ever before.

E1241: **Pathwise optimization for adaptive bridge-type estimators and its application to SDEs**
Presenter: **Alessandro De Gregorio**, University of Rome La Sapienza, Italy
Co-authors: Francesco Iafrate

The focus is on the bridge-type estimators arising from optimization problems with multiple adaptive l^q-penalties. By resorting to some tools arising from the nonconvex optimization theory, we introduce algorithms for computing the full solution path for the introduced estimators, for any possible value of the penalization parameter. We highlight that, up to our knowledge, this is the first attempt to introduce computational efficient methods in the bridge estimation setting. Furthermore, we discuss some applications of our approach to stochastic differential equations.

E1137: **Estimation of invariant density for a discretely observed diffusion: Impact of the sampling and of the asynchronicity**
Presenter: **Arnaud Gloter**, Universite d Evry Val d Essonne, France
Co-authors: Chiara Amorino

The purpose is to estimate in a non-parametric way the density π of the stationary distribution of a d-dimensional stochastic differential equation $(X_t)_{t\in[0,T]}$, for $d \geq 2$, from the discrete observations of a finite sample $X_{t_0},\ldots,X_{t_n}$ with $0 = t_0 < t_1 < \ldots < t_n = T$. We propose a kernel density estimator, and we study its convergence rates for the pointwise estimation of the invariant density under anisotropic Hölder smoothness constraints. We first find some conditions on the discretization step that ensures it is possible to recover the same rates as if the continuous trajectory of the process was available. As proven recently, such rates are minimax optimal and new in the context of density estimator. If such a condition on the discretization step is not satisfied, we also identify the convergence rate for the estimation of the invariant density. When the data are asynchronous, meaning that different components can be observed at different instants, the computation of the variance of the estimator is more difficult. We find conditions ensuring that this variance is comparable to the one of the continuous case. We also exhibit that the non-synchronicity of the data introduces additional bias terms in the study of the estimator.

EO676 Room Virtual R03 STATISTICS OF EXTREMES Chair: Jonathan El Methni

E0345: **Optimal pooling and distributed inference for the tail index and extreme quantiles**
Presenter: **Gilles Stupfler**, ENSAI - CREST, France
Co-authors: Abdelaati Daouia, Simone Padoan

The purpose is to investigate pooling strategies for tail index and extreme quantile estimation from heavy-tailed data. To fully exploit the information contained in several samples, we present general weighted pooled Hill estimators of the tail index and weighted pooled Weissman estimators of extreme quantiles calculated through a nonstandard geometric averaging scheme. We develop their large-sample asymptotic theory across a fixed number of samples, covering the general framework of heterogeneous sample sizes with different and asymptotically dependent distributions. Our results include optimal choices of pooling weights based on asymptotic variance and MSE minimization. In the important application of distributed inference, we prove that the variance-optimal distributed estimators are asymptotically equivalent to the benchmark Hill and Weissman estimators based on the unfeasible combination of subsamples, while the AMSE-optimal distributed estimators enjoy a smaller AMSE than the benchmarks in the case of large bias. We consider additional scenarios where the number of subsamples grows with the total sample size and effective subsample sizes can be low. We extend our methodology to handle serial dependence and the presence of covariates. Simulations confirm the statistical inferential theory of our pooled estimators. A real data application is discussed.

E0782: **Bias- and variance-corrected asymptotic Gaussian inference about extreme expectiles**
Presenter: **Antoine Usseglio-Carleve**, Avignon Universita, France
Co-authors: Gilles Stupfler, Abdelaati Daouia

The expectile is a prime candidate for being a standard risk measure in actuarial and financial contexts, for its ability to recover information about probabilities and typical behavior of extreme values as well as its excellent axiomatic properties. A series of recent papers have focused on expectile estimation at extreme levels, with a view to gathering essential information about low-probability, high-impact events that are of most interest to risk managers. Actual inference about extreme executives is a difficult question, however, due to their least squares formulation making them very sensitive to tail heaviness, even though the obtention of accurate confidence intervals is paramount if the expectile risk measure is to be used in practical applications. The focus is on asymptotic Gaussian inference about tail expectiles in the challenging context of heavy-tailed observations. We use an in-depth analysis of the proofs of asymptotic normality results for two classes of extreme expectile estimators to derive bias- and variance-corrected Gaussian confidence intervals. These, unlike previous attempts in the literature, are well-rooted in statistical theory and can accommodate underlying distributions that display a wide range of tail behaviors. A large-scale simulation study and real data analyses confirm the versatility of the proposed technique.

E1093: **High-dimensional variables clustering of a weakly dependent random process for sub-asymptotic maxima.**
Presenter: **Alexis Boulin**, Universite Cote d'Azur and Inria, Lemon, France
Co-authors: Elena Di Bernardino, Thomas Laloe, Gwladys Toulemonde

The dependence structure between extreme observations can be complex. For that purpose, we see clustering as a tool for learning the complex extremal dependence structure. We introduce the Asymptotic Independent block (AI-block) model, a model-based clustering where population-level clusters are clearly defined using independence of clusters' maxima of a multivariate random process. This class of models is identifiable allowing statistical inference. With a dedicated algorithm, we show that sample versions of the extremal correlation can be used to recover the clusters of variables without specifying the number of clusters. Our algorithm has a computational complexity that is polynomial in the dimension and it is shown to be strongly consistent in growing dimensions where observations are drawn from a stationary mixing process. This implies that groups can be learned in a completely nonparametric inference in the study of dependent processes where block maxima are only subasymptotic, i.e., approximately extreme value distributed.

E1225: **Extreme expectile estimation in heavy-tailed regression models**
Presenter: **Yasser Abbas**, Fondation Jean-Jacques Laffont, France
Co-authors: Abdelaati Daouia, Gilles Stupfler

Studying rare events at the heavy tails of Pareto-type distributions is a burgeoning science and has many applications both in and out of finance. Most attempts to tackle the subject involve quantile regression, which usually offers a natural way of examining the impact of covariates at different levels of the dependent variable. We argue, however, that quantiles are not well equipped to deal with sparsity around the tails, especially in the active field of risk management where they fail to satisfy the coherency axiom, and motivate their least-square analogues, expectiles, as a more appropriate alternative. We introduce versatile estimators of extreme conditional expectiles under an additive regression model with heavy-tailed noise and derive their asymptotic properties in a general setting. We then tailor the discussion to the linear and local linear estimation settings. We showcase the performance of our procedures in a detailed simulation study and apply them to a concrete dataset.

EO240 Room Virtual R04 DIRECTIONAL STATISTICS Chair: Toshihiro Abe

E0719: **Improve direct plug-in rule selector for circular kernel density estimation**
Presenter: **Yasuhito Tsuruta**, The University of Nagano, Japan

A circular kernel density estimation is a nonparametric method for exploring the density structure of circular data without model specifications because it flexibly changes its shape through the choice of the smoothing parameter. A huge smoothing parameter leads to undersmoothing, and the density plot looks like a multimodal density and brings wasteful zigzags. Whereas a small smoothing parameter leads to oversmoothing, and the density plot looks like a unimodal distribution and hides all non-unimodal distribution properties. It requires appropriately selecting a smoothing parameter. The optimal parameter as the minimizer of the mean integrated error depends on a functional of an underlying density. Therefore, there is a lot of research on estimating the optimal smoothing parameter. The direct plug-in rule (DPI) selector is the kernel functional estimator for the optimal parameter. However, DPI selector also requires choosing the pilot smoothing parameter. The minimizer of the mean squared error of its selector also depends on the functional of an underlying density. Therefore, a new kernel functional estimator for its pilot parameter is proposed. The proposed estimator is shown to be asymptotic normal and consistent. The numerical experiment is conducted to investigate the small sample characteristic of the proposed estimator.

E0882: **An extension of Johnson-Wehrly-type cylindrical distributions**
Presenter: **Yoichi Miyata**, Takasaki City University of Economics, Japan

The Johnson-Wehrly-type cylindrical distributions in which the circular marginal distributions are the sine-skewed wrapped Cauchy and the conditional distributions of a linear random variable are the Weibull are known to be identifiable, and have simple normalizing constants and easy random number generation algorithms. Extending the models, we propose a new family of cylindrical distributions in which the circular marginal distributions are the extended sine-skewed wrapped Cauchy. The proposed distributions also have simple normalizing constants and easy random number generation algorithms, and allow marginal distributions of a circular random variable to be skewed more significantly than that of the above models. Numerical examples of fitting this model to the Periwinkle data and a hidden Markov model with each component being the proposed distribution are shown.

E1462: **Complex-valued time-series models and their relation to directional statistics**
Presenter: **Takayuki Shiohama**, Nanzan University, Japan

Stationary time series fluctuation often exhibits periodic behavior and these patterns are summarized via a spectral density, which can be modeled using a circular distribution function. Several time-series models are studied in relation to a circular distribution. First, we illustrate how to model bivariate time-series data using complex-valued time series in the context of circular distribution functions. These models are then extended to have a skewed spectrum by incorporating a sine-skewing transformation. Further, two parameter estimation methods are introduced, and their asymptotic properties are investigated. These theoretical results are verified via a Monte Carlo simulation. In addition, real data analyses are performed to illustrate the applicability of the proposed models.

E1463: **New construction of a cylindrical distribution from two base distributions**
Presenter: **Tomoaki Imoto**, University of Shizuoka, Japan

In diverse scientific fields, data often appear that can be represented as points in the circumference of a unit circle, called circular data. In many cases, observations are made on linear and circular variables to discover some relationships like wind speed and direction. Such data is called cylindrical data. We propose a method for constructing a distribution for modeling cylindrical data. The difference from previous work is that the support of the linear variable of the constructed distribution can be arbitrary. This method is one of the specific marginals methods, and the conditional distribution of the circular variable becomes a well-studied distribution belonging to a family of sine-skewed distributions. The other research about estimation and application for fitting real data are also shown.

EO170 Room K2.31 (Nash Lec. Theatre) ADVANCES IN HETEROGENEOUS AND IMAGING DATA ANALYSIS Chair: Simon Vandekar

E1590: **Bayesian quantile latent factor on image regression**
Presenter: **Chuchu Wang**, The Chinese University of Hong Kong, Hong Kong
Co-authors: Qi Yang, Xiaoxiao Zhou, Xinyuan Song

A quantile latent factor-on-image (Q-LoI) regression model is considered to comprehensively investigate the relationship between the latent factor of interest and scalar and imaging predictors at different quantiles. The latent factor is characterized by several manifest variables through a confirmatory factor analysis model and then regressed on scalar and imaging covariates. We propose a two-stage method to conduct statistical inference. The first stage extracts leading features from the imaging data through the functional principal component analysis (FPCA) method. The second stage incorporates the extracted imaging features into the Q-LoI regression to examine the impacts of scalar and imaging covariates on the latent factor under various quantiles. A fully Bayesian method with Markov Chain Monte Carlo (MCMC) algorithms is developed for parameter estimation. Simulation studies demonstrate the satisfactory performance of the proposed method. An application to the Alzheimers disease study is presented to confirm the utility of our methodology.

E1595: **Bayesian order selection in heterogeneous hidden Markov models**
Presenter: **Yudan Zou**, The Chinese University of Hong Kong, Hong Kong
Co-authors: Xinyuan Song, Yiqi Lin

Hidden Markov models (HMMs) are valuable tools for analyzing longitudinal data due to their capability to describe dynamic heterogeneity. Conventional HMMs typically assume that the number of hidden states (i.e., the order of HMMs) is known or predetermined through criterion-based methods. However, prior knowledge about the order is often unavailable, and a pairwise comparison using criterion-based methods becomes increasingly tedious and computationally demanding when the model space enlarges. A few studies have considered simultaneously performing order selection and parameter estimation under the frequentist framework. Still, they focused only on homogeneous HMMs and thus cannot accommodate situations where potential covariates affect the between-state transition. A Bayesian double-penalized (BDP) procedure is proposed to conduct a simultaneous order selection and parameter estimation for heterogeneous HMMs. We develop a novel Markov chain Monte Carlo algorithm coupled with an efficient adjust-bound reversible jump scheme to address the challenges in updating the order. Simulation studies show that the proposed BDP procedure considerably outperforms the commonly used criterion-based methods. An application to the Alzheimer's Disease Neuroimaging Initiative study further confirms the utility of the proposed method.

E1706: **A joint mixed membership model for multivariate longitudinal and survival data**
Presenter: **Yuyang He**, The Chinese University of Hong Kong, Hong Kong

In longitudinal studies, a conventional approach to capture the individual heterogeneity of the population is to adopt a finite mixture model by assigning each subject to a single cluster. However, some subjects may not exactly correspond to one typical cluster, but instead, behave somewhere in between two (or more) clusters. We propose a new joint mixed membership model to address such heterogeneity and investigate the relationship between multivariate longitudinal and survival data. A vector of probability weights for characterizing partial membership is introduced both on (i) a mixed-effects model for describing the trajectories of longitudinal observations and (ii) a PH model for examining the effects of time-dependent risk factors on the hazard of interest. We develop a Bayesian joint estimation method coupled with efficient Markov chain Monte Carlo sampling schemes to perform statistical inference for the new joint model. The proposed approach is assessed through extensive simulation studies and an application to Alzheimer's Disease Neuroimaging Initiative study.

E1707: **Longitudinal mixed membership image-on-scalar model for learning the progression of Alzheimers disease**
Presenter: **Zhihao Wu**, The Chinese University of Hong Kong, Hong Kong

While magnetic resonance imaging (MRI) data has been widely used for the diagnosis and/or prediction of Alzheimers disease (AD), very little attention has been paid to the individual heterogeneity in terms of longitudinal MRI data. We propose a novel modeling framework for describing the dynamic pattern of longitudinal imaging data and use the proposed model to learn the progression of AD. First, a basis expansion approach is adopted to approximate the longitudinal images. Then, we introduce a vector of probability weights characterizing mixed membership to capture the individual heterogeneity. Finally, the approximated longitudinal images are modeled using regression models under typical membership. A Bayesian approach coupled with MCMC methods is developed to conduct statistical inference. Simulation results demonstrate a good performance of estimation under a medium sample size. The approach is applied to the Alzheimers Disease Neuroimaging Initiative (ADNI) to discover the dynamic pattern of brain regions in AD progression.

EO256 Room K2.40 ADVANCES IN FUNCTIONAL AND OBJECT DATA ANALYSIS Chair: Sonja Greven

E0881: **Functional linear regression for discretely observed data: from ideal to reality**
Presenter: **Fang Yao**, Peking University, China
Co-authors: Hang Zhou, Huiming Zhang

Despite extensive study on functional linear regression, there exists a fundamental gap in theory between the ideal estimation from fully observed covariate functions and the reality that one can only observe functional covariates discretely with noise. The challenge arises when deriving a sharp perturbation bound for the estimated eigenfunctions in the latter case, which renders existing techniques for functional linear regression not applicable. We use a pooling method to attain the estimated eigenfunctions and propose a sample-splitting strategy to estimate the principal component scores, which facilitates the theoretical treatment for discretely observed data. The slope function is estimated by approximated least squares, and we show that the resulting estimator attains the optimal convergence rates for both estimation and prediction when the number of measurements per subject reaches a certain magnitude of the sample size. This phase transition phenomenon differs from the known results for

the pooled mean and covariance estimation, and reveals the elevated difficulty in estimating the regression function. Numerical experiments, using simulated and real data examples, yield favourable results when compared with existing methods.

E0687: **Interpretable discriminant analysis for functional data supported on random non-linear domains**
Presenter: **Eardi Lila**, University of Washington, United States

A novel framework is proposed for the classification of functional data supported on non-linear, and possibly random, manifold domains. The motivating application is the identification of subjects with Alzheimer's disease from their cortical surface geometry and associated cortical thickness map. The proposed model is based upon a reformulation of the classification problem as a regularized multivariate functional linear regression model. This allows us to adopt a direct approach to the estimation of the most discriminant direction while controlling for its complexity with appropriate differential regularization. We apply the proposed method to a pooled dataset from the Alzheimer's Disease Neuroimaging Initiative and the Parkinson's Progression Markers Initiative, and are able to estimate discriminant directions that capture both cortical geometric and thickness predictive features of Alzheimer's Disease.

E1422: **Elastic full procrustes analysis of plane curves via Hermitian covariance smoothing**
Presenter: **Almond Stoecker**, Ecole polytechnique federale de Lausanne, Switzerland
Co-authors: Manuel Pfeuffer, Lisa Steyer, Sonja Greven

Determining the mean shape of a collection of curves is not a trivial task, in particular when curves are only irregularly/sparsely sampled at discrete points. We propose an elastic full Procrustes mean of shapes of (oriented) plane curves, which are considered equivalence classes of parameterized curves with respect to translation, rotation, scale, and re-parameterization (warping), based on the square-root-velocity framework. Identifying the real plane with the complex numbers, we establish a connection to covariance estimation in irregular/sparse functional data analysis and propose Hermitian covariance smoothing for (in)elastic full Procrustes mean estimation. We offer an implementation in the R package elastes and demonstrate the performance of the approach in a phonetic study on tongue shapes and in different realistic simulation settings, inter alia based on handwriting data.

E1286: **Continuous-time multivariate analysis**
Presenter: **Philip Reiss**, University of Haifa, Israel
Co-authors: Biplab Paul

The starting point for much of multivariate analysis (MVA) is an $n \times p$ data matrix whose n rows represent observations and whose p columns represent variables. Some multivariate data sets, however, may be best conceptualized not as n discrete p-variate observations, but as p curves or functions defined on a common time interval. Such a viewpoint may be useful for multivariate data observed at very high time resolution, with unequal time intervals, and/or with substantial missingness. We introduce a framework for extending techniques of multivariate analysis to such settings. The proposed framework rests on the assumption that the curves can be represented as linear combinations of basis functions such as B-splines. This is formally identical to the Ramsay-Silverman representation of functional data; but whereas functional data analysis extends MVA to the case of observations that are curves rather than vectors – heuristically, $n \times p$ data with p infinite – we are instead concerned with what happens when n is infinite. We demonstrate a new R package that translates the classical MVA methods of principal component analysis, Fisher's linear discriminant analysis, and k-means clustering to the above continuous-time setting. The methods are illustrated with a novel perspective on the well-known Canadian weather data set, as well as with applications to neurobiological and environmetric data.

EO070 Room K2.41 RECENT ADVANCES IN ANALYTICAL METHODS FOR LARGE-SCALE DATA Chair: Zhaoyuan Li

E1009: **Statistical physics approaches to the complex earth system**
Presenter: **Jingfang Fan**, Beijing Normal University, China

Global warming, extreme climate events, earthquakes and their accompanying socioeconomic disasters pose significant risks to humanity. Yet due to the nonlinear feedback, multiple interactions and complex structures of the Earth system, the understanding and, in particular, the prediction of such disruptive events represent formidable challenges to both scientific and policy communities. During the past years, the emergence and evolution of Earth system science has attracted much attention and produced new concepts and frameworks. Especially, novel statistical physics and complex networks-based techniques have been developed and implemented to substantially advance our knowledge of the Earth system, including climate extreme events, earthquakes and geological relief features, leading to substantially improved predictive performances. We present a comprehensive review of the recent scientific progress in the development and application of how combined statistical physics and complex systems science approaches such as critical phenomena, network theory, percolation, tipping points analysis, and entropy can be applied to complex Earth systems. Notably, these integrating tools and approaches provide new insights and perspectives for understanding the dynamics of the Earth systems.

E1030: **Minimum information dependence modeling**
Presenter: **Keisuke Yano**, The Institute of Statistical Mathematics, Japan
Co-authors: Tomonari Sei

A method of dependence modeling for a broad class of multivariate data is proposed. Our class is characterized by two orthogonal sets of parameters: the parameters of dependence and those of marginal distributions. We present the existence and uniqueness theorem for our model. To estimate the dependence parameter, we establish conditional inference together with a sampling procedure and show that conditional inference is asymptotically indistinguishable from the maximum likelihood inference. We also discuss the information-geometrical structure and the connection to the entropic optimal transport and the Schrodinger bridge problems. Finally, we illustrate an application to the earthquake data.

E1153: **A new CUSUM type procedure for sequential change detection**
Presenter: **Liyan Xie**, The Chinese University of Hong Kong, Shenzhen, China
Co-authors: George Moustakides, Yao Xie

The parametric online changepoint detection problem is studied, where the underlying distribution of the streaming data changes from a known distribution to an alternative that is of a known parametric form but with unknown parameters. We propose a joint detection/estimation scheme, which we call Window-Limited CUSUM, that combines the cumulative sum (CUSUM) test with a sliding window-based consistent estimate of the post-change parameters. We characterize the optimal choice of window size and show that the Window-Limited CUSUM enjoys first-order asymptotic optimality. Compared to existing schemes with similar asymptotic optimality properties, our test is far simpler in implementation because it can recursively update the CUSUM statistic by employing the estimate of the post-change parameters. A parallel variant is also proposed that facilitates the practical implementation of the test. Numerical simulations corroborate our theoretical findings.

E1006: **Change point inference for high-dimensional correlation matrix**
Presenter: **Zhaoyuan Li**, The Chinese University of Hong Kong, Shenzhen, China

The focus is on the problem of detecting and estimating a change point in the correlation matrix in a sequence of high dimensional vectors, where the dimension is substantially large compared to the sample size. We first propose a simulation-based approach to detect whether a change point exists. When we have witnessed a change point in the first step, a two-stage method is proposed to estimate the location of the change point. The first step involves reducing the dimension to identify elements of the correlation matrices corresponding to significant changes, where a simulation-based procedure generates a threshold. In the second step, we use the components after dimension reduction to determine the position of the change point. This method can efficiently estimate the change point located in the middle of the sequence and at the tails, which will be very useful for

online detection. Theoretical properties are developed for both approaches, and numerical studies are conducted to support the new methodology.

CI023 Room BH (SE) 2.12 ALTERNATIVE DATA IN FINANCE Chair: Serge Darolles

C0363: **Does alternative data improve financial forecasting?**
Presenter: **Thierry Foucault**, HEC Paris, France
Co-authors: Olivier Dessaint, Laurent Fresard

Existing research suggests that alternative data is mainly informative about short-term future outcomes. We show theoretically that the availability of short-term oriented data can induce forecasters to optimally shift their attention from the long-term to the short-term because it reduces the cost of obtaining short-term information. Consequently, the informativeness of their long-term forecasts decreases, even though the informativeness of their short-term forecasts increases. We test and confirm this prediction by considering how the informativeness of equity analysts' forecasts at various horizons varies over the long run and with their exposure to social media data.

C1759: **When are Google data useful to nowcast GDP: An approach via preselection and shrinkage**
Presenter: **Anna Simoni**, CNRS - CREST, France
Co-authors: Laurent Ferrara

Alternative data sets are widely used for macroeconomic nowcasting together with machine learning-based tools. The latter are often applied without a complete picture of their theoretical nowcasting properties. Against this background, a theoretically grounded nowcasting methodology is proposed that allows researchers to incorporate alternative Google Search Data (GSD) among the predictors and that combines targeted preselection, Ridge regularization, and Generalized Cross Validation. Breaking with most existing literature, which focuses on asymptotic in-sample theoretical properties, we establish the theoretical out-of-sample properties of our methodology and support them with Monte-Carlo simulations. We apply our methodology to GSD to nowcast GDP growth rate of several countries during various economic periods. Our empirical findings support the idea that GSD tends to increase nowcasting accuracy, even after controlling for official variables, but that the gain differs between periods of recessions and of macroeconomic stability.

C0365: **Alternative data for ESG events monitoring using NLP: Practical quantitative results**
Presenter: **Sylvain Forte**, SESAMm, France

The purpose is to analyze how negative ESG events (controversies) can be extracted using natural language processing technologies on millions of sources of web data in real time. We test several techniques to create equity long-short strategies that leverage this information to generate alpha.

CO330 Room S-1.01 MIXED-FREQUENCY METHODS IN FINANCE AND ECONOMICS Chair: Kris Boudt

C0817: **Improving the supervisors anti-money laundering risk rating approach using news event monitoring**
Presenter: **Kris Boudt**, UGent, VUB, VUA, Belgium
Co-authors: Olivier Delmarcelle, Pascal Ringoot

A tool is developed to support the regulator's risk assessment process of financial institutions in the context of money laundering and financing terrorism. As part of their mandate to safeguard financial stability, supervisors rate each institution and provide them advice on how to improve their policy. This is most often done using a manual analysis and expert judgement. Due to the yearly frequency of this process, it is important not to overlook important developments between successive reportings. It is shown how the integration of a news event and monitoring system in the Belgian supervisors' AML process reduces the manual work and can preselect relevant news articles to ease the risk rating. We specify how natural language processing and fuzzy matching techniques improve the selection process and how an articles importance is calculated by combining the keywords and institutions' relevance. Finally, we demonstrate how well this compares to the current news monitoring.

C0824: **Power enhancement in detecting sparse signals, with applications to correlated test statistics in finance**
Presenter: **Nabil Bouamara**, UCLouvain, Belgium
Co-authors: Sebastien Laurent, Shuping Shi

A simple tool is introduced to control for false discoveries and identify individual signals when there are many tests, the test statistics are correlated and the signals are potentially sparse. In such situations, the Cauchy combination test aims for a global statement over a set of null hypotheses by transforming and summing individual p-values. We unravel the combination test to find out which of the p-values trigger the global test to, for example in the context of time series data, timestamp rejections. We also revisit two multiple-hypothesis testing problems in financial econometrics for which the test statistics have either serial dependence or cross-sectional dependence. We conclude that using the raw p-values in another way boosts the power compared to the workhorse procedures and bypasses a lot of the drawbacks inherent in extreme value tests, quadratic forms, thresholding and screening techniques.

C0906: **An Islamic alter ego for Dow Jones Industrial Average portfolio**
Presenter: **Mulazim Ali Khokhar**, Vrije Universiteit Brussels, Belgium
Co-authors: Kris Boudt, Dawood Ashraf

A unique Islamic equity index is developed by using the qualitative and quantitative features extracted from the expert committee decisions about the inclusion/exclusion of stocks in the Dow Jones Industrial Average (DJIA) portfolios, the world's oldest index, using machine learning. While the DJIA is the reference portfolio for conventional investors, it is not acceptable for Islamic investors. Over the past years, more than half of its portfolio was invested in Shariah-incompliant stocks. A naive approach of applying Shariah compliance filters on DJIA portfolios renders unrealistic invested in only around 15 stocks. According to previous work, one needs around 30 stocks to be diversified. A similar problem was observed for the Dow Jones US universe, for which a range of Shariah-compliant portfolios was developed under the flagship of Dow Jones Islamic Market (DJIM) portfolios. Particularly the DJIM-titan 50 index, where the portfolio consists of 50 stocks. It restricts the universe to only the Shariah-compliant top 50 market capitalization stocks in the US and applies market capitalization weighting. Although DJIM-titan 50 portfolio constitutes Shariah-compliant blue-chip companies in the US, it lacks the sophistication in selection criteria and human expert judgement of DJIA expert committee. We intend to create simple rule-based algorithms of DJIA expert committee alter egos and create DJIA style Islamic portfolio and its smart beta alternatives.

C1939: **Covariance matrix regularization through resampling**
Presenter: **Kirill Dragun**, VUB, UGent, Belgium
Co-authors: Kris Boudt, Steven Vanduffel

Many covariance matrix estimators achieve higher reliability than the sample covariance matrix at the expense of positive semi-definiteness. A typical example is the element-wise estimation of the covariance matrix using the implied covariance from the estimated variance of a sum of two random variables. The resulting estimator can be called a pre-estimator which then needs refinement to be transformed into a positive semi-definite matrix. The most popular transformations are shrinkage and eigenvalue cleaning. We propose methods for adjustments of the covariance matrix estimates in a way to get it positive semidefinite, taking into account the distribution properties of the original estimator. While the primary goal of imposing the adjustment is to achieve positive semidefiniteness substantial accuracy gains are observed too.

CO090 Room Virtual R01 ADVANCES IN BAYESIAN COMPUTATIONAL METHODS Chair: David Nott

C0192: **Analytic natural gradient updates for Cholesky factor in Gaussian variational approximation**
Presenter: **Siew Li Linda Tan**, National University of Singapore, Singapore

Stochastic gradient methods have enabled variational inference for high-dimensional models and large datasets. However, the steepest ascent direction in the parameter space of a statistical model is actually given by the natural gradient, which premultiplies the widely used Euclidean gradient by the inverse of the Fisher information matrix. The use of natural gradients can improve convergence, but inverting the Fisher information matrix is daunting in high dimensions. In Gaussian variational approximation, natural gradient updates of the mean and precision matrix of the Gaussian distribution can be derived analytically, but do not ensure the precision matrix remains positively definite. To tackle this issue, we consider the Cholesky decomposition of the covariance or precision matrix, and derive analytic natural gradient updates of the Cholesky factor, which depend only on the first derivative of the log posterior density. Efficient natural gradient updates of the Cholesky factor are also derived under sparsity constraints representing different posterior correlation structures. As Adam's adaptive learning rate does not seem to pair well with natural gradients, we propose using stochastic normalized natural gradient ascent with momentum. The efficiency of the proposed methods is demonstrated using generalized linear mixed models.

C0864: **Fast variational inference for multinomial probit models**
Presenter: **Ruben Loaiza-Maya**, Monash University, Australia
Co-authors: Didier Nibbering

The multinomial probit model is often used to analyze choice behaviour. However, estimation with existing Markov Chain Monte Carlo (MCMC) methods is computationally costly, which limits its applicability to large choice data sets. A variational inference method is proposed that is fast, even when a large number of choice alternatives and observations are considered. Variational methods usually require an analytical expression for the unnormalized posterior density and an adequate choice of the variational family. Both are challenging to specify in a multinomial probit, which has a posterior that requires identifying restrictions and is augmented with a large set of latent utilities. We employ a spherical transformation on the covariance matrix of the latent utilities to construct an unnormalized augmented posterior that identifies the parameters, and use the conditional posterior of the latent utilities as part of the variational family. The proposed method is faster than MCMC, and can be made scalable to both a large number of choice alternatives and a large number of observations. The accuracy and scalability of our method are illustrated in numerical experiments and real purchase data with one million observations

C1047: **Spike-and-slab group lasso meets Bayesian P-splines**
Presenter: **Paul Bach**, Humboldt University Berlin, Germany
Co-authors: Nadja Klein

The Spike-and-Slab Group Lasso (SSGL) is combined with Bayesian P-splines to obtain a powerful Bayesian approach for effect selection and estimation in sparse high-dimensional additive models. The proposed method is able to decide whether a covariate has a linear, a nonlinear or no effect at all on the response. Moreover, it provides accurate effect estimates and valid posterior credible bands for uncertainty quantification. An interesting finding is that the original variant of the SSGL prior is not suitable in the present context because of severe MCMC mixing issues for the binary selection indicators. We attribute these issues to the sharp concentration of the Euclidean norm of the group Lasso distribution when the group dimension is not small and introduce a new variant of the SSGL prior for which MCMC mixing works much better. Another key feature of our approach is a new reparametrization that renders groupwise Gibbs updates extremely efficient. This reparametrization is closely related to the Demmler-Reinsch reparametrization but more applicable as it does not require full-rank design matrices. We compare the selection and estimation performance of the suggested method with that of several competitors in simulations and illustrate the applicability of our method by consideration of a real data example.

C1062: **Variational Bayes on manifolds and quantum speed-up**
Presenter: **Minh-Ngoc Tran**, University of Sydney, Australia

Most of the existing Variational Bayes (VB) algorithms is generally restricted to the case where the variational parameter space is Euclidean, which hinders the potentially broad application of the VB method. As the first contribution, we extend the scope of VB to the case where the variational parameter space is a Riemannian manifold. We develop an efficient manifold-based VB algorithm that exploits both the geometric structure of the constraint parameter space and the information geometry of the manifold of VB approximating probability distributions. The natural gradient is an essential component of efficient VB estimation, but it is prohibitively computationally expensive in high dimensions. As the second contribution, we propose a regression-based stochastic approximation of the natural gradient, a computationally efficient method with provable convergence guarantees under standard assumptions. This regression formulation enables further computational speedup through the use of quantum computation, particularly quantum matrix inversion. We demonstrate that the problem setup fulfils all the conditions required for quantum matrix inversion to deliver computational efficiency.

CO724 Room BH (SE) 1.01 PANEL DATA Chair: Martin Schumann

C0291: **On the incidental parameter problem in fractional response models with fixed effects**
Presenter: **Amrei Stammann**, Ruhr-University Bochum, Germany

The incidental parameter problem is studied in fixed effects quasi-maximum likelihood estimators for fractional response models. We uncover a special case that comprises many relevant empirical applications. Opposed to results known from the earlier literature, estimates for structural parameters and average partial effects often exhibit surprisingly small biases even when the number of time periods is small. We provide a theoretical explanation for this phenomenon and investigate the effectiveness of bias correction methods in simulation experiments.

C0530: **Bounds on average effects in discrete choice panel data models**
Presenter: **Cavit Pakel**, Bilkent University, Turkey
Co-authors: Martin Weidner

Average effects in discrete choice panel data models with individual-specific fixed effects are generally only partially identified in short panels. While consistent estimation of the identified set is possible, it generally requires very large sample sizes, especially when the number of support points of the observed covariates is large, such as when the covariates are continuous. We propose estimating outer bounds on the identified set of average effects. Our bounds are easy to construct, converge at the parametric rate, and are computationally simple to obtain even in moderately large samples, independent of whether the covariates are discrete or continuous. We also provide asymptotically valid confidence intervals on the identified set. Simulation studies confirm that our approach works well and is informative in finite samples. We also consider an application to labor force participation.

C0612: **Testing for equivalence of pre-trends in difference-in-differences estimation**
Presenter: **Martin Schumann**, Maastricht University, Netherlands
Co-authors: Holger Dette

The plausibility of the parallel trends assumption (PTA) in Difference-in-Differences estimation is usually assessed by a test of the null hypothesis that the difference between the means of both groups is constant over time before the treatment. However, failure to reject the null hypothesis does not imply the absence of differences in time trends between both groups. We provide three tests of equivalence that allow researchers to specify

a threshold below which differences in trends are deemed negligible. Our test procedures build on a simple regression and can thus be adapted to various situations of interest, e.g. heterogeneous treatment timings.

C0946: **Integrated likelihood based inference for dynamic binary choice panel data models with fixed effects**
Presenter: **Sofia Borodich Suarez**, University of Luxembourg, Luxembourg
Co-authors: Martin Schumann, Gautam Tripathi
An integrated likelihood approach is used to estimate the parameters and marginal effects in an AR(1) binary choice panel data model with fixed effects. Additional covariates in the model are treated as being predetermined each period, which allows for feedback from the current outcome to the future covariates.

C2040: **Nonparametric bootstrap correction for incidental parameter bias in GMM**
Presenter: **Yitian Li**, KU Leuven, Belgium
Co-authors: Geert Dhaene
The incidental parameter problem has been extensively studied in the context of parametric maximum likelihood estimation. In the last two decades, many methods have been developed to correct (or approximately correct) the bias of maximum likelihood estimates that arises when incidental parameters are present in the model. Nearly all these methods exploit the parametric likelihood structure of the model. In the more general context of GMM estimation, where a full parametric likelihood is lacking, the incidental parameter problem has been much less studied. We show that the nonparametric bootstrap can be used for approximate bias correction in the GMM framework. Our method also yields a novel bias correction in the likelihood setting as a special case. The bias correction can be applied directly to the GMM/ML estimator or to the estimating equations that define the GMM/ML estimator. We also show that the bias correction can be iterated, thereby reducing the asymptotic order of the bias and improving the coverage rate of confidence intervals at each iteration of the bootstrap. We discuss various numerical examples and simulations, including nonlinear models, and show that the method performs well.

CO699 Room BH (SE) 1.02 VOLATILITY AND OPTION PRICING MODELS Chair: Arnaud Dufays

C0595: **Factor dynamics, risk premia, and higher moments in multi-factor option pricing models**
Presenter: **Jeroen Rombouts**, ESSEC Business School, France
A class of multifactor Heston stochastic volatility models are considered to describe return and option dynamics jointly. The model can generate a wide range of stochastic volatility patterns, and flexible third and fourth-order moment dynamics. The estimation algorithm consists of an adapted particle MCMC sampler that allows fitting the models with long return series and large option panels. In the application, we estimate stochastic volatility models with up to three factors, jumps, and show that this additional flexibility generates more realistic volatility dynamics and risk premia. Furthermore, option fit greatly improves.

C0606: **A general framework for multifractal discrete stochastic volatility**
Presenter: **Arnaud Dufays**, EDHEC Business school, France
Co-authors: Maciej Augustyniak, Kassimou Abdoul Haki Maoude
Regime-switching processes are popular tools to interpret, model and forecast financial data. The Markov-switching multifractal (MSM) model has proved to be a strong competitor to the GARCH class of models for modeling the volatility of returns. In this model, volatility dynamics are driven by a latent high-dimensional Markov chain constructed by multiplying independent two-state Markov chains. We propose the multifractal discrete stochastic volatility (MDSV) model as a generalization of the MSM process and of other related high-dimensional hidden Markov models. Our model is intended to model financial returns jointly and realized volatilities, and therefore also extends existing high-dimensional Markov-switching processes to the joint setting. Our approach consists in building a high-dimensional Markov chain by the product of lower-dimensional Markov-chains which have a discrete stochastic volatility representation. The properties and structure of our model are studied theoretically, and it is shown that the MDSV process can be interpreted as a multi-component stochastic volatility model. An empirical study on 31 financial time series shows that the MDSV model can improve upon the realized EGARCH model in terms of fit and forecasting performance.

C1806: **Elicitability of marginal expected shortfall and related systemic-risk measures**
Presenter: **Jeremy Leymarie**, Edhec Business School, France
Co-authors: Ophelie Couperier, Olivier Scaillet, Sylvain Benoit
A risk measure, or more generally a statistical functional, is called elicitable if it can be defined as the minimizer of a suitable expected scoring function. The notion of elicitability (and identifiability) is explored for systemic-risk measures that are used to identify the financial institutions contributing the most to the overall risk in the financial system. Our elicitation framework applies to systemic-risk measures that are expressed as a function of the expected equity loss conditional on a financial crisis, such as the marginal expected shortfall (MES), the systemic expected shortfall (SES), or the systemic-risk measure SRISK. This property paves the way to the implementation of semiparametric M-estimation for the systemic-risk measures or to the comparison and backtesting of the systemic-risk models used by academics and policymakers to rank the systemically important financial institutions (SIFIs) whose failure might trigger a crisis in the entire financial system.

C0608: **Asymptotic efficiency for two-stage conditional M-estimators**
Presenter: **ELysee Aristide Houndetoungan**, Cy Cergy Paris Universite, France
Co-authors: Kassimou Abdoul Haki Maoude
Two-step estimation strategies are popular for dealing with many issues in applied econometrics, such as endogeneity, missing data, and latent variables. These approaches consist in estimating a control, which is used in a second estimation strategy. The asymptotic properties of the second stage estimator depend on the uncertainty in the first stage and can be challenging to establish. We study these properties for a general class of first-stage estimation procedures, where the second-stage estimator is an M-estimator. For example, our approach allows a Bayesian estimator or nonparametric estimator at the first stage. We develop a straightforward approach to consistently estimate the variance of the estimator at the second stage. Our method is computationally more attractive than the Bootstrap method as we do not perform several estimations.

CO700 Room BH (SE) 1.05 ALTERNATIVE DATA FOR ECONOMIC FORECASTING Chair: Luca Barbaglia

C0484: **Identifying monetary policy shocks: A natural language approach**
Presenter: **Thomas Drechsel**, University of Maryland, United States
A novel method is proposed for the identification of monetary policy shocks. By applying natural language processing techniques to documents that economists at the Federal Reserve prepare for Federal Open Market Committee meetings, we capture the information set available to the committee at the time of policy decisions. Using machine learning techniques, we then predict changes in the target interest rate conditional on this information set, and obtain a measure of monetary policy shocks as the residual. An appealing feature of our procedure is that only a small fraction of interest rate changes is attributed to exogenous shocks. We find that the dynamic responses of macroeconomic variables to our identified shocks are consistent with the theoretical consensus.

C0494: **Sentiment analysis of economic text: A lexicon-based approach**
Presenter: **Luca Barbaglia**, European Commission Joint Research Centre, Italy
Co-authors: Sergio Consoli, Sebastiano Manzan, Luca Tiozzo Pezzoli, Elisa Tosetti

The goal is to propose a dictionary specifically designed for textual applications in economics. We construct the dictionary with two important characteristics: 1) to have wide coverage of terms typically used in documents discussing economic and financial concepts, and 2) to provide a human-annotated sentiment score in the range [-1,1]. The sentiment score is a useful feature when the interest is to weigh words by their sentiment content as opposed to categorizing terms in positive and negative. We use the dictionary to construct a measure of economic pessimism and show that it captures the business cycle and correlates with measures of economic and financial uncertainty.

C0509: **Nowcasting Euro area GDP with news sentiment: A tale of two crises**
Presenter: **Julian Ashwin**, London Business School, United Kingdom
Co-authors: Lorena Saiz, Eleni Kalamara

The aim is to show that newspaper articles contain signals that can materially improve real-time nowcasts of real GDP growth for the euro area. Using articles from fifteen popular European newspapers, which are machine translated into English, we create daily sentiment metrics and assess their value for nowcasting, comparing to competitive and rigorous benchmarks. We find that newspaper text is especially helpful early in the quarter before other indicators are available. We also find that general-purpose sentiment measures perform better than more economics-focused ones in response to unanticipated events, and non-linear supervised models can help capture extreme movements in growth, but require sufficient training data in order to be effective.

C1317: **Nowcasting inflation using web searches**
Presenter: **Marco Colagrossi**, European Commission, Joint Research Centre, Italy
Co-authors: Luca Barbaglia, Sergio Consoli

Researchers are increasingly exploiting unconventional data sources to study phenomena for which timely and high-frequency data are not readily available. We exploit web search data obtained through a non-public Google API to improve fore- and now-casting performances of consumer price indexes (CPIs) across EU Member States. Starting from Google Search categories, we retrieve over seven thousand topics (which are aggregations of several web search queries that could be assigned to the same semantic domain) broadly related to the economy. Finally, using machine learning models, we show how by including the relevant set of searches, it is possible to improve fore- and now- casting performances of several CPIs across Europe.

CO650 Room BH (SE) 1.06 TIME SERIES: FORECASTING, NONLINEARITY AND MIXED FREQUENCY DATA Chair: Johan Lyhagen

C0628: **Test-based trimming for combined forecast**
Presenter: **Viktor Eriksson**, Uppsala University, Sweden

The common practice of combining all single forecasts falls short on not trimming off poor forecasts. We propose a new method for trimming in the combination of forecasts that utilizes a test statistic for differences in the AIC. Application on M4-data shows that test-based trimmed combination forecast typically has better forecasting performance than forecast produced by the best model. The improvement in forecasting performance by test-based trimming is greater for uniform combination weights than it is for Akaike combination weights.

C0618: **On the time-varying factor model and its application on finding minimum variance portfolio**
Presenter: **Chamika Porage**, Uppsala University, Sweden
Co-authors: Yukai Yang

A previous time-varying factor model is suggested by allowing the factor loadings to be smooth functions of random variables under certain conditions. We devise the corresponding consistent test for structural changes in factor loadings. Based on the extended time-varying factor model, we propose a new time-varying minimum variance portfolio in a large investment universe of assets. The method is suitable for short-term risk forecasting and hedging under a time-varying factor loadings context.

C0903: **Mixed Frequency data in a (S, s) pricing**
Presenter: **Jonas Andersson**, Norwegian School of Economics, Norway
Co-authors: Oivind A Nilsen, Hans J Skaug

In empirical analyses, prices are most often observed with a larger frequency than the explanatory variables. We overcome the mixed frequency issue by specifying a model where producers' monthly prices are functions of their (latent) monthly marginal costs, which are related to observed annual wage costs. The intermittency of the price changes is modelled using a stochastic (S, s) technique. We find that the mean distance between the frictionless price and the existing price has to be 33% before a price-adjustment process is initiated, which is smaller at the beginning of the year. The immediate pass-through of marginal costs to prices is approximately 0.30, i.e. significantly smaller than one, but larger in the long run. However, significant heterogeneity across sectors, firms and products is present. The analysis also shows that the intermittency blurs the picture when aggregating across time and/or firm/products. The rather low price-cost elasticity and the aggregation issue might have important implications for our understanding of the workhorse macro model, and in general, of sticky price models.

C1060: **Linearity testing in vector smooth transition autoregressive models when data are highly persistent**
Presenter: **Rickard Sandberg**, Stockholm School of Economics, Sweden

Asymptotic distributions are derived for linearity tests in vector smooth transition autoregressive (VSTAR) type of models in the presence of a unit root. The asymptotic distributions are non-standard because of the unit root assumption, and it is shown that the linearity hypothesis is rejected far too often (already up to 60% of the times at a 5% nominal significance level in bivariate systems) erroneously relying upon standard critical values. These findings will have practical implications because VSTAR models often are applied to data that are highly persistent, and the outcomes of standard linearity testing procedures in these cases should be interpreted with caution.

CO661 Room BH (SE) 2.05 ADVANCED STATISTICAL TOOLS IN SUSTAINABLE INSURANCE AND FINANCE Chair: Susanna Levantesi

C1174: **Deepening the relationship between ESG score and firms' performance via machine learning**
Presenter: **Susanna Levantesi**, Sapienza University of Rome, Italy
Co-authors: Rita DEcclesia, Valeria D Amato

Several firms are already adopting Environmental, Social, and Governance (ESG) into their governance, investment strategies, and risk management. Existing literature provides limited evidence of the relationship between the ESG score and the firm's profitability. We explore this matter by analyzing a sample of the companies constituting the EuroStoxx-600 index using different machine learning models. We aim to assess whether the ESG score significantly influences the firms' profitability measured by the earnings before interest and taxes (EBIT). We further deepen the relationship between ESG score and EBIT through machine learning interpretability toolboxes such as partial dependence plots and individual conditional expectation, which help to visualize the functional relationship between the predicted response and one or more features, and the Shapley value allowing us to examine the contribution of the feature to the prediction. Our findings show that the model can reach high levels of accuracy in detecting EBIT and that the ESG score is a promising predictor, compared to other traditional accounting variables.

C1094: **Loss given default in shipping finance: A machine learning approach**
Presenter: **Rita DEcclesia**, Sapienza University of Rome, Italy
Co-authors: Aida Salko

Different parametric and non-parametric modeling methods for estimating the Loss Given Default (LGD) of bank loans for shipping companies are analyzed. The shipping industry is subject to several other risks, which create the need to accurately measure the possible losses to estimate the LGDs for the banking industry. We use a unique database of defaulted loans in European banks involved in shipping finance. The aim is twofold: to compare the performance of alternative LGD modeling methodologies in shipping finance and to provide some insights into what drives LGD in the shipping industry. We find that non-parametric methods, predominantly random forest, lead to a remarkable increase in prediction accuracy and outperform the traditional statistical models in terms of both in-sample and out-of-sample results. To investigate the risk drivers in the shipping business, we use a variable importance measure built on the idea of permutation importance. We find the energy index to be paramount and the most important risk factor in estimating shipping finance LGD. We find that crude oil prices play a big role and may affect the financial health of shipping firms and then the LGDs of shipping loans.

C1259: **How to quantify the reputational risks due to ESG issues on the insurance companies activities**
Presenter: **Valeria D Amato**, University of Salerno, Italy

Insurance companies and pension funds are increasingly integrating ESG criteria in their business analysis. Nevertheless, only active greenwashing or explicit disregard for environmental protection, as well as only fulfilling reporting requirements, might lead to a big reputational issue for the company. Some studies highlight that reputational risk can quickly lead to actual litigation risks, such as regulatory sanctions or class actions. Actually, reputational risk can involve a significant loss of market value of portfolio assets arising from a controversial event, even though attributing losses to reputational risk may be technically challenging. The aim is to support actuaries and risk managers in identifying and evaluating reputational risks in the context of sustainability issues. We investigate if the ESG reputational risks amplify capital constraints, reducing firms growth opportunities, and increasing the firm probability of exiting the market. Meaningful research can arise from psychometrically sound measures of reputation. Conversely, to assess the value of reputation, we adopt the accounting approach in comparison with the financial one, specifically focusing on the positioning of insurance and pension companies.

C1442: **Sustainable health insurance: An application of GAMLSS for claims expenditure prediction**
Presenter: **Fabio Baione**, Sapienza University of Rome, Italy
Co-authors: Davide Biancalana

The Principles for Sustainable Insurance represents the milestone for insurance industries to manage and reduce Environmental, Social, and Governance (ESG) risks. To this aim, the interest in developing the so-called green insurance solutions represents a challenge for life and non-life insurance companies. Within the offer of significant social value covers, we can refer to products that supplement the public health service, designed to help manage the costs of treatment and assistance, as well as the reduction in earnings of customers in the event of serious illnesses or the loss of self-sufficiency. We apply Generalized additive models for location, scale, and shape (GAMLSS) to price health insurance which provides coverage for several categories of medical expenditure. It is worth noting that the coverage exhibits high repeatability. Hence the number of claims for each policyholder per year can be much greater than one, with a high probability. Moreover, depending on the claims category, we observe low or high frequencies and highly skewed claim size distributions. GAMLSS is a flexible class of regression models for analyzing data allowing for a relevant extension of distributions assumed for the response variable even in the presence of truncated or censored data. Given their flexibility and thriftiness, they can represent a valuable tool for solving actuarial problems even in the presence of big data.

CO346 Room BH (SE) 2.10 FINANCIAL ECONOMETRICS: MODELLING AND FORECASTING Chair: Vincenzo Candila

C0754: **Mixed-frequency quantile regressions to forecast value-at-risk and expected shortfall**
Presenter: **Vincenzo Candila**, University of Salerno, Italy
Co-authors: Lea Petrella, Giampiero Gallo

The use of quantile regression to calculate risk measures has been widely recognized in the financial econometrics literature. When data are observed at mixed-frequency, the standard quantile regression models are no longer adequate. We develop a model built on a mixed-frequency quantile regression to directly estimate the Value-at-Risk (VaR) and the Expected Shortfall (ES) measures. In particular, the low-frequency component incorporates information coming from variables observed at, typically, monthly or lower frequencies, while the high-frequency component can include a variety of daily variables, like market indices or realized volatility measures. The validity of the proposed model is then explored through a real data application using two energy commodities, that is, Crude Oil and Gasoline futures. We show that our model outperforms other competing specifications, using the most popular VaR and ES backtesting procedures.

C0942: **Sentiment analysis and NFT transaction dynamics**
Presenter: **Giorgia Rivieccio**, Parthenope University, Italy
Co-authors: Giovanni De Luca

A stunning paradigm of a benefit of technology is blockchain-based cryptocurrency art. In a world of digital art where everything can be freely copied and saved with a right-click, blockchain technology allowed for the creation of scarcity, promoting the expansion, bursting, and stability of the tumultuous market of artists, collectors, galleries, and curators. In this context, human coordination across Web3's financial, regulatory, and social norms is changing due to emerging market Non-Fungible Token (NFT). Despite their youth, NFTs have a $50 billion market worth and are quickly becoming crucial components of ownership in the digital sphere. NFT markets are volatile and hard to speculate on, much like conventional equities markets. We intend to develop a novel model to analyze the transaction dynamics, including market sentiment. We have collected data about this last year from the platform https://nonfungible.com to create a time-series of collectible (artwork) NFT transactions. We have then analyzed the univariate dynamics and studied the improvements in forecasting after including the co-movements with respect to cryptocurrency data, google trends data about NFT, and the sentiment scores from news extracted by Thomson Reuters concerning NFT textual data.

C1139: **Volatility and liquidity nexus in cryptocurrency markets**
Presenter: **Angelo Forino**, Sapienza University, Italy
Co-authors: Giacomo Morelli

The high fragmentation and weak regulation of cryptocurrency markets make the assessment of their liquidity and volatility difficult, especially at the high frequency. The nexus between liquidity and volatility in cryptocurrency markets is investigated. We propose the logMEM-MIDAS model, which accommodates the LogMEM model within the MIDAS context. This approach overcomes the shortcomings of traditional volatility models, which are inadequate to capture the impact of market liquidity on the volatility. In particular, the information about market liquidity is derived from easy-to-access transaction data through easy-to-compute estimators. We apply the model to real data considering two different cryptocurrency markets. Our findings show strong positive nexus between volatility and liquidity, which supports the perception of cryptocurrencies as speculative investments.

C1323: **Hierarchical Bayesian fuzzy clustering approach for high dimensional linear time-series**
Presenter: **Antonio Pacifico**, University of Macerata, Italy

A computational approach is developed to improve fuzzy clustering and forecasting performance when dealing with endogeneity issues and misspecified dynamics in high-dimensional dynamic data. Hierarchical Bayesian methods are used to structure linear time variations, reduce dimensionality, and compute a distance function capturing the most probable set of clusters among univariate and multivariate time series. Nonlinearities

involved in the procedure look like permanent shifts and are replaced by coefficient changes. Monte Carlo implementations are also addressed to compute exact posterior probabilities for each cluster chosen and then minimize the increasing probability of outliers plaguing traditional clustering time-series techniques. An empirical example highlights the strengths and limitations of the estimating procedure. Discussions with related works are also considered.

Monday 19.12.2022 14:40 - 16:20 Parallel Session P – CFE-CMStatistics

EO689 Room S-2.23 ECONOMETRICS: NEW DIRECTIONS Chair: Taoufik Bouezmarni

E1702: **On testing for independence between the generalized errors of several time series**
Presenter: **Bruno Remillard**, HEC Montreal, Canada
Test statistics are proposed for checking the independence between the generalized errors of several univariate time series models. These models include volatility models as well as regime-switching models. In order to obtain consistent tests, the statistics are constructed from lagged empirical processes whose asymptotic distribution is studied. Examples of applications from financial time series are given.

E1703: **Spatiotemporal Markov regime-switching models based on copulas**
Presenter: **Bouchra Nasri**, University of Montreal, Canada
The aim is to present copula-based Spatiotemporal Markov regime-switching models with covariates when the variables of interest are continuous, discrete, or zero-inflated. A simulation study to assess the performance of the models is presented.

E1723: **Copula-based multivariate expectile regression**
Presenter: **Karim Oualkacha**, UQAM, Canada
Co-authors: Aziz Lmoudden, Taoufik Bouezmarni
Expectiles summarize a distribution in a manner similar to quantiles. Expectile regression has recently gained great popularity, in part due to its attractive statistical and computational properties. Unfortunately, despite the renewed interest, it remains limited to single-output problems. To enhance this and gain insight into multivariate data, we build on a class of multivariate expectile loss functions to develop a unified and flexible copula-based multivariate expectile regression framework. Our approach provides a new class of multiple-output expectile regression estimators, which are unique solutions to convex risk minimization problems. We model the joint distribution of the multiple-output and the regressors using a copula model, which separates modelling the dependence and the marginal distributions. Then, we rewrite the multivariate expectile regression loss function in terms of the copula and the marginal distributions. We prove the asymptotic properties of our estimators (weak convergence and i.i.d. representation). We demonstrate the effectiveness of our approach through simulation studies and by analyzing the Fourth Dutch Growth data.

E1729: **Change-point tests and estimators for gradually changing dependence structures based on Kendalls tau**
Presenter: **Felix Camirand Lemyre**, Universita de Sherbrooke, Canada
Co-authors: Jean-Francois Quessy
Suppose that pairs $(X_1,Y_1),\ldots,(X_n,Y_n)$ of independent observations are subject to a gradual change in their stochastic behaviour in the sense that for given $K_1 < K_2$, between 1 and n, the underlying joint distribution of a given pair is F before K_1, G after $K_2 > K_1$, and gradually moving from F to G between the two times of change K_1 and K_2. This setup elegantly generalizes the usual abrupt change model, which is usually assumed in the change-point analysis. Under this configuration, asymptotically unbiased estimators of Kendall's tau up to the change and after the change are derived, as well as tests and estimators of gradual change points related to these measures of association. The asymptotic behaviour of the introduced estimators and test statistics as n goes to infinity is rigorously investigated, in particular by demonstrating a general result (of independent interest) concerning weighted indexed U-statistics computed under heterogeneous data. The sampling properties of the proposed estimators of the Kendall taus, tests for change-point detection and estimator of times of change are studied in a simulation study that considers various scenarios of gradual changes in dependence. The usefulness of the introduced tools is illustrated on a multivariate time series of monthly consumer price index in the United States.

EO184 Room S-1.04 MACHINE LEARNING IN THE BEHAVIORAL SCIENCES Chair: Andreas Alfons

E1122: **What can statistics do for Open Science and what can Open Science do for statistics?**
Presenter: **Sabine Hoffmann**, Ludwig-Maximilians-Universitaet Muenchen, Germany
In recent years, the reliability of scientific findings has been investigated through replication and multi-analyst studies, in which multiple teams of researchers are asked to answer the same research question on the same data set. These two types of studies provide increasing evidence that classical statistical methods which only focus on sampling uncertainty convey a disproportionate level of certainty and thereby yield overconfident results, leading to what has been referred to as replication or statistical crisis in science. While these issues have fuelled debate and received considerable attention both in the scientific community and beyond, the involvement of statisticians in finding causes and solutions to this crisis has been surprisingly limited. We will give an overview of ways in which statisticians can contribute to methodological challenges in the Open Science movement and ways in which the Open Science movement can help tackle longstanding methodological challenges. In particular, we will give an overview of ideas on how we can make evidence from different study designs and different analysis strategies comparable, how we can derive uncertainty intervals that account for sampling uncertainty and analytical variability and how we can assess the extent of selective reporting in methodological research.

E1419: **Identifying periods of careless responding in surveys: A deep learning approach**
Presenter: **Max Welz**, Erasmus University Rotterdam, Netherlands
Co-authors: Andreas Alfons
Rating-scale datasets collected from surveys are paramount to empirical research. However, some respondents may not comply with survey instructions due to deficiencies in survey design or lack of motivation. This phenomenon is known as careless responding (CR). CR is a major threat to internal validity and should therefore be screened for. Existing methods for detecting CR are designed to identify respondents who respond carelessly throughout the survey. However, recent work suggests that the longer a survey takes, the higher the likelihood that a large proportion of all respondents will eventually start responding carelessly. Thus, we are interested in identifying when a respondent becomes careless (if at all) rather than trying to detect respondents who respond carelessly throughout the survey. Correspondingly, we propose a novel method for identifying the periods of carelessness (or a lack thereof) of each respondent. The proposed method uses the deep learning technique of autoencoders in combination with response times. By means of extensive numerical experiments, we find that our proposed method achieves high reliability in correctly identifying periods of careless responding and discriminates well between careless and regular respondents. Our method seems to perform particularly well in long surveys, which are common in psychology and health sciences, where it is likely that a large proportion of all respondents eventually responds carelessly due to fatigue.

E1111: **Modelling customer journeys with transformers**
Presenter: **Luuk van Maasakkers**, Erasmus University Rotterdam, Netherlands
Co-authors: Dennis Fok, Bas Donkers
A customer journey is a sequence of customer- and firm-initiated actions that may or may not end in a particular target event, such as a purchase being made by the customer. Firms collect data on these customer journeys by tracking online clicks and monitoring other customer behavior. This data can be used to learn how firm-initiated actions affect customer behavior, which enables a firm to optimize its marketing efforts. We propose a neural network-based method to model the dynamics in individual customer journeys over time. More specifically, we model the next action of a customer as a function of all prior touchpoints, both firm- and customer-initiated. Not only the type of the next touchpoint is modelled, but also

the time until the next customer-initiated action. While much of the previous literature has modeled the sequence of all touchpoints, both firm- and customer-initiated, our model solely aims at predicting customer-initiated touchpoints; firm-initiated touchpoints only serve as explanatory input. In this way, we can estimate the effects of the type and timing of firm-initiated touchpoints on the customer's behavior. This enables us to optimize the firm actions, as we can simulate how customers would respond to new firm policies.

E0640: **Learning with subset stacking**
Presenter: **Hakan Akyuz**, Erasmus University Rotterdam, Netherlands
Co-authors: Ilker Birbil, Sinan YILDIRIM, Kaya Gokalp

A new regression algorithm is proposed that learns from a set of input-output pairs. Our algorithm is designed for populations where the relation between the input variables and the output variable exhibits a heterogeneous behavior across the predictor space. The algorithm starts with generating subsets that are concentrated around random points in the input space. This is followed by training a local predictor for each subset. Those predictors are then combined in a novel way to yield an overall predictor. We call this algorithm "LEarning with Subset Stacking" or LESS, due to its resemblance to the method of stacking regressors. We compare the testing performance of LESS with state-of-the-art methods on several datasets. Our comparison shows that LESS is a competitive supervised learning method. Moreover, we observe that LESS is also efficient in terms of computation time and it allows a straightforward parallel implementation.

EO492 Room S-1.06 RECENT ADVANCES IN REINFORCEMENT LEARNING (VIRTUAL) Chair: Jiayi Wang

E0231: **Bayesian basket of bandits**
Presenter: **Eric Laber**, Duke University, United States

Contextual bandit models are a primary tool for sequential decision-making with applications ranging from clinical trials to e-commerce. While there are multiple bandit algorithms which achieve optimal regret and show strong performance on benchmark problems, algorithm selection and tuning in any given application remains a major open problem. We propose the Bayesian Basket of Bandits (B3), a meta-learning algorithm which automatically ensembles a set (basket) of candidate algorithms and tuning procedures to produce a learning algorithm which dominates all algorithms in the basket. The method works by treating the evolution of a bandit algorithm as a Markov decision process in which the states are posterior distributions over model parameters and subsequently applying approximate Bayesian dynamic programming to learn an optimal ensemble. We derive both Bayesian and frequentist convergence results for the cumulative discounted utility. In simulation experiments, our proposed method provides lower regret than state-of-the-art algorithms, including Thompson Sampling, Upper Confidence Bound methods, and information-directed sampling.

E0720: **Regret bounds for information-directed reinforcement learning**
Presenter: **Botao Hao**, Deepmind, United States

Information-directed sampling (IDS) has revealed its potential as a data-efficient algorithm for reinforcement learning (RL). However, the theoretical understanding of IDS for Markov Decision Processes (MDPs) is still limited. We develop novel information-theoretic tools to bound the information ratio and cumulative information gain about the learning target. Our theoretical results shed light on the importance of choosing the learning target such that the practitioners can balance the computation and regret bounds. As a consequence, we derive prior-free Bayesian regret bounds for vanilla-IDS which learns the whole environment under tabular finite-horizon MDPs. In addition, we propose a computationally efficient regularized IDS that maximizes an additive form rather than the ratio form and show that it enjoys the same regret bound as vanilla-IDS. With the aid of rate-distortion theory, we improve the regret bound by learning a surrogate, less informative environment. Furthermore, we extend our analysis to linear MDPs and prove similar regret bounds for Thompson sampling as a by-product.

E0840: **Projected state-action balancing weights for offline reinforcement learning**
Presenter: **Jiayi Wang**, The University of Texas at Dallas, United States
Co-authors: Zhengling Qi, Raymond Ka Wai Wong

Offline policy evaluation (OPE) is considered a fundamental and challenging problem in reinforcement learning (RL). The focus is on the value estimation of a target policy based on pre-collected data generated from a possibly different policy, under the framework of infinite-horizon Markov decision processes. Motivated by the recently developed marginal importance sampling method in RL and the covariate balancing idea in causal inference, we propose a novel estimator with approximately projected state-action balancing weights for the policy value estimation. We obtain the convergence rate of these weights, and show that the proposed value estimator is semi-parametric efficient under technical conditions. In terms of asymptotics, our results scale with both the number of trajectories and the number of decision points at each trajectory. As such, consistency can still be achieved with a limited number of subjects when the number of decision points diverges. In addition, we develop a necessary and sufficient condition for establishing the well-posedness of Bellman operator in the off-policy setting, which characterizes the difficulty of OPE and may be of independent interest. Numerical experiments demonstrate the promising performance of our proposed estimator.

E0956: **Pessimism in the face of confounders: Provably efficient offline RL in partially observable Markov decision processes**
Presenter: **Zhuoran Yang**, Yale University, United States

Offline reinforcement learning (RL) is studied in partially observable Markov decision processes. In particular, we aim to learn an optimal policy from a dataset collected by a behavior policy which possibly depends on the latent state. Such a dataset is confounded in the sense that the latent state simultaneously affects the action and the observation, which is prohibitive for existing offline RL algorithms. To this end, we propose the Proxy variable Pessimistic Policy Optimization (`P3O`) algorithm, which addresses the confounding bias and the distributional shift between the optimal and behavior policies in the context of general function approximation. At the core of `P3O` is a coupled sequence of pessimistic confidence regions constructed via proximal causal inference, which is formulated as minimax estimation. Under a partial coverage assumption on the confounded dataset, we prove that `P3O` achieves an $n^{-1/2}$ suboptimality, where n is the number of trajectories in the dataset. To our best knowledge, `P3O` is the first provably efficient offline RL algorithm for POMDPs with a confounded dataset.

EO673 Room S-1.27 COMPLEX JOINT AND MULTIVARIATE MODELS WITH MEDICAL APPLICATIONS Chair: Michael Daniels

E0282: **Flexible evaluation of surrogacy in Bayesian adaptive platform studies**
Presenter: **Michael Sachs**, University of Copenhagen, Denmark
Co-authors: Erin Gabriel, Alessio Crippa, Michael Daniels

Trial-level surrogates are useful tools for improving the speed and cost-effectiveness of trials, but surrogates that have not been properly evaluated can cause misleading results. The evaluation procedure is often contextual and depends on the type of trial setting. As Bayesian adaptive platform studies are becoming more popular, methods for surrogate evaluation using them are needed. These studies also offer a rich data resource for surrogate evaluation that would not normally be possible. They also offer a set of statistical issues, including heterogeneity of the study population, treatments, implementation, and even potentially the quality of the surrogate. We propose the use of a hierarchical Bayesian semiparametric model for the evaluation of potential surrogates using nonparametric priors for the distribution of true effects based on Dirichlet process mixtures. The motivation for this approach is to flexibly model relationships between the treatment effect on the surrogate and the treatment effect on the outcome and also to identify potential clusters with differential surrogate value in a data-driven manner. We demonstrate how our method can be used in a simulated illustrative example based on an ongoing platform study in prostate cancer.

E0412: **Multivariate shared parameter mixed-effects location scale (MELS) models for intensive longitudinal data**
Presenter: **Donald Hedeker**, University of Chicago, United States

Intensive longitudinal data using Ecological Momentary Assessment (EMA) are obtained to study subjective experiences within changing environmental contexts. The intensive longitudinal data allow one to characterize a subject's mean and variance and specify models for both. We focus on an EMA study of dual users (i.e., both combustible and electronic cigarette users), where interest is on characterizing mood associated with these nicotine products, and whether the subject's mood response relates to future nicotine product use. For this, the MELS model of the EMA data includes random subject effects for the mean (i.e., location), which characterize the subject's differential mood response to combustible and electronic cigarettes. A random effect for the subject's variability (i.e., scale) is also included to characterize the subject's mood consistency/erraticism. These random location and scale effects are then shared in a bivariate longitudinal mixed model of post-EMA nicotine product use (both combustible and electronic cigarettes).

E0733: **Bayesian feature selection in joint models with application to cardiovascular disease cohorts**
Presenter: **Mirajul Islam**, University of Florida, United States
Co-authors: Michael Daniels, Juned Siddique

Cardiovascular disease (CVD) cohorts collect data longitudinally to study the association between CVD event times and risk factors. An important area of scientific research is to understand better what features of CVD risk factor trajectories are associated with disease. We develop methods for feature selection in joint models where the features are viewed as a two-level variable selection problem with multiple features and multiple longitudinal processes. We modify a Bayesian sparse group selection prior for the joint modeling framework to select features both at the group level (CVD risk factor) and within a group (features of a longitudinal risk factor). We apply our methods to the Atherosclerosis Risk in Communities (ARIC) study data, a population-based and prospective cohort study consisting of 15,792 participants measured four times at three-year intervals, where it is important to investigate which characteristics of CVD risk factor trajectories are associated with death from CVD.

E0885: **Describing complex disease progression with latent class models for multivariate longitudinal markers and times-to-event**
Presenter: **Cecile Proust-Lima**, INSERM, France
Co-authors: Tiphaine Saulnier, Viviane Philipps, Alexandra Foubert-Samier

Some diseases are characterized by numerous markers of progression. Although not specific, this is mainly the case in neurodegenerative diseases where pathological brain changes may induce multiple clinical signs on which the progression of a patient is assessed. For instance, Multiple System Atrophy (MSA), a rare neurodegenerative synucleinopathy, is characterized by various combinations of progressive autonomic failure and motor dysfunction, and by a very poor prognosis with a median survival of a few years after diagnosis. Describing the progression of such complex and multi-dimensional diseases is particularly difficult. One has to simultaneously account for the assessment of multivariate markers over time, the occurrence of clinical endpoints, and the highly suspected heterogeneity between patients partly due to the difficulty of formally diagnosing the disease. Yet, such description is crucial for understanding the natural history of the disease, staging patients diagnosed with the disease, unraveling subphenotypes, and predicting the prognosis. Through the example of MSA progression, we show how a latent class approach, implemented in the R package lcmm, can help describe complex disease progression measured by multiple repeated markers and clinical endpoints, and identify subphenotypes for exploring new pathological hypotheses.

EO562 Room K0.16 NETWORK AND HIGH DIMENSIONAL DATA ANALYSIS Chair: Jing Lei

E0976: **Population-level balance in signed networks**
Presenter: **Weijing Tang**, Harvard University, United States

Statistical network models are useful for understanding the underlying formation mechanism and characteristics of complex networks. However, statistical models for signed networks have been largely unexplored. In signed networks, there exist both positive (e.g., like, trust) and negative (e.g., dislike, distrust) edges, which are commonly seen in real-world scenarios. The positive and negative edges in signed networks lead to unique structural patterns, which pose challenges for statistical modeling. We introduce a statistically principled latent space approach for modeling signed networks and accommodating the well-known balance theory, i.e., "the enemy of my enemy is my friend" and "the friend of my friend is my friend". The proposed approach treats both edges and their signs as random variables, and characterizes the balance theory with a novel and natural notion of population-level balance. This approach guides us towards building a class of balanced inner-product models, and towards developing scalable algorithms via projected gradient descent to estimate the latent variables. We also establish non-asymptotic error rates for the estimates, which are further verified through simulation studies. In addition, we apply the proposed approach to an international relation network, which provides an informative and interpretable model-based visualization of countries during World War II.

E1188: **Linear regression and its inference on noisy network-linked data**
Presenter: **Tianxi Li**, University of Virginia, United States
Co-authors: Can Minh Le

Linear regression on network-linked observations has been an essential tool in modeling the relationship between response and covariates with additional network structures. Previous methods either lack inference tools or rely on restrictive assumptions on social effects and usually assume that networks are observed without errors. A regression model with nonparametric network effects is proposed. The model does not assume that the relational data or network structure is exactly observed and can be provably robust to network perturbations. Asymptotic inference framework is established under a general requirement of the network observational errors, and the robustness of this method is studied in the specific setting when the errors come from random network models. We discover a phase-transition phenomenon of the inference validity concerning the network density when no prior knowledge of the network model is available while also showing a significant improvement achieved by knowing the network model. Simulation studies are conducted to verify these theoretical results and demonstrate the advantage of the proposed method over existing work in terms of accuracy and computational efficiency under different data-generating models. The method is then applied to middle school students' network data to study the effectiveness of educational workshops in reducing school conflicts.

E1189: **Conformal-based hypothesis testing using rank sum statistics**
Presenter: **Jing Lei**, Carnegie Mellon University, United States

The problem of testing the equality of the conditional distribution is considered for a response variable given a set of covariates between two populations. Such a testing problem is related to transfer learning and causal inference. We develop a nonparametric procedure by combining recent advances in conformal prediction with some new ingredients, such as a novel choice of conformity score and data-driven choices of weight and score functions. To our knowledge, this is the first successful attempt to use conformal prediction for testing statistical hypotheses beyond exchangeability. Our method is suitable for modern machine learning scenarios where the data has high dimensionality and large sample sizes, and can be effectively combined with existing classification algorithms to find good weight and score functions. The performance of the proposed method is demonstrated in synthetic and real data examples.

E1303: **Constructing local cell-specific networks from single-cell data**
Presenter: **Xuran Wang**, Icahn School of Medicine at Mount Sinai, United States
Co-authors: David Choi, Kathryn Roeder

Gene co-expression networks yield critical insights into biological processes, and single-cell RNA sequencing provides an opportunity to target

inquiries at the cellular level. However, the sparsity and heterogeneity of transcript count present challenges when constructing gene networks using traditional estimation techniques. These methods fail to detect complex co-expression patterns and obscure the heterogeneity across cell populations. We develop a method to estimate a cell-specific network (CSN) for every single cell that facilitates testing for differences in network structure between cell groups. Average CSNs provide stable estimates of network structure and detect gene block structure better than traditional measures. New downstream analysis methods using CSNs utilize more fully the information contained within them. We examined the evolution of gene networks in fetal brain cells and compared the CSNs of cells sampled from autism spectrum disorder and control subjects to reveal intriguing patterns in gene co-expression.

EO066 Room K0.18 RECENT ADVANCES IN BAYESIAN CAUSAL INFERENCE Chair: Erica Moodie

E0690: **Mediation analysis with external summary data on total effect**
Presenter: **Bhramar Mukherjee**, University of Michigan, United States

As modern assaying technologies continue to improve, environmental health studies are increasingly measuring endogenous omics data to study intermediary biological pathways of outcome-exposure associations. Mediation analysis is often carried out when there is well-established literature showing the statistical and practical significance of the association between an exogenous exposure and a health outcome of interest, or the total effect. For example, there are a plethora of studies associating maternal phthalate exposure with preterm delivery, and researchers are now trying to characterize the mechanisms by which phthalate exposure impacts final gestational age. The existing methodology for performing mediation analyses does not leverage the rich external information available on the total effect. We show that incorporating external summary-level information on the total effect improves the estimation efficiency of the natural direct and indirect effects, and is a function of the partial R-squared comparing the outcome model with and without the mediators. Additionally, we discuss how to handle incongruous external information which can arise from transportability violations or fundamentally different adjustment sets. The proposed framework blends mediation analysis with modern data integration techniques.

E0898: **RVine copula as virtual baseline generator in agent-based modeling in clinical research**
Presenter: **Nicolas Savy**, Toulouse Institute of Mathematics, France
Co-authors: Philippe Saint-Pierre

Simulation is today considered the third pillar of science, a peer alongside theory and experimentation. Indeed, simulation is a relevant means to analyze complex systems. In medical research, a wide range of questions can be investigated through simulation involving various simulation tools. In a series of articles, the authors have reflected on the use of agent-based models for clinical research. The common idea is to use the enormous amount of information available about the patients of interest, the disease of interest, the drug of interest and the trial design in order to build a stochastic model mimicking the course of clinical research. The strategy consists of identifying design weaknesses, measuring the performance of a trial within a predefined framework while reducing the number of logistical barriers. Two classes of models have been identified: first, models to generate baseline values for a patient cohort and second, models to define the evolution of the outcome(s) of interest. The first class of models are called virtual baseline generators and the second class execution models. Our focus here is on the first model class. Virtual Baseline Generators essentially consist of Monte-Carlo generators of covariate vectors. We will discuss the relevance of the RVine copula as a VBG. These models are attractive because they are able to capture the correlation structure of a training dataset and generate data based on that structure.

E1271: **Bayesian sensitivity analysis**
Presenter: **Aad van der Vaart**, TU Delft, Netherlands

Some Bayesian methods are reviewed for sensitivity analysis for deviations from the assumptions in causal inference or missing data problems. We discuss a particular method based on nonparametric modelling.

E1538: **Bayesian inference for functional parameters**
Presenter: **David Stephens**, McGill University, Canada
Co-authors: Vivian Meng

Bayesian semiparametric inference targets finite-dimensional parameters that are functionals of the (unknown) data-generating distribution, which, to avoid inconsistency due to misspecification, should be modelled in a non-parametric fashion. We will introduce a general framework for performing inference in this setting that ensures compatibility between the inherent finite and infinite-dimensional specifications that comprise the inference. We will illustrate the methodology with examples from causal inference, including regression-based adjustment and inverse probability weighting.

EO706 Room K0.19 RECENT ADVANCES IN CAUSAL INFERENCE Chair: Peng Ding

E0224: **Distributional robustness, replicability, and causality**
Presenter: **Dominik Rothenhaeusler**, Stanford University, United States
Co-authors: Yujin Jeong

How can we draw trustworthy scientific conclusions? One criterion is that a study can be replicated by independent teams. While replication is critically important, it is arguably insufficient. If a study is biased for some reason and other studies recapitulate the approach, then findings might be consistently incorrect. It has been argued that trustworthy scientific conclusions require disparate sources of evidence. However, different methods might have shared biases, making it difficult to judge the trustworthiness of a result. We formalize this issue by introducing a "distributional uncertainty model", which captures biases in the data collection process. Distributional uncertainty is related to other concepts in causal inference, such as confounding and selection bias. We show that a stability analysis on a single data set allows for the construction of confidence intervals that account for both sampling uncertainty and distributional uncertainty. The proposed method is inspired by a stability analysis that is advocated for by many researchers in causal inference.

E0747: **No star is good news: A unified look at rerandomization based onp-values from covariate balance tests**
Presenter: **Anqi Zhao**, National University of Singapore, Singapore
Co-authors: Peng Ding

Randomized experiments balance all covariates on average and provide the gold standard for estimating treatment effects. Chance imbalances nevertheless exist more or less in realized treatment allocations, subjecting subsequent inference to possibly large variability and conditional bias. Modern social and biomedical scientic publications often require the reporting of covariate balance tables with not only covariate means by treatment group but also the associated p-values from signicance tests of their dierences. The practical need to avoid small p-values renders balance check and rerandomization by hypothesis testing standards an attractive tool for improving covariate balance in randomized experiments. Despite the intuitiveness of such practice and its possibly already widespread use in reality, the existing literature knows little about its implications on subsequent inference, subjecting many effectively rerandomized experiments to possibly inecient analyses. To fill this gap, we examine a variety of potentially useful schemes for rerandomization based on p-values (ReP) from covariate balance tests, and quantify their impact on subsequent inference. Specically, we focus on three estimators of the average treatment effect from the unadjusted, additive, and fully interacted linear regressions of the outcome on treatment, respectively, and derive their respective asymptotic sampling properties under ReP.

E0829: **Design-based anytime-valid causal inference**
Presenter: **Iavor Bojinov**, Harvard Business School, United States
Co-authors: Dae Woong Ham
Many organizations run online randomized experiments to drive product innovation and augment decision-making. Traditionally, these have a fixed time horizon during which experimental units (the customers) receive either a treatment (the new version) or a control (the standard offering). At the end of the experiment, an analyst determines the effectiveness of the treatment relative to the control by, for example, computing the average treatment effect (ATE) and an associated confidence interval. However, as customers arrive sequentially over time, partial results are available throughout the study; unfortunately, peaking at these data invalidates the subsequent statistical inference. To overcome this, companies have started using methods to compute confidence sequences, sequences of confidence intervals that are uniformly valid over time and allow analysts to monitor the ATE continuously. We develop design-based any-time valid asymptotic confidence sequences for three settings. First, we consider subjects arriving independently so that time indexes individual units. Second, we consider running a time series experiment in which the same unit receives multiple treatments over time. Third, we consider panel experiments in which multiple units receive multiple treatments over time. Across the three settings, we allow the treatment assignment to dynamically update based on the observed historical data. Our work is partially motivated by a collaboration with Netflix.

E0794: **Horizontal and vertical regressions: From duels to duals**
Presenter: **Dennis Shen**, University of California, Berkeley, United States
Co-authors: Peng Ding, Jasjeet Sekhon, Bin Yu
A central goal in social science is to evaluate the causal effect of a policy. In this pursuit, researchers often organize their observations in a panel data format where a subset of units are exposed to a policy for some time periods while the other units are unaffected. The information across time and space motivates two general approaches: (i) horizontal regression (i.e., unconfounedness), which exploits time series patterns, and (ii) vertical regression (e.g., synthetic controls), which exploits cross-sectional patterns. Although conventional wisdom states that the two are fundamentally different, we prove that they yield numerically identical estimates under several standard settings. Within this regime, we study properties of the estimator from three model-based perspectives—horizontal, vertical, and their mixture—and construct corresponding confidence intervals that offer new approaches to inference. Our results highlight how the choice of randomness relates to the choice of estimand. We show that these insights carry over to the design-based framework as well.

EO674 Room K0.20 MULTIVARIATE METHODS: JOINT DIAGONALIZATION AND PROJECTION PURSUIT Chair: Aurore Archimbaud

E1068: **Subspace independent component analysis: Finding clustering structures in a low dimensional space**
Presenter: **Jeffrey Durieux**, Erasmus University, Netherlands
K-means is an often-used clustering method. However, applying K-means to a data set may fail to uncover clusters due to presence of masking variables and the curse of dimensionality. A commonly used workaround is to apply PCA to the data prior to performing cluster analysis, a practice called Tandem Analysis (TA). A vulnerability of TA is that PCA does not guarantee to preserve the cluster structure present in the original data, jeopardizing the usefulness of subsequent cluster analysis. Multiple authors have provided procedures that reduce the dimensionality of a data set and perform cluster analysis on the reduced data, all aiming to find suitable low-dimensional representations of data while also keeping cluster structures intact. We present a novel approach to reducing dimensionality and performing cluster analysis on the low dimensional representation of the data called Subspace Independent Component Analysis (SICA). The method is described and thoroughly tested in systematically manipulated simulation studies where we compare it to related methods. Results show that SICA is a fast procedure that extracts components from the data that preserve cluster structures, but that performance depends on characteristics of the data. In addition, the correctness of the clusterings obtained through SICA is high, although it does not always outperform currently available methods.

E1302: **Invariant coordinate selection as preprocessing for clustering**
Presenter: **Aurore Archimbaud**, Erasmus University Rotterdam, Netherlands
Co-authors: Andreas Alfons, Klaus Nordhausen, Anne Ruiz-Gazen
Dimension reduction is an important preprocessing step in the multivariate analysis field, likely improving the identification of clusters. The well-known Principal Component Analysis (PCA), is one of the most famous dimension reduction techniques, but it may not be the best choice for clustering purposes. An alternative approach, Invariant Component Selection (ICS), relies on the simultaneous diagonalization of two scatter matrices. It goes beyond PCA by finding directions of interest through the optimization of general kurtosis measures and returns affine invariant components. Two challenging steps are the choice of the pair of scatter matrices and the selection of the components to retain. Some theoretical results have already been derived that guarantee that under some elliptical mixture models, the structure of the data can be highlighted on a subset of the first and/or last components. ICS has received little attention concerning clustering tasks. We evaluate the performance of several well-known clustering algorithms with ICS as a preprocessing step. We consider different combinations of scatter matrices, components selection approaches and the impact of outliers, on some simulations and some benchmark data sets.

E0678: **Refining invariant coordinate selection via local projecton pursuit**
Presenter: **Lutz Duembgen**, University of Bern, Switzerland
Invariant coordinate selection (ICS) is a powerful tool for finding potentially interesting projections of multivariate data. In some cases, some of the projections proposed by ICS come close to really interesting ones, but little deviations can result in a blurred view which does not reveal the feature (e.g. a clustering) which would otherwise be clearly visible. To remedy this problem, we propose an automated and localized version of projection pursuit (PP). Precisely, our local search is based on gradient descent applied to estimated differential entropy as a function of the projection matrix. We illustrate the method with two- and three-dimensional projections.

E1108: **Blind recovery of sources for multivariate space-time random fields**
Presenter: **Christoph Muehlmann**, Technical University of Vienna, Austria
Co-authors: Sandra De Iaco, Klaus Nordhausen
With advances in modern world technology, huge datasets that show dependencies in space as well as in time frequently occur in practice. As an example, several monitoring stations at different geographical locations track hourly concentration measurements of a number of air pollutants for several years. Such a dataset contains thousands of multivariate observations. Thus, proper statistical analysis needs to account for dependencies in space and time between and among the different monitored variables. To simplify the consequent multivariate spatio-temporal statistical analysis, it might be desirable to find linear transformations of the original data that result in easily interpretative, spatio-temporally uncorrelated processes that are also highly likely to have real physical meaning. Blind source separation (BSS) is a statistical methodology that is concerned with finding so-called latent processes that exactly meet the former requirements. BSS was already successfully used for sole temporal and sole spatial data, but, it was not yet introduced for the spatio-temporal case. BSS is reviewed and a generalization of BSS for second-order stationary multivariate spatio-temporal random fields (stBSS) is proposed. Two novel estimators (stAMUSE and stSOBI) which solve the formulated problem are also provided.

EO652 Room K0.50 DESIGN AND ANALYSIS OF COMPLEX EXPERIMENTS: THEORY AND APPLICATIONS Chair: MingHung Kao

E1594: **Differentiating various patient groups across parkinsonism spectrum via a new dantzig-selector-type screener**
Presenter: **Frederick Kin Hing Phoa**, Academia Sinica, Taiwan
Co-authors: Jing-Wen Huang, Yan-Han Lin, Shau Ping Lin, Yi-Tzang Tsai, Ming-Che Kuo, Koji Ueda, Ruey-Meei Wu

A robust, efficient and easily accessible differential diagnosis (D/D) scheme of patient groups from the parkinsonism spectrum is presented. The scheme includes a newly proposed Biomedical Oriented Logistic Dantzig Selector (BOLD Selector) for identifying robust biomarkers for D/D from the main effect model of a supersaturated experiment with binary response. To test the scheme's robustness via a published lipidomic dataset, double cross-validation is implemented to assess the generalizability of the model and single out suitable tuning parameters of the BOLD Selector. A newly generated proteomic dataset is investigated by profiling plasma EV proteins by LC/MS-MS from Parkinson's disease without dementia (PDND), PD with mild cognitive impairment (PD-MCI), PD with dementia (PDD), multiple system atrophy (MSA) and healthy controls (HC), using a multi-stage binary tree that employs a BOLD Selector at each stage to derive the prediction formulas for D/D of various patient groups. Biomarker candidates are engaged in the lipid metabolic pathway relevant to alpha-synucleinopathy, the pathological hallmark for both PD and MSA, indicating the promise of the BOLD Selector. Not only can it identify robust biomarkers with pathophysiological significance, thus facilitating D/D, but it can also pave the way towards identifying disease-relevant targets.

E0943: **Fast approximation of the Shapley values based on order-of-addition experimental designs**
Presenter: **Wei Zheng**, University of Tennessee, United States
Co-authors: Liuqing Yang, Yongdao Zhou, Haoda Fu, Min-Qian Liu

Shapley value is originally a concept in econometrics to fairly distribute both gains and costs to players in a coalition game. In recent decades, its application has been extended to other areas, such as marketing, engineering and machine learning. However, its heavy computational burden has been long recognized but rarely investigated. Specifically, in a d-player coalition game, calculating a Shapley value requires the evaluation of $d!$ or 2^d marginal contribution values, depending on whether we are taking the permutation or combination formulation of the Shapley value. Hence it becomes infeasible to calculate the Shapley value when d is reasonably large. A common remedy is to take a random sample of the permutations to surrogate for the complete list of permutations. We find an advanced sampling scheme can be designed to yield a much more accurate estimation of the Shapley value than simple random sampling (SRS). Our sampling scheme is based on combinatorial structures in the field of design of experiment (DOE), particularly the order-of-addition experimental designs for the study of how the orderings of the components would affect the output. We show that the obtained estimates are unbiased and consistent, sometimes even deterministically recover the original Shapley value. Both theoretical and simulation results show that our DOE-based sampling scheme outperforms SRS in terms of estimation accuracy.

E0994: **Experimental designs for functional data analysis**
Presenter: **MingHung Kao**, Arizona State University, United States

Functional data analysis (FDA) has gained much popularity in extracting useful information from repeated measurements collected at various points in a domain, such as time. A crucial step for rendering a precise and valid inference is to have a high-quality sampling schedule to sample informative data from the underlying random function. We are concerned with this design problem for FDA, and propose efficient computational approaches for obtaining good designs to rein in cost. Our proposed approach generates high-quality designs to allow a precise recovery of the underlying function, as well as precise prediction with functional regressions.

E1011: **Pilot study designs for sparse functional data**
Presenter: **Ping-Han Huang**, Arizona State University, United States
Co-authors: MingHung Kao

In sparse functional data analysis (SFDA), the number of repeated measures per subject is often limited by practical constraints such as costs. In light of this issue, a number of approaches have been developed to find (locally) optimal designs to help increase the efficiency of SFDA. The success of these design methods greatly hinges on the accurate prior information from pilot studies. However, the selection of a good pilot study design remains unclear. The aim is to fill this gap by proposing new hybrid designs that combine some combinatorial designs with a 'notorious' type of FDA designs. Through simulation studies, we demonstrate that our proposed designs can outperform the widely used simple random designs to facilitate the use of previously developed locally optimal design approaches.

EO715 Room S0.03 RECENT ADVANCES IN SPATIAL AND SPATIO-TEMPORAL STATISTICS (VIRTUAL) Chair: Thomas Neyens

E1107: **Bayesian spatial and spatiotemporal cluster Regression modeling**
Presenter: **Andrew Lawson**, Medical University of South Carolina, United States

Bayesian cluster modeling can be approached in a variety of ways. Explicit parametric cluster models can be assumed, or as an alternative, exceedance probabilities can be used, which are associated with an underlying standard risk model. These are often limited in their ability to link predictors to clustering effects directly. We will examine the use of mixture models whereby the spatial or spatio-temporal risk surface consists of a mixture of uncorrelated and correlated effects. The mixing probability is allowed to be spatially or spatio-temporally dependent, and is linked to a linear or non-linear predictor via a standard link function (logit, probit etc.). An example of cluster modeling for respiratory cancer incidence at a local spatial scale will be provided.

E1610: **Practical implementation of Hilbert space reduced-rank Bayesian Gaussian processes for spatial and temporal data**
Presenter: **Gabriel Riutort-Mayol**, Foundation for the Promotion of Health and Biomedical Research of Valencia Region (FISABIO), Spain
Co-authors: Paul-Christian Burkner, Michael Riis Andersen, Arno Solin, Aki Vehtari

Gaussian processes (GPs) are powerful non-parametric probabilistic models for stochastic functions, widely used for spatial and temporal data. However, the direct implementation entails a complexity that is computationally intractable when the number of observations is large, especially when estimated with fully Bayesian methods such as Markov chain Monte Carlo. A low-rank approximate Bayesian GP based on a basis function approximation via Laplace eigenfunctions for stationary covariance is implemented. The approach is simple and exhibits an attractive computational complexity due to its linear structure. However, the number of basis functions used in the approximation grows exponentially, and consequently, the computational requirements, with respect to the number of input dimensions. Practical guidelines on how to select the key factors of the method, such as the number of basis functions and the boundary factor, are provided. Furthermore, diagnostics for checking that the selected factors are adequate given the data to accurately fit the model are proposed. On that basis, an iterative procedure to achieve accurate approximation performance with minimal computational costs is also developed. Several illustrative examples of the performance and applicability of the method for simulated and real data in uni-dimensional (e.g. time-series data) and multi-dimensional (e.g. spatial and spatio-temporal data) cases are presented, using the probabilistic programming language Stan.

E1642: **Generalized spatial conditional overdispersion models: Applications and spatio-temporal extensions**
Presenter: **Mabel Morales Otero**, University of the Basque Country, Facultad de Economia y Empresa, Spain
Co-authors: Vicente Nunez-Anton

The purpose is to introduce and review the generalized spatial conditional overdispersion models, where regression structures are specified both for the conditional mean and for the overdispersion parameter structures. In these models, the possible spatial correlation in the data is modelled by

incorporating a spatial term in these regression structures by means of a parameter that directly estimates the intensity of the spatial association. We illustrate their performance for Poisson and binomially distributed responses by fitting them to the infant mortality rates and the mothers' postnatal screening period in Colombia datasets. Moreover, we also include their comparison with the widely used BYM and BYM2 models. In addition, we propose a direct spatio-temporal extension of the spatial conditional models, where we include the spatial lag of the response variable for each time unit in the linear predictor. Finally, we also propose the temporally varying spatial lag coefficient models, which allow the coefficient for the spatial term to vary with time. In order to illustrate their behavior, we apply our proposals, for Poisson distributed responses, to the respiratory hospital admissions in Glasgow data and, for binomially distributed responses, to the Georgia low birth weight data. In each case, we compare the performance of such models with the widely used Knorr-Helds models.

E1719: **Modelling species communities through space and time using opportunistic datasets**
Presenter: **Maxime Fajgenblat**, KU Leuven & UHasselt, Belgium
Co-authors: Robby Wijns, Luc De Meester, Thomas Neyens

Understanding how species communities are shaped by local and regional processes is one of the grand aims of ecology. Community datasets are, however, typically challenging to analyse due to their high dimensionality. Recent statistical advances increasingly facilitate the joint analysis of multispecies data and have given rise to several influential joint species distribution models (JSDMs). The rising popularity of JSDMs coincides with the increasing availability of crowdsourced biodiversity data through citizen science initiatives. Most JSDMs, however, cannot deal with the challenges intrinsic to these opportunistic data sources, such as imperfect and heterogeneous detection probabilities. We developed a spatio-temporal joint species distribution model that flexibly acknowledges imperfect detection. Specifically, we combined a spatio-temporal latent factor approach with a comprehensive site-occupancy approach to model occupancy and detection patterns across the considered species. We performed Bayesian inference through the probabilistic programming language Stan and applied the developed model to large datasets on invertebrate occurrences in Belgium. By doing so, we were able to gain insights at both the species and the community level. We argue that extending joint species distribution models to flexibly accommodate imperfect detection enables the study of species communities at an unprecedented scale due to their ability to harness a wider variety of datasets.

EO544 Room S0.11 RANDOM MATRICES IN STATISTICS AND ECONOMETRICS (VIRTUAL) Chair: Andrej Srakar

E0625: **Limit results for distributed estimation in invariant subspaces in multiple networks interference and PCA**
Presenter: **Runbing Zheng**, North Carolina State University, United States
Co-authors: Minh Tang

The problem of estimating the left and right singular subspaces is studied for a collection of heterogeneous random graphs with a shared common structure. We analyze an algorithm that first estimates the orthogonal projection matrices corresponding to these subspaces for each individual graph, then computes the average of the projection matrices, and finally finds the matrices whose columns are the eigenvectors corresponding to the d largest eigenvalues of the sample averages. We show that the algorithm yields an estimate of the left and right singular vectors whose row-wise fluctuations are normally distributed around the rows of the true singular vectors. We then consider a two-sample hypothesis test for the null hypothesis that two graphs have the same edge probabilities matrices against the alternative hypothesis that their edge probabilities matrices are different. Using the limiting distributions for the singular subspaces, we present a test statistic whose limiting distribution converges to a central chi-square (resp. non-central chi-square) under the null (resp. alternative) hypothesis. Finally, we adapt the theoretical analysis for multiple networks to the setting of distributed PCA; in particular, we derive normal approximations for the rows of the estimated eigenvectors using distributed PCA when the data exhibit a spiked covariance matrix structure.

E0705: **A uniform bound on the operator norm of sub-Gaussian random matrices and its applications**
Presenter: **Grigory Franguridi**, University of Southern California, United States
Co-authors: Roger Moon

For an $N \times T$ random matrix $X(\beta)$ with weakly dependent uniformly sub-Gaussian entries $x_{it}(\beta)$ that may depend on a possibly infinite-dimensional parameter $\beta \in \mathbf{B}$, we obtain a uniform bound on its operator norm of the form $E \sup_{\beta \in \mathbf{B}} ||X(\beta)|| \leq CK \left(\sqrt{\max(N,T)} + \gamma_2(\mathbf{B}, d_{\mathbf{B}}) \right)$, where C is an absolute constant, K controls the tail behavior of (the increments of) $x_{it}(\cdot)$, and $\gamma_2(\mathbf{B}, d_{\mathbf{B}})$ is Talagrand's functional, a measure of multi-scale complexity of the metric space $(\mathbf{B}, d_{\mathbf{B}})$. We illustrate how this result may be used for estimation that seeks to minimize the operator norm of moment conditions as well as for estimation of the maximal number of factors with functional data.

E0742: **Spectral universality in regularized linear regression with nearly deterministic design matrices**
Presenter: **Rishabh Dudeja**, Harvard University, United States
Co-authors: Subhabrata Sen, Yue M Lu

Statistical properties of many high-dimensional estimation tasks empirically exhibit universality with respect to the underlying design matrices. Specifically, matrices with very different constructions seem to behave identically if they share the same spectrum and have "generic" singular vectors. We prove this universality phenomenon for the performance of convex regularized least squares (RLS) estimators for linear regression. The contributions are two-fold: (1) We introduce a notion of universality classes for design matrices, defined through deterministic conditions that fix the spectrum of the matrix and formalize the heuristic notion of generic singular vectors; (2) We show that for all matrices in the same universality class, the dynamics of the proximal gradient algorithm for the regression problem, and the performance of RLS estimators themselves (under additional strong convexity conditions) are asymptotically identical. In addition to including i.i.d. Gaussian and rotational invariant matrices as special cases, our universality class also contains highly structured, strongly correlated, and even nearly deterministic matrices. Examples include randomly signed incoherent tight frames, and randomly subsampled Hadamard transforms. Due to this universality result, the performance of RLS estimators on many structured matrices with limited randomness can be characterized using the rotationally invariant ensemble as an equivalent yet mathematically tractable surrogate.

E1180: **Core shrinkage covariance estimation for matrix-variate data**
Presenter: **Peter Hoff**, Duke University, United States

A separable covariance model for a random matrix provides a parsimonious description of the covariances among the rows and the columns of the matrix, and permits likelihood-based inference with very small sample size. However, in many applications, the assumption of exact separability is unlikely to be met, and data analysis with a separable model may overlook or misrepresent important dependence patterns in the data. We propose a compromise between separable and unstructured covariance estimation. We show how the set of covariance matrices may be uniquely parametrized in terms of the set of separable covariance matrices and a complementary set of "core" covariance matrices, where the core of a separable covariance matrix is the identity matrix. This parametrization defines a Kronecker-core decomposition of a covariance matrix. By shrinking the core of the sample covariance matrix with an empirical Bayes procedure, we obtain an estimator that can adapt to the degree of separability of the population covariance matrix.

EO044 Room S0.12 STATISTICS IN NEUROSCIENCE I Chair: Russell Shinohara

E0854: **Accurate estimation of individual functional brain connectivity and topology via ICA with empirical population priors**
Presenter: **Amanda Mejia**, Indiana University, United States

Co-authors: Daniel Spencer, Ani Eloyan

A primary objective in resting-state fMRI studies is the localization of functional areas (i.e., resting-state networks) and the functional connectivity (FC) between them. These spatial and temporal properties of brain organization have been shown to be related to disease progression, development, and aging. Independent component analysis (ICA) is a popular tool to estimate functional areas and their FC. However, due to the high noise levels and short scan duration of typical fMRI data, subject-level ICA results tend to be noisy. Thus, group-level functional areas are often used in lieu of subject-specific ones, ignoring inter-subject variability in functional topology. These group-average maps also often form the basis for estimating FC, leading to potential bias in FC estimates given the topological differences in underlying functional areas. An alternative to these two extremes (noisy subject-level ICA and one-size-fits-all group ICA) is Bayesian hierarchical ICA, wherein information shared across subjects is leveraged to improve the subject-level estimation of spatial maps and FC. Functional connectivity template ICA is a computationally convenient hierarchical ICA framework using empirical population priors on the spatial configuration and connectivity between functional brain networks. These priors can be derived from large fMRI databases or holdout data. The proposed approach is validated through simulations and functional MRI data from the Human Connectome Project.

E0937: **Robust and reproducible group-level neuroimage analysis in R with the pbj package**

Presenter: **Simon Vandekar**, Vanderbilt University, United States

Co-authors: Kaidi Kang, Neil Woodward, Anna Huang, Maureen McHugo, Shawn Garbett, Jeremy Stephens, Russell Shinohara, Armin Schwartzman, Jeffrey Blume

Recent simulation studies have identified group-level neuroimaging statistical inference methods that make minimal assumptions about the data and consistently control error rates for cluster and voxel-wise inference. These include permutation and bootstrap procedures that can use sandwich covariance estimators to robustly account for spatial correlation in the images. Until now, there was no software available to implement these robust analyses within R. We present the pbj R package, a validated tool to perform group-level neuroimage analysis completely within R. pbj can be combined with tools available through Neuroconductor and CRAN to perform reproducible and interactive analyses. The theory and validation of the pbj software implementation are presented. A brief tutorial of neuroimage analysis using the pbj package is given.

E0978: **A time-varying AR, bivariate DLM of functional near-infrared spectroscopy data**

Presenter: **Timothy Johnson**, University of Michigan, United States

Functional near-infrared spectroscopy (fNIRS) is a relatively new neuroimaging technique. It is a low-cost, portable, and non-invasive method to measure brain activity via the blood oxygen level-dependent signal. Similar to fMRI, it measures changes in the level of blood oxygen in the brain. Its time resolution is much finer than fMRI; however, its spatial resolution is much courser—similar to EEG or MEG. fNIRS is finding widespread use on young children who cannot remain still in the MRI magnet, and it can be used in situations where fMRI is contraindicated—such as with patients who have cochlear implants. Furthermore, fNIRS measures the concentration of both oxygenated and deoxygenated hemoglobin, both of which may be of scientific interest. We propose a fully Bayesian time-varying autoregressive model to analyze fNIRS data within the multivariate DLM framework. The hemodynamic response function is modeled with the canonical HRF and the low-frequency drift with a variable B-spline model (both locations and number of knots are allowed to vary). Both the model error and the auto-regressive processes vary with time. Via simulation studies, we show that this model naturally handles motion artifacts and gives good statistical properties. The model is then applied to a fNIRS data set.

E1925: **Mass univariate relative risk regression for longitudinal binary-valued neuroimaging data**

Presenter: **Thomas Nichols**, University of Oxford, United Kingdom

Co-authors: Petya Kindalova, Michele Veldsman, Ioannis Kosmidis

There is growing interest in binary-valued brain images from MRI. Binary image data can identify the tissue damaged by a stroke, multiple sclerosis lesions in white matter, or bright spots called white matter hyperintensities. We recently proposed a mass univariate approach to modelling crossectional data that addresses the problem of low base rate with a penalised maximum likelihood approach with a probit or logistic model. We consider the additional challenge of longitudinal data. Users often want to interpret results as relative risks instead of odds-ratios, but a log-link with binomial variance function may lead to estimation instabilities when event probabilities are close to 1. To address these issues, we use generalized estimating equations with log-link regression structures with identity variance function and unknown dispersion parameter, with a penalty on the GEE of the gradient of the Jeffreys prior to avoid infinite parameter estimates. Our findings from extensive simulation studies show significant improvement over the standard log-link generalized estimating equations by providing finite estimates and achieving convergence when boundary estimates occur. The real data application on UK Biobank brain lesion maps further reveals the instabilities of the standard log-link generalized estimating equations for a large-scale data set and demonstrates the clear interpretation of relative risk in clinical applications.

EO711 Room S0.13 DYNAMIC RANDOM OBJECTS Chair: Wolfgang Polonik

E0290: **Intrinsic Riemannian functional data analysis**

Presenter: **Zhenhua Lin**, National University of Singapore, Singapore

Co-authors: Lingxuan Shao, Fang Yao

Riemannian functional data, in which functions take values in a nonlinear Riemannian manifold, pose new challenges to functional data analysis. To overcome these challenges, an intrinsic framework to analyze densely/sparsely observed Riemannian functional data is developed. The framework features the following innovative components: a frame-independent covariance operator/function, a smooth vector bundle termed covariance vector bundle, a parallel transport and a smooth bundle metric on the covariance vector bundle. The introduced intrinsic covariance function links the estimation of covariance structure to smoothing problems that involve raw covariance observations derived from sparsely observed Riemannian functional data, while the covariance vector bundle provides a mathematical foundation for formulating the smoothing problems. The parallel transport and the bundle metric together make it possible to measure the fidelity of fit to the covariance function. They also play a critical role in quantifying the quality of estimators for the covariance function. As an illustration, based on the proposed framework, a local linear smoothing estimator is developed for the covariance function.

E1307: **Coevolving latent space network with attractors models for polarization**

Presenter: **Eric Kolaczyk**, McGill University, Canada

A broadly applicable class of coevolving latent space network with attractors (CLSNA) models is developed, where nodes represent individual social actors assumed to lie in an unknown latent space, edges represent the presence of a specified interaction between actors, and attractors are added in the latent level to capture the notion of attractive and repulsive forces. We apply the CLSNA models to understand the dynamics of partisan polarization on social media, where we expect US Republicans and Democrats to increasingly interact with their own party and disengage with the opposing party. Using longitudinal social networks from the social media platforms Twitter and Reddit, we investigate the relative contributions of positive (attractive) and negative(repulsive) forces among political elites and the public, respectively. Our goals are to disentangle the positive and negative forces within and between parties and explore if and how they change over time. Our analysis confirms the existence of partisan polarization in social media interactions among both political elites and the public. Moreover, while positive partisanship is the driving force of interactions across the full periods of study for both the public and Democratic elites, negative partisanship has come to dominate Republican elites interactions since the run-up to the 2016 presidential election.

E1319: **Some recent advances on regression models in shape data**
Presenter: **Alfred Kume**, University of Kent, United Kingdom

Shape data are naturally represented as points in non-Euclidean spaces due to the natural process of their generation and the inherited invariances that need to be applied during the inferential process. For example, rotation invariance imposes a nonlinear constraint on the shape observations, which can typically be a collection of some landmark coordinates in 2 or 3-dimensional spaces. The general approaches with regression models applied to such data are listed. Among all the approaches developed so far, there has not been much focus on the generality of landmark correlations. This is due to the fact that a general covariance structure leads to identifiability issues with the respective parameters. Some advances to that end will be reported here and will be illustrated with some data examples.

E1329: **Testing for global covariate effects in dynamic interaction event networks**
Presenter: **Alexander Kreiss**, Leipzig University, Germany
Co-authors: Wolfgang Polonik, Enno Mammen

In statistical network analysis, it is common to observe so-called interaction data. Such data is characterized by actors forming the vertices and interacting along the edges of the network, where edges are randomly formed and dissolved over the observation horizon. In addition, covariates are observed, and the goal is to model the impact of the covariates on the interactions. We distinguish two types of covariates: global, system-wide covariates (i.e. covariates taking the same value for all individuals, such as seasonality) and local, dyadic covariates modeling interactions between two individuals in the network. Existing continuous time network models are extended to allow for comparing a completely parametric model and a model that is parametric only in the local covariates but has a global non-parametric time component. This allows, for instance, to test whether global time dynamics can be explained by simple global covariates like weather, seasonality etc. The procedure is applied to a bike-sharing network by using weather and weekdays as global covariates and distances between the bike stations as local covariates.

EO514 Room Virtual R01 ANALYSIS OF MULTILAYER NETWORKS Chair: Marianna Pensky

E0959: **Alternating minimization algorithm for clustering mixture multilayer network**
Presenter: **Teng Zhang**, University of Central Florida, United States
Co-authors: Marianna Pensky

A Mixture Multilayer Stochastic Block Model (MMLSBM) is considered, where layers can be partitioned into groups of similar networks, and networks in each group are equipped with a distinct Stochastic Block Model. The goal is to partition the multilayer network into clusters of similar layers, and to identify communities in those layers. The MMLSBM and a clustering methodology, TWIST, based on regularized tensor decomposition have been recently introduced. A different technique is presented, an alternating minimization algorithm (ALMA), that aims at simultaneous recovery of the layer partition, together with the estimation of the matrices of connection probabilities of the distinct layers. Compared to TWIST, ALMA achieves higher accuracy both theoretically and numerically.

E1506: **Clustering in diverse multiplex network model**
Presenter: **Marianna Pensky**, University of Central Florida, United States

A recent multilayer network model is introduced, where all layers of the network have the same collection of nodes and are equipped with the Stochastic Block Models. In addition, all layers can be partitioned into groups with the same community structures, although the layers in the same group may have different matrices of block connection probabilities. The model generalizes previous work on multilayer networks where the same community structure persists in all layers, as well as the Mixture Multilayer Stochastic Block Model, where the layers in the same group have identical matrices of block connection probabilities.

E0770: **Sparse subspace clustering in diverse multiplex network model**
Presenter: **Majid Noroozi**, University of Memphis, United States
Co-authors: Marianna Pensky

The DIverse MultiPLEx (DIMPLE) network model is considered, where all layers of the network have the same collection of nodes and are equipped with the Stochastic Block Models. In addition, all layers can be partitioned into groups with the same community structures, although the layers in the same group may have different matrices of block connection probabilities. The DIMPLE model generalizes a multitude of papers that study multilayer networks with the same community structures in all layers, as well as the Mixture Multilayer Stochastic Block Model (MMLSBM), where the layers in the same group have identical matrices of block connection probabilities. While spectral clustering was previously applied to the proxy of the adjacency tensor, Sparse Subspace Clustering (SSC) is used for identifying groups of layers with identical community structures. Under mild conditions, the latter leads to strongly consistent between-layer clustering. In addition, SSC allows to handle much larger networks than the original methodology and is perfectly suitable for the application of parallel computing.

E1517: **Spectral methods for multiplex networks: An introduction to unfolded spectral embedding**
Presenter: **Andrew Jones**, University of Bristol, United Kingdom

As a generic model for a network, we consider a set of entities together with a function describing their pairwise interactions. In the case of a multiplex network, this function is vector-valued (with each component corresponding to a specific feature of the interaction), and thus we may naturally view such networks as 3-dimensional tensors. Building upon existing spectral methods for graph analysis, unfolded spectral embedding (USE) is a novel technique which exploits this tensor structure to allow us to identify behavioural trends among the underlying entities by aggregating across all layers of interaction, while simultaneously providing us with layer-by-layer representations for comparison. We will briefly introduce USE, discuss how it can be applied to multiplex analogues of stochastic blockmodel graphs (including some asymptotic results concerning the output of the embedding), and present examples of its use in studying real-world dynamic graph networks.

EO360 Room Virtual R02 ADVANCES IN MULTIVARIATE DATA ANALYSIS AND DIMENSION REDUCTION Chair: Abdul-Nasah Soale

E0454: **Clustered covariate regression**
Presenter: **Emmanuel Tsyawo**, Universite Mohammed VI Polytechnique, Morocco
Co-authors: Abdul-Nasah Soale

High dimensionality is an increasingly occurrent phenomenon in model estimation. A common approach to handling high-dimensionality is regularisation-based methods that impose sparsity, and require that several elements of the high-dimensional parameter be zero. However, sparsity cannot always be assumed or easily verified in given empirical contexts. Severe bias and misleading inference may occur when sparsity does not hold. The Grouped Parameter Estimator is introduced, which generalises the notion of sparsity. It remains valid even if the support of the high-dimensional parameter is bounded away from zero. Monte Carlo simulations demonstrate the estimator's high approximative ability of the high-dimensional parameter, improved precision, reduced bias, and a favourably competitive performance relative to competing estimators.

E1224: **Subspace estimation with automatic dimension and variable selection in sufficient dimension reduction**
Presenter: **Jing Zeng**, University of Science and Technology of China, China
Co-authors: Qing Mai, Xin Zhang

Sufficient dimension reduction (SDR) methods target finding lower-dimensional representations of a multivariate predictor to preserve all the information about the conditional distribution of the response given the predictor. The reduction is commonly achieved by projecting the predictor

onto a low-dimensional subspace. The smallest such subspace is known as the Central Subspace (CS) and is the key parameter of interest for most SDR methods. We propose a unified and flexible framework for estimating the CS in high dimensions. The approach generalizes a wide range of model-based and model-free SDR methods to high-dimensional settings, where the CS is assumed to involve only a subset of the predictors. We formulate the problem as a quadratic convex optimization so that the global solution is feasible. The proposed estimation procedure simultaneously achieves the structural dimension selection and coordinate-independent variable selection of the CS. Theoretically, our method achieves dimension selection, variable selection, and subspace estimation consistency at a high convergence rate under mild conditions. We demonstrate the effectiveness and efficiency of our method with extensive simulation studies and real data examples.

E1762: **Nonparametric finite mixture: Applications in contaminated trials**
Presenter: **Solomon Harrar**, University of Kentucky, United States

Investigating the differential effect of treatments in groups defined by patient characteristics is of paramount importance in personalized medicine research. In randomized clinical trials, participants are first classified as having or not having the characteristic of interest by using diagnostic tools, but such classifiers may not be perfectly accurate. The impact of diagnostic misclassification in statistical inference has been recently investigated in parametric model contexts and shown to introduce severe bias in the estimation of treatment effects. The problem is addressed in a fully nonparametric setting. Methods for estimating and testing meaningful yet nonparametric treatment effects are developed. Consistent estimators and asymptotic distributions are provided for the misclassification error rates as well as the treatment effects. The proposed methods are applicable for outcomes measured in ordinal, discrete or continuous scales. The methods do not require any assumptions, such as the existence of moments, on the distribution of the data. Simulation results show significant advantages of the proposed methods in terms of bias reduction, coverage probability and power. The applications of the proposed methods are illustrated with gene expression profiling of bronchial airway brushings in asthmatic and healthy control subjects.

E0473: **A selective review of sufficient dimension reduction for multivariate response regression**
Presenter: **Abdul-Nasah Soale**, University of Notre Dame, United States

Sufficient dimension reduction (SDR) estimators with multivariate response are reviewed. A wide range of SDR methods are characterized as inverse regression SDR estimators or forward regression SDR estimators. The inverse regression family includes pooled marginal estimators, projective resampling estimators, and distance-based estimators. On the other hand, ordinary least squares, partial least squares, and semiparametric SDR estimators are discussed as estimators from the forward regression family.

EO721 Room Virtual R03 ADVANCES IN MARKOV CHAIN MONTE CARLO Chair: Kshitij Khare

E1187: **On some variations of Riemannian manifold and Lagrangian Monte Carlo**
Presenter: **Vivekananda Roy**, Iowa State University, United States

Diffusions based and Hamiltonian dynamics-based methods such as the Metropolis adjusted Langevin algorithms (MALA), and Hamiltonian Monte Carlo (HMC) algorithms have emerged as powerful Metropolis-Hastings algorithms. IWe consider some variations of the Riemannian manifold HMC (RMHMC) and Lagrangian Monte Carlo(LMC) methods. In particular, we investigate the mixtures of the LMC and RMHMC transition kernels with the manifold MALA kernels. The resulting algorithms are shown to converge at a geometric rate under certain conditions. The algorithms are illustrated using several examples.

E1222: **Two-component Gibbs samplers: Convergence rate and asymptotic variance**
Presenter: **Qian Qin**, University of Minnesota, United States
Co-authors: Galin Jones

Deterministic-scan and random-scan Gibbs samplers with two components are compared. In terms of convergence rate, the deterministic-scan version is superior. On the other hand, in terms of asymptotic variance, the random-scan version is more robust. The comparison takes into account the computation cost of the MCMC algorithms.

E1282: **Convergence properties of data augmentation algorithms for high-dimensional robit regression**
Presenter: **Saptarshi Chakraborty**, State University of New York at Buffalo, United States
Co-authors: Sourav Mukherjee, Kshitij Khare

The logistic and probit link functions are common choices for binary regression problems. However, they are not robust to the presence of outliers. The robit link function, defined as the inverse CDF of the Student's t-distribution, provides a robust alternative to them. A multivariate normal prior for the regression coefficients is customary for Bayesian inference in robit regression models. The resulting posterior density is intractable, and a Data Augmentation (DA) Markov chain is used to generate approximate samples from the desired posterior. Establishing geometric ergodicity for this Markov chain is important as it provides theoretical guarantees for asymptotic validity of MCMC standard errors for desired posterior expectations/quantiles. Previous work established the geometric ergodicity of this robit DA chain, assuming restrictions on the sample size n, the number of predictors p, and design matrix X. We show that the robit DA Markov chain is trace-class for arbitrary choices of n, p, and X, and the prior mean and variance parameters. The trace-class property implies geometric ergodicity. Moreover, it allows us to conclude that the sandwich robit chain obtained by inserting an inexpensive extra step in between the two steps of the DA chain is strictly better than the robit DA chain in an appropriate sense, and enables the use of methods to estimate the spectral gap of trace class DA chains.

E1320: **Asynchronous and distributed data augmentation for massive data settings**
Presenter: **Kshitij Khare**, University of Florida, United States

Data augmentation (DA) algorithms are slow in massive data settings due to multiple passes through the entire data. We address this problem by developing a DA extension that exploits asynchronous and distributed computing. The extended DA algorithm is called Asynchronous and Distributed (AD) DA, with the original DA as its parent. Any ADDA is indexed by a parameter $r \in (0,1)$ and starts by dividing the entire data into k disjoint subsets and storing them on k processes. Every iteration of ADDA augments only an r-fraction of the k data subsets with some positive probability and leaves the remaining $(1-r)$-fraction of the augmented data unchanged. The parameter draws are obtained using the r-fraction of new and $(1-r)$-fraction of old augmented data. We show that the ADDA Markov chain is Harris ergodic with the desired stationary distribution under mild conditions on the parent DA algorithm. We demonstrate that ADDA is significantly faster than its parent for many (k,r) choices in three representative models. We also establish the geometric ergodicity of the ADDA Markov chain for all three models, which yields asymptotically valid standard errors for estimates of desired posterior quantities.

EO440 Room Virtual R04 STATISTICAL MODELING, DESIGN, AND INFERENCE Chair: Subir Ghosh

E1171: **Confidence sets for a level set in linear regression**
Presenter: **Fang Wan**, Lancaster University, United Kingdom
Co-authors: Wei Liu, Frank Bretz

Regression modelling is the workhorse of statistics, and there is vast literature on the estimation of the regression function. It is realized in recent years that in regression analysis, the ultimate aim may be the estimation of a level set of the regression function, instead of the estimation of the regression function itself. The published work on the estimation of the level set has thus far focused mainly on nonparametric regression, especially on point estimation. The construction of confidence sets for the level set of linear regression is considered. In particular, $1-\alpha$ level upper, lower

and two-sided confidence sets are constructed for the normal-error linear regression. It is shown that these confidence sets can be easily constructed from the corresponding $1-\alpha$ level simultaneous confidence bands. It is also pointed out that the construction method is readily applicable to other parametric regression models where the mean response depends on a linear predictor through a monotonic link function, which includes generalized linear models, linear mixed models and generalized linear mixed models. Therefore the method is widely applicable. Examples are given to illustrate the method.

E1267: **Flexible handling of dependence for signal identification using functional ANOVA**
Presenter: **David Causeur**, Institut Agro, France
Co-authors: Ching-Fan Sheu

Functional near-infrared spectroscopy (fNIRS) uses the absorption of near-infrared light by hemoglobin to record changes in blood oxygenation as signals of functional brain activity. For designs in which subjects are instructed to execute a specific mental task under different experimental conditions with pre-determined levels for covariates, fNIRS provides real-time cerebral hemodynamic responses for studying neural correlates of task-related experimental variables. Data obtained from such designs are discretized observations of the hemodynamic curves on a high-resolution time scale. Testing for overall group mean differences among curves or, more generally, relationships between curves and explanatory variables can be addressed by using functional Analysis of Variance (fANOVA) procedures in a general multivariate linear regression framework where additional assumptions are made to account for the regularity of mean curves and for the strong time-dependence across residuals. How way time dependence is modeled in such fANOVA testing procedures is crucial and should account for the interplay between the pattern of regression parameter curves and the distribution of the time correlations. To address the challenging issue of identifying time points for which the association signal is nonzero, we propose a fANOVA testing procedure that flexibly handles dependence and deduces optimal signal detection procedures.

E1316: **Two-dimensional P-spline smoothing for spatial analysis of plant breeding trials**
Presenter: **Piepho Hans-Peter**, University of Hohenheim, Germany

Large agricultural field trials may display irregular spatial trends that cannot be fully captured by a purely randomization-based analysis. For this reason, paralleling the development of analysis-of-variance procedures for randomized field trials, there is a long history of spatial modelling for field trials, starting with the early work of Papadakis on nearest neighbour analysis, which can be cast in terms of first or second differences among neighbouring plot values. This kind of spatial modelling is amenable to a natural extension using splines. We consider the P-spline framework, focussing on model options that are easy to implement in linear mixed model packages. Two examples serve to illustrate and evaluate the methods. A key conclusion is that first differences are rather competitive with second differences. Second differences require special attention regarding the representation of the null space of the smooth terms for spatial interaction. An unstructured variance-covariance structure is required to ensure invariance to the translation and rotation of eigenvectors associated with that null space. We develop a strategy that permits fitting this model with ease, but the approach is more demanding than that needed for fitting models using first differences. Hence, even though in other areas, second differences are very commonly used in the application of P-splines, our conclusion is that with field trials, first differences have advantages for routine use.

E1926: **Approximate I-optimal designs for polynomial models over the unit ball**
Presenter: **Linda Haines**, University of Cape Town, South Africa

The focus is on approximate I-optimal designs for full polynomial models over the unit ball. The designs have support on an appropriate set of spheres concentric or coincident with the boundary of the unit ball and place weights on the uniform distributions over those spheres. The result is stated in a single sentence in a paper in 1977, but has not been revisited since. The indicated proof is formalized and the requisite designs are constructed using an approach which emanates from Euclidean design theory. Comparisons of the approximate I-optimal designs with their D-optimal counterparts are made and a Pareto approach to obtaining a design which is a compromise between D- and I-efficiency is introduced. Examples which reinforce the findings are presented throughout.

EO717 Room Virtual R05 OPTIMAL TRANSPORT: RECENT THEORETICAL ADVANCES Chair: Nabarun Deb

E0657: **Limit theorems for smooth Wasserstein distances**
Presenter: **Kengo Kato**, Cornell University, United States
Co-authors: Ziv Goldfeld, Sloan Nietert, Gabriel Rioux

The Wasserstein distance is a metric on a space of probability measures that have seen a surge of applications in statistics, machine learning, and applied mathematics. However, statistical aspects of Wasserstein distances are bottlenecked by the curse of dimensionality, whereby the number of data points needed to be estimated accurately grows exponentially with dimension. Gaussian smoothing was recently introduced as a means to alleviate the curse of dimensionality, giving rise to a parametric convergence rate in any dimension, while preserving the Wasserstein metric and topological structure. To facilitate valid statistical inference, we develop a comprehensive limit distribution theory for the empirical smooth Wasserstein distance. The limit distribution results leverage the functional delta method after embedding the domain of the Wasserstein distance into a certain dual Sobolev space, characterizing its Hadamard directional derivative for the dual Sobolev norm, and establishing weak convergence of the smooth empirical process in the dual space. To estimate the distributional limits, we also establish the consistency of the nonparametric bootstrap.

E1070: **Wasserstein gradient flows of entropic optimal transport**
Presenter: **Lenaic Chizat**, EPFL, Switzerland

Entropic Optimal Transport (EOT) is a modification of the Optimal Transport problem that is statistically and computationally more tractable, and that is a building block of several useful tools in machine learning (Sinkhorn divergence, barycenters of measures, trajectory inference, etc). We study the well-posedness, and the long-time behavior of Wasserstein gradient flows of functionals involving EOT on a compact domain. The well-posedness relies on a stable result for the solutions of the dual EOT problem, and the long-time behavior follows from a nonlinear generalization of the convergence of overdamped Langevin dynamics via log-Sobolev inequalities. Finally, we will discuss applications to the grid-free computation of regularized Wasserstein barycenters and the problem of trajectory inference.

E1231: **Convergence rates for regularized optimal transport via quantization**
Presenter: **Stephan Eckstein**, ETH Zurich, Switzerland
Co-authors: Marcel Nutz

A simple approach is showcased to obtain sharp convergence rates for the convergence of divergence-regularized optimal transport as the regularization parameter vanishes. The approach is based on quantization, where we balance an increasingly large discrete approximation of the marginals against a decreasing regularization parameter. The methodology is flexible and applicable to different divergences, multi-marginal problems and a large class of cost functions. Among others, this yields the sharp leading-order term for entropically regularized 2-Wasserstein distance under just $(2+\delta)$-moment assumption on the marginals.

E1290: **Convergence of Wasserstein metric under data dependence in multi-dimension**
Presenter: **Debarghya Mukherjee**, Princeton University, United States
Co-authors: Nabarun Deb

The Wasserstein distance is a powerful tool in modern machine learning to metrize the space of probability distributions in a way that considers the

domain's geometry. Therefore, a lot of attention has been devoted to understanding rates of convergence for Wasserstein distances based on i.i.d data. However, often in machine learning applications, especially in reinforcement learning, object tracking, performative prediction, and other online learning problems, observations are received sequentially, rendering some inherent temporal dependence. Motivated by this observation, we attempt to understand the problem of estimating Wasserstein distances using the natural plug-in estimator based on stationary beta-mixing sequences, a widely used assumption in the study of dependent processes. Our rates of convergence results are applicable under both short and long-range dependence. As expected, under short-range dependence, the rates match those observed in the i.i.d. case. Interestingly, however, even under long-range dependence, we can show that the rates can match those in the i.i.d. case provided the (intrinsic) dimension is large enough. Our analysis establishes a non-trivial trade-off between the degree of dependence and the complexity of certain function classes on the domain. The key technique in our proofs is a blend of the big-block-small-block method coupled with Berbees lemma and chaining arguments for suprema of empirical processes.

EO739 Room Virtual R06 ADVANCES IN STATISTICAL METHODS FOR COMPLEX GENETIC/GENOMIC DATA Chair: Yuehua Cui

E0826: **Significance tests based on neural networks with applications to genetic association studies**
Presenter: **Xiaoxi Shen**, Texas State University, United States
Co-authors: Chang Jiang, Lyudmila Sakhanenko, Qing Lu

Despite the great success of applications of neural networks in many different fields, such as natural language processing and image recognition, there is a lack of research that focuses on the interpretation of neural network models. Two hypothesis testing methods based on neural networks with one hidden layer will be introduced to conduct significance tests of input features. The asymptotic distributions for both test statistics are simple, so it is easy to apply in real data analysis. The validity of the asymptotic distributions is investigated via simulations, and we applied our proposed tests to perform a genetic association analysis on the sequencing data from Alzheimer's Disease Neuroimaging Initiative (ADNI).

E0832: **High dimensional mediation analysis via difference in coefficients with applications in genetics**
Presenter: **Qi Zhang**, University of New Hampshire, United States

High-dimensional mediation analysis has been enjoying increasing popularity, largely motivated by the scientific problems in genomics and biomedical imaging. Previous literature has been primarily focused on mediator selection. There has also been work on estimating the overall indirect effect for low dimensional exposure. We aim at estimation and inference of the overall indirect effect for high dimensional exposures and high dimensional mediators. We propose MedDiC, a novel debiased estimator of the high dimensional overall indirect effect based on the difference-in-coefficients approach. We evaluate the proposed method using intensive simulations, and find the MedDiC provides valid inference and offers higher power and shorter computing time than the competitors for both low dimensional and high dimensional exposures. We also apply MedDiC to a mouse f2 dataset for diabetes study, and a dataset composed of diverse maize inbred lines for flowering time, and show that MedDiC yields more biologically meaningful gene lists. The results are reproducible across different measures of identical biological signal.

E1208: **Graph neural networks for multimodal single-cell data integration**
Presenter: **Yuying Xie**, Michigan State University, United States

Recent advances in multimodal single-cell technologies have enabled simultaneous acquisitions of multiple omics data from the same cell, providing deeper insights into cellular states and dynamics. However, it is challenging to learn the joint representations from the multimodal data, model the relationship between modalities, and, more importantly, incorporate the vast amount of single-modality datasets into the downstream analyses. To address these challenges and correspondingly facilitate multimodal single-cell data analyses, three key tasks have been introduced: Modality prediction, Modality matching and joint embedding. We present a general Graph Neural Network framework scMoGNN to tackle these three tasks and show that scMoGNN demonstrates superior results in all three tasks compared with the state-of-the-art and conventional approaches. All implementations of our methods have been integrated into DANCE package.

E1423: **Causal inference with Mendelian randomization for longitudinal traits**
Presenter: **Yuehua Cui**, Michigan State University, United States

Mendelian Randomization uses genetic variants as instrument variables to determine whether an observational association between a risk factor and an outcome is consistent with a causal effect. The use of Mendelian Randomization reduces regression bias and provides a more reliable estimate of the likely underlying causal relationship between an exposure and a disease outcome. Most current Mendelian Randomization methods are focused on cross-sectional phenotypic traits. Longitudinal studies track the same sample at different time points and have a number of advantages over cross-sectional studies. It would be possible for researchers to learn more about 'cause and effect' relationships when incorporating time information. We propose a time lag model to investigate the delayed causal effects in a longitudinal study. We assume that both the current and past values of exposure contribute to the current outcome. In order to select the duration of delay included in the model, an algorithm is developed for the variable selection purpose. The point-wise testing and simultaneous testing are developed to test the existence of causal effects. The method was illustrated via simulation studies and an application to a real dataset.

EO556 Room Virtual R07 OPPORTUNITIES AND CHALLENGES OF NEUROIMAGING DATA Chair: Aaron Scheffler

E0227: **Biobank-scale imaging genetics for human health: Findings, perspectives, and resources**
Presenter: **Bingxin Zhao**, University of Pennsylvania, United States

In recent years, the UK Biobank study has scanned over 40,000 participants' brains and bodies using magnetic resonance imaging (MRI). In addition, publicly available imaging genetic datasets also emerge from several other independent studies. We collected massive individual-level MRI data from different data resources, harmonized image processing procedures, and conducted biobank-scale genetic studies for various traits resulting from diverse imaging modalities. We showcase novel clinical findings from our analyses, such as the shared genetic influences between brain and heart health. We also discuss current progress, future topics, and publicly available resources in this research area. More information about our studies is available at the Brain Imaging Genetics Knowledge Portal.

E1221: **Scan once, analyse many: Using large open-access neuroimaging datasets to understand the brain**
Presenter: **Christopher Madan**, University of Nottingham, United Kingdom

Brain imaging data are readily available. Researchers are now sharing data more than ever before. Additionally, large-scale data-collecting initiatives are underway with the vision that many future researchers will use the data for secondary analyses. An overview of available datasets and some example use cases are provided. Example use cases include examining individual differences, more robust findings, reproducibility–both in public input data and availability as a replication sample, and methods development.

E1344: **Bayesian integration of multi-modal imaging data and efficient inference with random data compression**
Presenter: **Rajarshi Guhaniyogi**, Texas A & M university, United States
Co-authors: Aaron Scheffler

Clinical researchers often collect multiple images from separate modalities (sources) to investigate fundamental questions of human health that are inadequately explained by considering one image source at a time. Viewing the collection of images as multiple objects, the successful integration of multi-object data produces a sum of information greater than the individual parts, but this integration can be challenging due to the complexity induced by the different topological structures of the objects. We will show a novel joint prior formulation that integrates information from networks

and structural images to draw inferences on brain regions significantly related to the language score predictive of Primary Progressive Aphasia. The principled Bayesian framework allows precise characterization of the uncertainty in ascertaining a region being actively related to the language score. Our framework is implemented using an efficient Markov Chain Monte Carlo algorithm. Empirical results with simulated data illustrate substantial inferential gains of the proposed framework over its popular competitors. Our framework yields new insights into the relationship of brain regions with PPA, offering neuro-degeneration pathways for PPA. We will also show strategies to draw scalable inferences with large data using random data compression approach.

E2015: **A vine copula change point model for neuroimaging studies and their computational reproducibility**
Presenter: **Ivor Cribben**, Alberta School of Business, Canada

A new methodology called Vine Copula Change Point (VCCP) is introduced to estimate change points in the network structure between multivariate time series. It uses vine copulas, various state-of-the-art segmentation methods to identify multiple change points, and a likelihood ratio test or the stationary bootstrap for inference. The vine copulas allow for various forms of dependence, including tail, symmetric and asymmetric dependence, which has not been explored before in the dynamic analysis of neuroimaging data. We will also discuss some recent work on reproducibility in statistics by attempting to reproduce the results in 93 published papers in prominent journals utilizing functional magnetic resonance imaging (fMRI) data during the 2010-2021 period.

EO555 Room Virtual R08 ADVANCES IN NONPARAMETRIC STATISTICS FOR LARGE-SCALE DATASET Chair: Shan Yu

E0643: **Optimal nonparametric inference with two-scale distributional nearest neighbors**
Presenter: **Lan Gao**, University of Tennessee Knoxville, United States
Co-authors: Emre Demirkaya, Yingying Fan, Jinchi Lv, Patrick Vossler, Jingbo Wang

The weighted nearest neighbors (WNN) estimator has been popularly used as a flexible and easy-to-implement nonparametric tool for mean regression estimation. The bagging technique is an elegant way to form WNN estimators with weights automatically generated to the nearest neighbors; we name the resulting estimator as the distributional nearest neighbors (DNN) for easy reference. Yet, there is a lack of distributional results for such an estimator, limiting its application to statistical inference. Moreover, when the mean regression function has higher-order smoothness, DNN does not achieve the optimal nonparametric convergence rate, mainly because of the bias issue. We provide an in-depth technical analysis of the DNN, based on which we suggest a bias reduction approach for the DNN estimator by linearly combining two DNN estimators with different subsampling scales, resulting in the novel two-scale DNN (TDNN) estimator. We prove that the two-scale DNN estimator enjoys the optimal nonparametric rate of convergence in estimating the regression function under the fourth-order smoothness condition. We further go beyond estimation and establish asymptotic normality for DNN and two-scale DNN estimators. For the practical implementation, we also provide variance estimators and a distribution estimator using the jackknife and bootstrap techniques for the two-scale DNN.

E0871: **Big spatial data learning: A parallel solution**
Presenter: **Shan Yu**, University of Virginia, United States
Co-authors: Guannan Wang, Lily Wang

Nowadays, we are living in the era of Big Data. A significant portion of big data is big spatial data captured through advanced technologies or large-scale simulations. Explosive growth in spatial and spatiotemporal data emphasizes the need for developing new and computationally efficient methods and credible theoretical support tailored for analyzing such large-scale data. Parallel statistical computing has proved to be a handy tool when dealing with big data. However, it is hard to execute the conventional spline regressions in parallel. We will present a novel parallel smoothing technique for generalized partially linear spatially varying coefficient models, which can be used under different hardware parallelism levels. Moreover, conflated with concurrent computing, the proposed method can be easily extended to the distributed system. The newly developed method is evaluated through several simulation studies and an analysis of the US Loan Application Data.

E1332: **Structure identification of space-time epidemic models**
Presenter: **Zhiling Gu**, Iowa State University, United States
Co-authors: Guannan Wang, Xinyi Li, Lily Wang

Epidemiological models are a vital tool for understanding a course of an epidemic and making predictions. And model complexity is a crucial factor that significantly affects the forecast accuracy of pandemics. Specifically, at the early stage of a pandemic, a simple model is usually preferred due to the sparsity of cases. As the disease progresses, a more complex model with a significant amount of flexibility can better capture the heterogeneities and complexity of the underlying process. Therefore, it is of great interest to develop an analytic tool that can automatically balance the simplicity and flexibility of epidemiological models. We consider a class of space-time epidemic models (STEMs) to investigate spatial-temporal patterns in disease spread at the area level. Based on this flexible modeling framework, we develop a structure identification method to adjust the model complexity by automatically detecting predictors with linear, nonlinear, and spatially varying effects on the response. Moreover, we investigate the theoretical properties of the proposed method. We show the consistency of different types of estimators and asymptotic normality for estimators of the linear components. The proposed method is evaluated by Monte Carlo simulation studies and applied to analyze the COVID-19 outbreak.

E1497: **Inference for nonparanormal partial correlation via regularized rank-based nodewise regression**
Presenter: **Yumou Qiu**, Iowa State University, United States

Partial correlation is a common tool in studying conditional dependence for Gaussian distributed data. However, the partial correlation being zero may not be equivalent to conditional independence under non-Gaussian distributions. We propose a statistical inference procedure for partial correlations under the high-dimensional nonparanormal (NPN) model, where the observed data are normally distributed after certain monotone transformations. The nonparanormal partial correlation is the partial correlation of the normal transformed data under the NPN model, which is a more general measure of conditional dependence. We estimate the NPN partial correlations by regularized nodewise regression based on the empirical ranks of the original data. A multiple testing procedure is proposed to identify the nonzero NPN partial correlations. The proposed method can be carried out by a simple coordinate descent algorithm for lasso optimization. It is easy to implement and computationally more efficient compared to the existing methods for estimating NPN graphical models. Theoretical results are developed to show the asymptotic normality of the proposed estimator and to justify the proposed multiple-testing procedure. Numerical simulations and a case study on brain imaging data demonstrate the utility of the proposed procedure and evaluate its performance compared to the existing methods.

EO643 Room K2.31 (Nash Lec. Theatre) SAFE & TRUSTWORTHY PREDICTIVE MODELING Chair: Stathis Gennatas

E1948: **Multi-modal prototype learning for interpretable multivariable time series classification**
Presenter: **Reza Abbasi Asl**, University of California, San Francisco, United States

Multivariable time series classification problems are increasing in prevalence and complexity in a variety of domains, such as biology and finance. While deep learning methods are an effective tool for these problems, they often lack interpretability. We propose a novel modular prototype learning framework for multivariable time series classification. In the first stage, encoders extract features from each variable independently. Prototype layers identify single-variable prototypes in the resulting feature spaces. The next stage represents the multivariable time series sample points in terms of their similarity to these single-variable prototypes. This results in an inherently interpretable representation of multivariable patterns, on which prototype learning is applied to extract representative examples, i.e. multivariable prototypes. We validate our framework on a simulated dataset with embedded patterns, as well as a real human activity recognition problem. Our framework attains comparable or superior

classification performance to existing time series classification methods on these tasks. On the simulated dataset, we find that our model returns interpretations consistent with the embedded patterns. Moreover, the interpretations learned on the activity recognition dataset align with domain knowledge.

E1974: **Making predictions under hypothetical interventions in clinical prediction models**
Presenter: **Niels Peek**, The University of Manchester, United Kingdom

The methods with which clinical prediction models are usually developed mean that neither the parameters nor the predictions should be interpreted causally. For many applications, this is perfectly acceptable. However, when these models are used to support clinical decision-making, there is often a need for predicting outcomes under hypothetical interventions. We aimed to compare methodological approaches for predicting individual-level cardiovascular risk under three hypothetical interventions: smoking cessation, reducing blood pressure, and reducing cholesterol. We used data from the PREDICT prospective cohort study in New Zealand to calculate cardiovascular risk in a primary care setting. We compared three strategies to estimate absolute risk under hypothetical interventions: (a) conditioning on hypothetical interventions in non-causal models; (b) integrating existing prediction models with causal effects estimated using observational causal inference methods; and (c) integrating existing prediction models with causal effects reported in published literature.

E2018: **How can AI and ML disrupt critical care?**
Presenter: **Romain Pirracchio**, UCSF, United States

The importance of critical care has been highlighted during the Covid-19 pandemic. It has also highlighted the fact that ICU resources are limited and should be used wisely. Ever since its creation in the mid-1900s, critical care has typically been a curative healthcare service. Indeed, patients are admitted to the ICU if they present organ dysfunctions that can only be treated in the ICU environment. Although supportive care is much needed in the most severe patients, an alternative approach where patients at risk of deterioration are identified early could help prevent organ dysfunction from happening and potentially improve the outcome, and also help better identify the patients who need to be treated in an ICU environment. Recent developments in predictive analytics may promote the much-needed transition from a purely curative to more of a preventative paradigm for ICU care delivery.

EO583 Room K2.40 RECENT DEVELOPMENTS ON FUNCTIONAL DATA ANALYSIS (VIRTUAL) Chair: Sara Lopez Pintado

E0587: **Regularized halfspace depth for functional Data**
Presenter: **Hyemin Yeon**, Iowa State University, United States
Co-authors: Xiongtao Dai, Sara Lopez Pintado

Data depth is a powerful nonparametric tool originally proposed to rank multivariate data from center outward. For multivariate data, one of the most archetypical depth notions is the halfspace depth by Tukey. In the last few decades, notions of depth have been proposed for functional data. However, the halfspace depth by Tukey cannot be extended to handle functional data because of a degeneracy issue. In our work, we propose a new halfspace depth for functional data and avoid degeneracy by regularization. The halfspace projection directions are constrained to have a small reproducing kernel Hilbert space norm. Desirable theoretical properties of the proposed depth, such as isometry invariance, maximality at center, monotonicity relative to a deepest point, and upper semi-continuity, are established. Moreover, the proposed regularized halfspace depth can rank functional data with a varying emphasis in shape or magnitude, depending on the regularization. A new outlier detection approach is also proposed, which is capable of detecting both shape and magnitude outliers. It is applicable to trajectories in L^2, a highly general space of functions including non-smooth trajectories. Based on an extensive numerical study, our methods are shown to perform well in terms of detecting outliers of different types. Three real data examples showcase the proposed depth notion.

E1067: **Quantile regression for longitudinal functional data with application to feed intake of lactating sows**
Presenter: **Maria Laura Battagliola**, EPFL, Switzerland
Co-authors: Helle Sorensen, Anders Tolver, Ana-Maria Staicu

A model framework and estimation methodology are introduced for quantile regression in scenarios with clustered or longitudinal data and functional covariates. The proposed quantile regression model uses a time-varying regression coefficient function to quantify the association between covariates and quantile level of interest, and includes subject-specific intercepts to incorporate within-subject dependence. Estimation relies on spline representation of the unknown coefficient functions, and can be carried out with existing software. The proposed method is studied numerically in a simulation study that covers a wide range of situations, and bootstrap procedures for bias adjustment and computation of standard errors are introduced. The work is motivated by a study on lactating sows, where the main interest is the influence of temperature, measured throughout the day, on the lower quantiles of feed intake. Analysis of the lactation data indicates, among others, that the influence of temperature increases during the lactation period.

E1439: **Nonparametric functional data modeling of pharmacokinetic processes with applications in dynamic PET imaging**
Presenter: **Todd Ogden**, Columbia University, United States
Co-authors: Baoyi Shi

Modeling a pharmacokinetic process typically involves solving a system of linear differential equations and estimating the parameters upon which the functions depend. In order for this approach to be valid, it is necessary that a number of fairly strong assumptions hold, assumptions involving various aspects of the kinetic behavior of the substance being studied. In many situations, such models are understood to be simplifications of the "true" kinetic process. While in some circumstances, such a simplified model may be a useful (and close) approximation to the truth, in some cases, important aspects of the kinetic behavior cannot be represented. We present a nonparametric approach, based on principles of functional data analysis, to modeling of pharmacokinetic data. We illustrate its use through application to data from a dynamic PET imaging study of the human brain.

E1478: **Global depths for irregularly observed multivariate functional data**
Presenter: **Wenlin Dai**, Renmin University of China, China
Co-authors: Zhuo Qu, Marc Genton

Two frameworks for multivariate functional depth based on multivariate depths are introduced. The first framework is multivariate functional integrated depth, whereas the second framework, multivariate functional extremal depth, is extended from the extremal depth for univariate functional data. In each framework, global and local multivariate functional depths are proposed. Properties of population multivariate functional depths and the consistency of the finite sample depths to their population versions are proved. In addition, the estimation of finite sample depths under irregularly observed time grids is investigated. Finally, the simplified sparse functional boxplot and the simplified intensity sparse functional boxplot are proposed for visualization without the need for data reconstruction. A simulation study demonstrates the advantages of global multivariate functional depths over local multivariate functional depths in outlier detection and faster running time. An application to cyclone track data demonstrates the excellent performance of our global multivariate functional depths.

CO644 Room S-2.25 MACHINE LEARNING IN FINANCE Chair: Anastasija Tetereva

C0841: **Infinite sparse factor stochastic volatility model**
Presenter: **Martina Zaharieva**, CUNEF Universidad, Spain

A sparse factor multivariate stochastic volatility model is proposed, in which the sparsity of the loading matrix is achieved by introducing the Indian buffet process, a Bayesian nonparametric prior defining a distribution over infinite binary matrices. The benefit of the infinite-dimensional latent process is twofold. First, inducing sparsity prior reduces the dimensionality of the problem, and second, the number of active factors is determined by the data itself and a priori set to infinity. Both, the diagonal elements of the covariance matrix of the idiosyncratic term, and the active factors follow univariate stochastic volatility processes. Each latent volatility is sampled independently and in parallel by means of a particle filtering and smoothing technique.

C0883: **When firms open up: Identifying value relevant textual disclosure using simBERT**
Presenter: **Christian Breitung**, Technical University of Munich, Germany
Co-authors: Sebastian Mueller

By introducing simBERT, a novel semantically sensitive similarity measure for textual data, we find that international annual reports contain value-relevant information that investors do not timely price. We measure the value relevance of international corporate disclosures by constructing a portfolio that is long in stocks with a low- and short in stocks with a high level of semantically new information. Such a portfolio yields a highly significant yearly abnormal return of 8.52%. We observe a higher value relevance of textual disclosure in developed countries, which we trace back to stricter securities laws standards. Our findings thus indicate that tighter regulation promotes the disclosure of value-relevant accounting information. We further find evidence that analysts update their earnings forecasts and recommendations in accordance with textual changes in firm reports. This suggests that analysts contribute to market efficiency by conveying qualitative information from accounting statements to the public.

C1285: **Intrinsic factor risk premia and tests of asset pricing models**
Presenter: **Alberto Quaini**, Columbia University, United States
Co-authors: Fabio Trojani, Ming Yuan

An intrinsic factor risk premium is given by the factor covariance with the maximum Sharpe ratio return out of a set of test assets. When the maximum Sharpe ratio is finite, intrinsic risk premia are well-defined and equal to zero for any factor, implying a zero or a vanishing correlation with returns. If a factor is not tradable, intrinsic risk premia are not consistently estimated by established cross-sectional two-step estimation methods. Therefore, we introduce an Oracle intrinsic risk premium estimator, which is asymptotically normal and consistently selects all intrinsically priced factors. Using our estimation and inference methodology based on intrinsic risk premia, we study a family of one to six-factor asset pricing models from the factor zoo. In this context, we clarify the key role of the interplay between misspecification and factor tradeability for understanding the pricing of factor risk and comparing asset pricing models.

C1388: **Training economic tracking portfolios using trees of assets**
Presenter: **Onno Kleen**, Erasmus University Rotterdam, Netherlands
Co-authors: Anastasija Tetereva, Rasmus Lonn

Stock returns can help in nowcasting and forecasting economic variables such as consumption, inflation, or unemployment. Typically, economic tracking portfolios are based on aggregated portfolio returns. A modified version of regression trees is introduced that allows us to employ individual stocks to track economic variables. We build cross-sections of managed portfolios that serve as tracking portfolios, and that can be used in asset pricing models. Empirical results show that our portfolios display superior tracking properties across various market conditions.

CO675 Room S-1.01 ASSET PRICING Chair: Benjamin Holcblat

C0821: **Debt-stabilizing properties of GDP-linked securities: A macro-finance perspective**
Presenter: **Sarah Mouabbi**, Banque de France, France
Co-authors: Jean-Paul Renne, Jean-Guillaume Sahuc

The debt-stabilizing properties of indexing debt to GDP are studied using a consumption-based macro-finance model. To this end, we derive quasi-analytical pricing formulas for any type of bond or equity by exploiting the discretization of the state-space, which makes large-scale simulations tractable. Such pricing formulas are feasible thanks to an approximation of the risk-neutral dynamics – a novelty in this class of models. Three results stand out. First, GDP-linked security prices would embed sizable and time-varying risk premiums of about 40 basis points. Second, for a fixed budget surplus, issuing GDP-linked securities does not necessarily imply more beneficial debt-to-GDP ratios in the medium- to long-run. Third, the debt-stabilizing budget surplus is more predictable under such issuances at the expense of being higher on average. Our findings call into question the view that GDP-linked securities tame debt.

C0902: **Dynamic financial constraints in presence of uncertainty: Theory and evidence from credit lines**
Presenter: **Niklas Amberg**, Sveriges Riksbank, Sweden

Using a comprehensive Swedish credit registry, firms throughout the size distribution are documented to have access to substantial amounts of unused borrowing capacity via credit lines. This finding seems to conflict with the notion that financial constraints are widespread in the economy, but we show that uncertainty can help reconcile the apparent contradiction. We begin by constructing a model in which firms optimally choose not to borrow up to the limit when facing uncertainty about future productivity and access to external financing; this choice results from a trade-off between the benefits of borrowing today and the expected cost of illiquidity tomorrow. We then empirically test and confirm the main predictions of the model, in particular, that credit-line utilisation rates are negatively related to idiosyncratic uncertainty. Our findings imply that financial constraints need to be assessed dynamically, since a firm can be financially constrained even when its borrowing constraint is not binding in a static sense.

C1426: **A greenwashing index**
Presenter: **Elise Gourier**, ESSEC, France
Co-authors: Helene Iung-Mathurin

Textual analysis of articles in the Wall Street Journal is used to construct indices of attention to greenwashing over the last forty years, in aggregate and by industry. We measure to which extent greenwashing contaminates the management and hedging of climate change risk.

C1813: **Dissecting anomalies in conditional asset pricing**
Presenter: **Valentina Raponi**, IESE Business School, Spain
Co-authors: Paolo Zaffaroni

A methodology is developed for estimating and testing the effect of anomalies in conditional asset pricing models when premia vary over time. By showing that conventional approaches are ill-suited to estimate time-varying anomalies premia, we develop a new method based on simple ordinary and weighted least square estimation and provide closed-form standard errors that can be used to make inferences on the premia parameters. To quantify the effect and the economic significance of anomalies, a new cross-sectional R-squared test is also proposed. Using a dataset of 20,000 individual US stock returns, we find that most of the anomalies, although statistically significant, explain only a small fraction (less than 10%) of the cross-sectional variation of expected returns. Moreover, their effect tends to be more important during economic and financial crises.

CO462 Room S-1.22 ADVANCES IN ECONOMETRICS Chair: Rustam Ibragimov

C1959: **Joint-VaR: A new conditional risk measure**
Presenter: **Elisabetta Mensali**, University of Bologna, Italy
Co-authors: Leopoldo Catania, Alessandra Luati

The Joint Value at Risk (JVaR) is defined as the quantile of an asset return distribution, given an upper tail event affecting its log-volatility. The purpose of JVaR is to measure financial risk under a volatility stress scenario. A distinguishing feature of the proposed risk measure is that conditioning events are latent. The relations with the VaR and the CoVaR, that is, the VaR conditional to some observed event, are made explicit. The properties of JVaR are studied based on a stochastic volatility representation of the underlying process. We prove that JVaR is leverage consistent, i.e. it is an increasing function of the dependence parameter in the stochastic representation. The difference between the JVaR and its reference state, represented by the VaR, provides a natural tool for monitoring risk. A feasible class of semiparametric M-estimators is introduced by exploiting the elicitability of quantiles and the stochastic ordering theory. Consistency and asymptotic normality of the proposed JVaR M-estimator are derived in two steps based on the pair (VaR, JVaR), and its finite-sample properties are illustrated in a simulation study. Empirical results with S&P500 data show that accounting for extreme volatility levels is relevant to characterize the evolution of risk better.

C1972: **Impact of machine learning-based traders on the high-frequency stock market**
Presenter: **Kirill Mansurov**, Saint-Petersburg State University, Russia
Co-authors: Alexander Semenov, Dmitry Grigoriev, Rustam Ibragimov

The role of self-learning agents in multi-agent models on financial markets is investigated. We develop an agent-based simulation model of a stock market, and in addition to the agents with fixed strategies used in previous research, we introduce an agent with a self-learning strategy. To model the behavior of such an agent, we use deep reinforcement learning algorithms, namely Deep Deterministicpolicy gradient (DDPG). Next, we conduct a comparative analysis of the results of the constructed model with outcomes of previously proposed models, as well as with the characteristics of real markets. To conduct a comparative analysis, we use stylized facts of asset return that allow us to evaluate and compare the characteristics of the markets. Our results show that a model with a self-learning agent gives a better approximation to the real market than a model with classic agents. In particular, unlike the model with classical agents, the model with a self-learning agent turns out to be not so heavy-tailed. Thus, we demonstrate that for a complete understanding of market processes, simulation models should take into account self-learning agents that have a significant presence on the SP 500 index.

C1969: **Mixed integer optimization for time series change points detection**
Presenter: **Alexander Semenov**, University of Florida and Saint Petersburg State University, United States
Co-authors: Artem Prokhorov, Anton Skrobotov

Recent advances in mixed-integer optimization (MIO) methods are used to develop a framework for identifying and estimating structural breaks in time series. The framework requires a transformation of the classical structural break detection problem into a Mixed Integer Quadratic Programming problem. MIO is capable of finding provably optimal solutions to this problem using a well-known optimization solver. The framework allows us to determine the unknown number of structural breaks. In addition to that, we demonstrate how to accommodate a specific required number of structural breaks, or a minimal required number of breaks. We demonstrate the effectiveness of our approach through extensive numerical experiments on synthetic and real-world data. We examine optimal and sub-optimal solutions to the problem, and the effect of tuning the parameters. We show how to choose the tuning parameters and compare our results with established econometric methods.

C1977: **Intergenerational transmission of tail inequality in incomes**
Presenter: **Paul Kattuman**, University of Cambridge, United Kingdom
Co-authors: Joseph Gatus, Rustam Ibragimov

It is known that over the last many decades, income inequality has increased substantially. It is also known that much of the increase has been in the upper tail of the income distribution. It would be useful to know to what extent the increase in upper-tail income inequality is intergenerationally transmitted. The tail index of the income distribution is a summary measure of inequality among top incomes. We examine the direction and extent to which the intergenerational mobility estimated for the United Kingdom using the British Household panel survey accounts for the changes in the tail index of the income distribution of the parent's generation to the tail index of the income distribution of the child's generation. We elucidate the methodological contribution and discuss findings and implications for the evolution of upper-tail income inequality.

CO478 Room Safra Lecture Theatre NOVEL APPROACHES TO TIME SERIES FORECASTING Chair: Anindya Roy

C1023: **Zero-crossings of time series: Sign-prediction, mean-square error and a holding-time constraint**
Presenter: **Marc Wildi**, University of Applied Sciences Zurich, Switzerland

An extension of classic time series approaches is proposed, which addresses zero-crossings of a zero-mean stationary time series. Specifically, we subject the original optimization criterion to a novel holding-time constraint which conditions the expected duration between consecutive crossings. Formally, the solution to this prediction problem is obtained by assigning more weight to components of the classic estimate, which are in better accordance with the constraint. Besides an analysis based on simple forecast and signal extraction exercises, we also provide an application of our novel approach to business-cycle analysis. The latter example illustrates that improved smoothing capability (noise suppression) does not necessarily conflict with timeliness (relative lead) in a real-time nowcasting setting.

C1147: **Multivariate direct filter analysis for co-integrated processes**
Presenter: **Tucker McElroy**, Census Bureau, United States
Co-authors: Marc Wildi

Real-time signal extraction for multivariate time series extracts trends, cycles, and seasonal patterns by utilizing information from related series. Optimal real-time signal extraction filters can be obtained through the multivariate direct filter analysis (MDFA) methodology. Two advances to MDFA are discussed: (i) integration constraints and (ii) co-integration constraints. A new approach is used, whereby integration/co-integration constraints are incorporated into the optimization problem by transforming the real-time filters. The new techniques are demonstrated in construction and employment data.

C1196: **Transformer-based models for time series forecasting**
Presenter: **Thu Nguyen**, University of Maryland Baltimore County, United States

Time series forecasting has long been a key and well-studied area of academic research and has many important applications in topics such as commercial decision-making in retail, finance, product development and planning, biological sciences and medicine, to name a few. In contrast to traditional methods for time series forecasting, which focus on parametric models informed by domain expertise and rely heavily on well-designed features, modern machine learning methods, especially deep learning, which gained popularity in recent times, try to learn temporal dynamics in a purely data-driven manner. With the increasing data availability and computing power in recent times, as well as their success in practical applications and competitions, deep learning-based forecasting models have become a popular choice for many time series forecasting tasks. We present a new class of models inspired by Transformer based architecture for the time series forecasting tasks. The proposed models' architecture is based on the attention mechanism, which has received increased interest in applying for time series-related applications. In addition to employing

new attention mechanisms, we also utilize some ideas from classical time series methods to learn complex patterns and dynamics from time series data. The proposed models are flexible enough to apply to both univariate and multivariate time series data.

C1090: **Adjusting for seasonality using point process models**
Presenter: **Anindya Roy**, U.S. Census Bureau, United States
Co-authors: Tucker McElroy

Data publishing agencies are increasingly publishing data collected at a higher frequency than in the past. This has resulted in several published mixed-frequency time series data. Seasonal adjustment of mixed frequency time series is a challenging problem. We use a marked point process model to model the series as aggregation at different frequency levels. We use a transformation of the intensity measure to excise periodic components, thereby adjusting the data to be non-seasonal. The method is illustrated with daily, weekly, and monthly time series.

CO298 Room K2.41 TOPICS IN TIME SERIES ECONOMETRICS Chair: Kanchana Nadarajah

C1053: **Fully data-driven non-parametric estimation of Toeplitz covariance matrices**
Presenter: **Karolina Klockmann**, University of Vienna, Austria
Co-authors: Tatyana Krivobokova

Estimating the Toeplitz covariance matrix of a single realization of a stationary stochastic process is a central problem in time series econometrics. It is well known that the sample auto-covariance matrix is inconsistent in the spectral norm, so regularized versions, such as the tapered or banded covariance estimators, have been proposed. However, such estimators are not guaranteed to be positive definite, and data-driven choices of the regularization parameters are not available. We present an estimator for the Toeplitz covariance matrix and its inverse, which overcome these drawbacks. First, we derive an alternative version of the Whittle likelihood based on the Discrete Cosine Transform matrix, which is shown to asymptotically diagonalize Toeplitz matrices. Using variance stabilizing transforms, we transform the resulting Gamma regression problem into an approximate Gaussian regression setting for the log-spectral density. The resulting estimators for the Toeplitz covariance matrix and its inverse are positive definite and all regularization parameters are data-driven. As our main result, we show that our estimators attain the minimax optimal convergence rate under the spectral norm for Gaussian stationary time series. The performance of our estimators is demonstrated in simulations and real data analysis.

C1130: **Bootstrap specication tests for GARCH processes with nuisance parameters on the boundary**
Presenter: **Indeewara Perera**, University of Sheffield, United Kingdom

Tests are developed for the correct specification of the conditional variance function in GARCH models when the true parameter may lie on the boundary of the parameter space. The test statistics considered are of Kolmogorov-Smirnov and Cramer-von Mises type, and are based on a certain empirical process marked by centered squared residuals. The limiting distributions of the test statistics depend on unknown nuisance parameters in a non-trivial way, making the tests difficult to implement. We, therefore, introduce a novel bootstrap procedure which is shown to be asymptotically valid under general conditions, irrespective of the presence of nuisance parameters on the boundary. The proposed bootstrap approach is based on shrinking the parameter estimates used to generate the bootstrap sample toward the boundary of the parameter space at a proper rate. It is simple to implement and fast in applications, as the associated test statistics have simple closed-form expressions. Although the bootstrap test is designed for a data-generating process with fixed parameters (i.e., independent of the sample size n), we also discuss how to obtain valid inference for sequences of DGPs with parameters approaching the boundary. A simulation study demonstrates that the new tests have excellent finite sample properties. Two data examples illustrate the implementation of the proposed tests in applications.

C1378: **Ordinal pattern-based time series analysis**
Presenter: **Annika Betken**, University of Twente, Netherlands
Co-authors: Herold Dehling, Alexander Schnurr, Jannis Buchsteiner, Jeannette Woerner, Ines Nuessgen

In time series analysis, ordinal patterns describe the relative position of consecutive data points generated by a stochastic process. Among other things, estimators are considered for the probabilities of occurrence of ordinal patterns (ordinal pattern probabilities) in time series. We investigate the statistical properties of these estimators in discrete-time Gaussian processes and establish limit theorems that describe the asymptotic distribution of the considered estimators. Additionally, we consider a measure of dependence between two-time series that is based on ordinal patterns (so-called ordinal pattern dependence). Ordinal pattern dependence can be considered as a non-parametric and non-linear counterpart to Pearson's correlation that allows for modeling leverage effects and dependence structures between financial time series.

C1383: **Estimation and prediction in misspecified fractionally integrated models with an unknown mean**
Presenter: **Kanchana Nadarajah**, University of Sheffield, United Kingdom
Co-authors: Gael Martin, Indeewara Perera, Donald Poskitt

The aim is to explore the impact of misspecification of the short memory dynamics on estimation and prediction in a fractionally integrated model with an unknown mean. In particular, we derive the limiting distributions of three parametric estimators, namely, exact Whittle, time-domain maximum likelihood, and the conditional sum of squares (CSS), under common misspecification of the short memory dynamics. We also show that, conditional on the use of a consistent estimator of the mean, these estimators converge to the same pseudo-true value and that their asymptotic distributions are identical to those of two alternative estimators that are mean invariant: the frequency domain maximum likelihood and discrete Whittle (DWH) estimators. We further derive the properties of a linear predictor under misspecification. We show that the linear predictor for zero-mean processes is biased, and the mean squared forecast error depends on the true and pseudo-true value of the fractional differencing parameter. A Monte Carlo simulation study shows that the DWH estimator of the pseudo-true value of the fractional differencing parameter has the best overall performance in terms of bias and mean squared error in finite samples, across a range of misspecification designs. In terms of finite sample forecast performance, DWH also exhibits the smallest forecast error and mean squared forecast error.

Monday 19.12.2022 16:50 - 18:30 Parallel Session Q – CFE-CMStatistics

EO667 Room S-2.25 FLEXIBLE BAYESIAN MODELLING FOR BIOSTATISTICS Chair: Tommaso Rigon

E0673: **Multiple hypothesis screening via mixtures of non-local distributions with applications to genomic datasets**
Presenter: **Francesco Denti**, Universita' Cattolica del Sacro Cuore, Italy

The analysis of large-scale datasets, especially in biomedical contexts, frequently involves a principled screening of multiple hypotheses. The celebrated two-group model jointly models the distribution of the test statistics with mixtures of two competing densities, the null and the alternative distributions. We investigate the use of non-local densities to specify alternative distributions that enforce separation from the null and thus refine the screening procedure. Parametric and nonparametric model specifications are proposed. With a simulation study, we exhibit how our model compares with both well-established and state-of-the-art alternatives in terms of various operating characteristics. Finally, to illustrate the versatility of our method, we show the results of differential expression analyses conducted on publicly-available datasets from genomic studies of heterogeneous nature.

E0835: **Tree-based models for high-dimensional compositional data in microbiome studies**
Presenter: **Zhuoqun Wang**, Duke University, United States

The human gut microbiome is associated with various diseases and health outcomes. A key characteristic of microbiome compositional data is its large and complex cross-sample heterogeneity. Appropriately accounting for these variance components is critical for several common inference tasks, including identifying latent structures, carrying out hypothesis testing on cross-group differences, and modeling dynamics, but is complicated by the key features of microbiome compositional data, including high-dimensionality, sparsity, and compositionality. These characteristics incur the need for structural constraints on covariance modeling while maintaining the analytical and computational tractability of the resulting models and methods. We present recently proposed methods that aim to utilize a tree structure – namely the phylogeny of the microbial species – to incorporate flexible covariance components while maintaining computational scalability. In particular, we present probabilistic models for microbiome compositional data based on the logistic-tree normal (LTN) distribution and demonstrate their wide applicability in a range of applications, including mixed-effects modeling, covariance estimation, and differential abundance analysis.

E1750: **Hierarchical modeling using Bayesian additive regression trees**
Presenter: **Vittorio Orlandi**, Duke University, United States
Co-authors: Alexander Volfovsky

Hierarchically structured data, in which units belong to various known or unknown groups, pervades many fields ranging from macroeconomics, to education, to medicine. Hierarchical models attempt to learn group-specific effects, relating units in the same group, while allowing for the sharing of information across groups. We propose using Bayesian Additive Regression Trees (BART), a flexible nonparametric method, to model such data. While BART has previously been used to model data with a group structure, past approaches treat group membership indicators no differently from covariates, which can lead to inefficient tree structures and high variance estimates. We show the advantages of our approach over this simpler alternative on simulated data and also present an extension of our method, based on a latent variable formulation, that addresses the case of unknown groups. We demonstrate this behavior in the United Network for Organ Sharing (UNOS) data.

E1377: **Assessing covariate balance in matched observational studies with high-dimensional categorical variables**
Presenter: **Massimiliano Russo**, Harvard Medical School, United States

Inferring causal effects in matched non-randomized studies requires that the exposure groups are approximately balanced. This implies that the two groups share a similar joint distribution of many confounding factors. There is vast literature on how to achieve such balance, but less attention has been devoted to assessing if the balance has been reached at different scales, including high-order interactions of the confounding factors. Focusing on the common case of multivariate categorical data, we describe a global test for balance that assesses if the joint probability mass function differs between the exposure groups, estimating a Bayes Factor via either an efficient Gibbs sampler or a variational approximation. We discuss explicit control of type-I error via a permutation scheme. If the imbalance is detected, local tests with explicit false discovery rate control are performed to detect which terms or interaction terms are responsible for the imbalance. We compare our methods with popular competitors, and discuss improved performance in simulations and real cohort studies.

EO635 Room S-1.04 RECENT ADVANCES IN HIGH-DIMENSIONAL INFERENCE Chair: Chao Zheng

E0559: **Multiple-splitting projection test for high dimensional mean vectors**
Presenter: **Xiufan Yu**, University of Notre Dame, United States

A multiple-splitting projection test (MPT) is proposed for one-sample mean vectors in high-dimensional settings. The idea of the projection test is to project high-dimensional samples to a 1-dimensional space using an optimal projection direction such that traditional tests can be carried out with projected samples. However, the estimation of the optimal projection direction has not been systematically studied in the literature. We bridge the gap by proposing a consistent estimation via regularized quadratic optimization. To retain the type I error rate, we adopt a data-splitting strategy when constructing test statistics. To mitigate the power loss due to data-splitting, we further propose a test via multiple splits to enhance the testing power. We show that the p-values resulting from multiple splits are exchangeable. Unlike existing methods which tend to combine dependent p-values conservatively, we develop an exact level α test that explicitly utilizes the exchangeability structure to achieve better power. Numerical studies show that the proposed test well retains the type I error rate and is more powerful than state-of-the-art tests.

E0838: **Multiple autocovariance change-points problems in high-dimensional time series**
Presenter: **Yuan Ke**, University of Georgia, United States

A framework is established to study multiple autocovariance change-points problems in high-dimensional, piecewise stationary, and heavy-tailed time series. First, we propose an element-wise truncated autocovariance estimator for high dimensional and nonstationary time series. We prove the estimator enjoys nice nonasymptptic and asymptotic properties when the time series data exhibits nonlinear temporal dependency and heavy-tailedness. Next, we introduce a moving sum statistic and a recursive segmentation algorithm to consistently detect the number and locations of autocovariance change-points in high-dimensional time series. The detection threshold in the algorithm is selected by a block-wise Gaussian multiplier bootstrap method. Further, we study the inference for the existence of a change-point around a pre-specified location and false discovery rate control for multiple autocovariance change-points detection. The superior empirical performance of the proposed methods is evaluated by various simulated and real data examples.

E0857: **Learning Graphical Model with Uniform Performance via Distributional Robust Optimization**
Presenter: **Youngseok Song**, EPFL, Switzerland
Co-authors: Wen Zhou

Learning large graphical models under distributional shifts such as unknown heavy-tailed contamination or latent heterogeneous sub-populations has become a major challenge in statistics and machine learning. Utilizing the Wasserstein-2 ball to define the shift of unknown data generation distribution from the true graph, we formulate this problem as a high-dimensional nodewise distributionally robust regression, whose computationally tractable dual problem is established as a square-root elastic net regression. We develop an iterative algorithm that lends a substantial advantage in

computational time. We study the finite sample upper bounds of the graph recovering against general distributional shifts, revealing the trade-off between the distributional robustness and the convergence rates. Extensive numerical experiments, together with real data applications, demonstrate the advantage of the proposed method compared to peer methods against a variety of distributional shifts.

E1865: **Generalised dynamic factor models for high dimensional spatio-temporal random fields on a network**
Presenter: **Chao Zheng**, University of Southampton, United Kingdom

High dimensional datasets containing records of spatio-temporal structures are of interest in many applications e.g., brain imaging, meteorology, and marketing research. We consider a dimensionality reduction technique using the generalized dynamic factors models. Different from the conventional approaches in time series where representation theory by using Wold's theorem and the concept of "innovations" is very natural, we have to take into account the spatial dependence where there is no unique definition of the concept of "innovation". To this end, we derive our generalized factor model by working on the spectral theory of random fields on 3D-grids. We established the rigorous definitions of multivariate space-time processes, its spectral representation theory, and a consistent spectral density estimator. Numerical and real-data examples are also provided.

EO516 Room S-1.06 ADVANCES IN SEMIPARAMETRIC ESTIMATION AND FINANCIAL DATA ANALYSIS Chair: Wendun Wang

E0279: **Recovering latent linkage structures and spillover effects with structural breaks in panel data models**
Presenter: **Wendun Wang**, Erasmus University Rotterdam, Netherlands

The aim is to capture time-varying spillover effects in a panel data setting. We consider panel models where the outcome of a unit depends not only on its own characteristics but also on the characteristics of other units (spillover effects). The effect of own characteristics can be unit-specific or homogeneous (common effects). We allow the linkage structure, i.e., which units interact with which, to be latent. Moreover, the structure and the spillover effects may both change at an unknown break point. To estimate the breakpoint, linkage structure, and spillover and common effects, we solve a penalized least squares optimization and employ double machine learning procedures to improve the convergence and inference. We establish the super consistency of the breakpoint estimator, which allows us to make inferences on other parameters as if the breakpoint was known. We illustrate the theory via simulated data.

E0311: **Flexible regularized estimating equations: Some new perspectives**
Presenter: **Yue Zhao**, University of York, United Kingdom
Co-authors: Archer Yang, Yuwen Gu, Jun Fan

Some observations about the equivalences between regularized estimating equations, fixed-point problems and variational inequalities are made: (a) A regularized estimating equation is equivalent to a fixed-point problem, specified via the proximal operator of the corresponding penalty; (b) A regularized estimating equation is equivalent to a (generalized) variational inequality. Both equivalences extend to any estimating equations with convex penalty functions. To solve large-scale regularized estimating equations, it is worth pursuing computation by exploiting these connections. While fast computational algorithms are less developed for regularized estimating equations, there are many efficient solvers for fixed-point problems and variational inequalities. In this regard, we apply some efficient and scalable solvers which deliver a hundred-fold speed improvement. These connections can lead to further research in both computational and theoretical aspects of the regularized estimating equations.

E0321: **Semiparametric efficiency in deep instrumental variable models**
Presenter: **Zuofeng Shang**, New Jersey Institute of Technology, United States

Endogeneity is fundamentally important as many empirical applications may suffer from the omission of explanatory variables, measurement error, or simultaneous causality. Recently, researchers propose a Deep Instrumental Variable (IV) framework based on deep neural networks to address endogeneity, demonstrating superior performances to existing approaches. We aim to understand the empirical success of the Deep IV theoretically. Specifically, we consider a two-stage estimator using deep neural networks in the linear instrumental variables model. By imposing a latent structural assumption on the reduced form equation between endogenous variables and instrumental variables, the first-stage estimator can automatically capture this latent structure and converge to the optimal instruments at the minimax optimal rate, which is free of the dimension of instrumental variables. Given that the network architectures are well chosen, we further show that the second-stage estimator achieves the semiparametric efficiency bound. In comparison with classical methods, the deep IV method does not require an explicit functional form and has a faster convergence rate, which is more computationally tractable in modeling the interactions between IVs. Simulation experiments on synthetic data and a real-world application will be provided.

E1559: **Beta sorted portfolio**
Presenter: **Weining Wang**, University of York, United Kingdom

Beta-sorted portfolios comprising assets with similar covariation to selected risk factors are a popular tool in empirical finance to analyze models of (conditional) expected returns. Despite their widespread use, little is known of their statistical properties in contrast to comparable procedures such as two-pass regressions. We formally investigate the properties of beta-sorted portfolio returns by casting the procedure as a two-step nonparametric estimator with a nonparametric first step and a beta-adaptive portfolio construction. The framework rationalizes the well-known estimation algorithm with precise economic and statistical assumptions on the general data-generating process and characterizes its key features. We study beta-sorted portfolios for both a single cross-section as well as for aggregation over time (e.g., the grandmean), offering conditions that ensure consistency and asymptotic normality along with new uniform inference procedures allowing for uncertainty quantification and testing of various relevant hypotheses in financial applications. We also highlight some limitations of current empirical practices and discuss what inferences can and cannot be drawn from returns to beta-sorted portfolios for either a single cross-section or across the whole sample. Finally, we illustrate the functionality of our new procedures in an empirical application

EO591 Room S-1.22 ADVANCES IN ECOLOGICAL STATISTICS Chair: Vianey Leos Barajas

E1064: **Bayesian causal inference in zero-inflated citizen science data**
Presenter: **Ben Swallow**, University of St Andrews, United Kingdom
Co-authors: Marie-Abele Bind

The use of causal inference in observational studies in ecology is an area of significant potential, with a need to understand drivers of changing environments associated with climate change and human influence. In order to allocate the observed changes directly to underlying causes, causal methods aim to construct a pseudo-experiment to emulate the conditions of a randomised trial. We present a method for Bayesian causal inference for data exhibiting zero-inflation based on a potential outcome formulation. The approach enables the estimation of an average treatment effect that can be allocated to the active treatment of interest. The method is applied to long-term data from UK citizen science surveys, in which zero-inflation is a regular problem, to enable the estimation of predator pressure on the abundances of a variety of avian species.

E1394: **Bayesian semi-supervised hidden Markov models for animal movement**
Presenter: **Vianey Leos Barajas**, University of Toronto, Canada

Hidden Markov models (HMMs) provide a flexible framework to model time series data where the observation process $Y[t]$ is taken to be driven by an underlying latent state process $Z[t]$, assumed to be a finite-state Markov chain. Applied to the study of animal movement, HMMs assume that the movements observed by an animal are the realizations of the animal's underlying (often unobserved) behavior. In this sense, the number

of states chosen is hoped to reflect the distinct numbers of behaviors that are able to be identified from the time series alone. However, in a fully unsupervised framework, the states can be, at best, taken to be proxies of the underlying behavioral process, and, at worst, a flexible framework to capture structure in the data without connecting to biological reality. To overcome this difficulty, we can move toward a semi-supervised framework where parts of the time series are labeled in a manner to validate the behavioral processes captured. Here we discuss how different labeling designs can improve both parameter and state estimation for HMMs applied to animal movement data.

E1428: **Incorporating body condition into the analysis of animal movement**
Presenter: **Marco Antonio Gallegos Herrada**, University of Toronto, Canada
Co-authors: Vianey Leos Barajas, Juan Morales

A long-sought goal in ecology is to connect movement with population dynamics. Especially for ungulates, there is a known link between conditions (e.g. fat reserves) and the probability of survival and reproduction. Assuming a particular genetic makeup and physiology, the condition reflects the history of behavioural decisions, including movement and habitat use. However, the condition of an animal can also have a direct implication on the types of movements that it performs and the habitats that it visits. Using Merino sheep as a case study, we present a model that allows for the interaction of movement and condition over time. For the movement dynamics, we use discrete-time, finite-state hidden Markov models (HMMs), with the positional data of the sheep serving as the observation process and the underlying state process serving as a proxy for behaviors of interest. To incorporate condition as a potential covariate affecting the movement, and thus behavioral, process, we make use of physiological equations that describe the evolution of body fat in order to predict daily values of the condition process, which are typically recorded once a month. The physiological equations are expressed as a function of the states inferred by HMM, as well as the distance that the sheep travels.

E1514: **Individual random effect models: Accounting for survivorship bias**
Presenter: **Ruth King**, University of Edinburgh, United Kingdom
Co-authors: Blanca Sarzo, Rachel McCrea

Survivorship bias arises when conclusions are drawn conditional on only the surviving individuals, whilst failing to correct for those individuals who have not survived. The issue has been well studied in many fields, such as economy, construction, forestry, health etc., but has been less well explored within the context of capture-recapture studies. We explore the implications related to survivorship bias that may arise in relation to individual heterogeneity models that are commonly fitted to capture-recapture data. The survivorship bias is manifested within these studies in that weaker individuals are more likely to die at a younger age compared to stronger individuals who may survive for longer within the study period. This implies that weaker individuals have a smaller probability of being observed within the study compared to stronger individuals, thus leading to an overestimate in the survival probabilities. We will initially discuss the impact of survivorship bias on associated survival probabilities within the common Cormack-Jolly-Seber model before describing how we can correct for this issue within capture-recapture studies when individuals are of known age when they are initially observed. To demonstrate the approach, we will initially consider simulated data, before applying the developed approach to data collected on ibex.

EO114 Room S-1.27 ADVANCES IN LONGITUDINAL DATA ANALYSIS Chair: Sanjoy Sinha

E0330: **Optimal designs in mixed ANCOVA models for longitudinal data**
Presenter: **Xiaojian Xu**, Brock University, Canada
Co-authors: Sanjoy Sinha

The construction of optimal designs for linear mixed models with covariates is investigated when involving longitudinal data. Random effects are employed to accommodate the clusters. We consider both the treatment effects as well as continuous covariates in the model. The goal of the designs is to optimally select the levels of covariates as well as the proportions of the sample units allocated to each treatment within a given total sample size. Both D- and A-optimality are chosen to be the design criteria. Although the estimators can be given with analytic forms if normality is assumed, the optimal designs depend on the unknown parameters involved in the variance components. Therefore, we apply both two-stage and sequential approaches. The problem of interest can be formulated in an ANCOVA framework, and the following specific problems are addressed: (i) optimal allocations for treatment groups if heteroscedastic random effects appear when the covariate levels are specified; (ii) optimal designing the levels of covariates for balanced design if random effects appear to be homoscedastic; and (iii) optimizing both the allocations for treatment groups and design levels for the covariates in the ANCOVA models with possible heteroscedasticity.

E0741: **Behavioural change models for disease transmission**
Presenter: **Rob Deardon**, University of Calgary, Canada

The COVID-19 pandemic has illustrated both the utility and limitation of using epidemic models for understanding and forecasting disease spread. One of the many difficulties in modelling epidemic spread is that caused by behavioural change in the underlying population. This can be a major issue in public health since, as we have seen during the COVID-19 pandemic, behaviour in the population can change drastically as infection levels vary, both due to government mandates and personal decisions. Such changes in the underlying population result in major changes in transmission dynamics of the disease, making the modelling challenges. However, these issues arise in agriculture and public health, as changes in farming practices are also often observed as disease prevalence changes. We propose a model formulation where time-varying transmission is captured by the level of alarm in the population and specified as a function of the past epidemic trajectory. The model is set in a data-augmented Bayesian framework, as epidemic data are often only partially observed, and we can utilize prior information to help with parameter identifiability. We investigate the estimability of the population alarm across a wide range of scenarios, using both parametric functions and non-parametric Gaussian processes and splines. The benefit and utility of the proposed approach are illustrated through an application to COVID-19 data from New York City.

E0716: **Clustering longitudinal matrix-variate count data**
Presenter: **Sanjeena Dang**, Carleton University, Canada

Three-way data structures or matrix-variate data are commonly generated in biological studies. In RNA sequencing, three-way data structures are obtained when high-throughput transcriptome sequencing data are collected for n genes at r conditions over p time points. Matrix variate distributions offer a natural way to model three-way data, and mixtures of matrix variate distributions can be used to cluster three-way data. Clustering of gene expression data is carried out as means of discovering gene co-expression networks. A family of a mixture of matrix variate Poisson-log normal distributions is introduced for clustering longitudinal read counts from RNA sequencing. By considering the matrix variate structure, the number of covariance parameters to be estimated is reduced, and the components of resulting covariance matrices provide a meaningful interpretation. To account for the longitudinal nature of the data, a modified Cholesky decomposition is utilized in the covariance structure. Furthermore, a parsimonious family of models are developed by imposing constraints on elements of these decompositions. The models are applied to both real and simulated data.

E0524: **Constrained inference in mixed models for clustered data**
Presenter: **Sanjoy Sinha**, Carleton University, Canada

Mixed models are commonly used for analyzing clustered data, including longitudinal data and repeated measurements. Unrestricted full maximum likelihood (ML) methods have been extensively studied in the literature for analyzing generalized, linear, and mixed models. However, constraints or parameter orderings may occur in practice. In such cases, we can improve the efficiency of a statistical method by incorporating parameter constraints into ML estimation and hypothesis testing. We will discuss constrained inference with generalized linear mixed models (GLMMs)

under linear inequality constraints. Methods will be assessed using both Monte Carlo simulations and actual survey data from a health study.

EO160 Room K0.16 NOVEL APPROACHES ON MODELING AND INFERENCE OF NETWORK DATA Chair: Wen Zhou

E0471: **Root and community inference on Markovian models of networks**
Presenter: **Min Xu**, Rutgers University, United States

Preferential attachment (PA) is a popular way of modeling random networks in which the network starts as a single node which we call the root node, and at every new time step, a new node and new edges are added to the network; this dynamic captures the growth/recruitment process that underlies many real-world networks. Given only a single snapshot of the final network G, we study the problem of constructing confidence sets for the early history, in particular the root node, of the unobserved growth process; the root node can be patient zero in a disease infection network or the source of fake news in a social media network. In the case where the graph is a PA tree contaminated with random noise edges, we propose an inference algorithm based on Gibbs sampling that scales to networks with millions of nodes and provide a theoretical analysis showing that the expected size of the confidence set is small, so long as the noise level is not too large. We also propose variations of the model in which multiple growth processes occur simultaneously, reflecting the growth of multiple communities, and we use these models to provide a new approach to community detection.

E1318: **Identification and estimation of network statistics with missing link data**
Presenter: **Matthew Thirkettle**, Rice University, United States

Informative bounds are obtained on network statistics in a partially observed network whose formation is explicitly modeled. Partially observed networks are commonplace due to, for example, partial sampling or incomplete responses in surveys. Network statistics (e.g., centrality measures) are not point identified when the network is partially observed. Worst-case bounds on network statistics can be obtained by letting all missing links take values zero and one. We dramatically improve on the worst-case bounds by specifying a structural model for network formation. An important feature of the model is that we allow for positive externalities in the network-formation process. The network-formation model and network statistics are set identified due to multiplicity of equilibria. We provide a computationally tractable outer approximation of the joint identified region for preferences determining network-formation processes and network statistics. In a simulation study on Katz-Bonacich centrality, we find that worst-case bounds that do not use the network formation model are 44 times wider than the bounds we obtain from my procedure.

E1626: **Identification and estimation of network models with nonparametric unobserved heterogeneity**
Presenter: **Andrei Zeleneev**, University College London, United Kingdom

Homophily based on observables is widespread in networks. Therefore, homophily based on unobservables (fixed effects) is also likely to be an important determinant of the interaction outcomes. Failing to properly account for latent homophily (and other complex forms of unobserved heterogeneity, in general) can result in inconsistent estimators and misleading policy implications. To address this concern, we consider a network model with nonparametric unobserved heterogeneity, leaving the role of the fixed effects and the nature of their interaction unspecified. We argue that the outcomes of the interactions can be used to identify agents with the same values as the fixed effects. The variation in the observed characteristics of such agents allows me to identify the effects of the covariates, while controlling for the impact of the fixed effects. Building on these ideas, we construct several estimators of the parameters of interest and characterize their large sample properties. The suggested approach is not specific to the network context and applies to general two-way models with nonparametric unobserved heterogeneity, including large panels. A Monte-Carlo experiment illustrates the usefulness of the suggested approaches and supports the large sample theory findings.

E1785: **Trading off accuracy for speedup: Multiplier bootstraps for subgraph counts**
Presenter: **Purnamrita Sarkar**, U. T. Austin, United States

A new class of multiplier bootstraps is proposed for count functionals, ranging from a fast, approximate linear bootstrap tailored to sparse, massive graphs to a quadratic bootstrap procedure that offers refined accuracy for smaller, denser graphs. For the fast, approximate linear bootstrap, we show that $\sqrt{n}$-consistent inference of the count functional is attainable in certain computational regimes that depend on the sparsity level of the graph. Furthermore, even in more challenging regimes, we prove that our bootstrap procedure offers valid coverage and vanishing confidence intervals. For the quadratic bootstrap, we establish an Edgeworth expansion and show that this procedure offers higher-order accuracy under appropriate sparsity conditions. We complement our theoretical results with a simulation study and real data analysis and verify that our procedure offers state-of-the-art performance for several functionals.

EO684 Room K0.18 CAUSAL INFERENCE AND MACHINE LEARNING Chair: Oliver Dukes

E1113: **Generalizing treatment effects with incomplete covariates**
Presenter: **Imke Mayer**, Charite Universitaetsmedizin Berlin, Germany
Co-authors: Julie Josse

The focus is on the problem of generalizing a causal effect estimated on a randomized controlled trial (RCT) to a target population described by a set of variables from observational data. Available methods such as inverse propensity sampling weighting are not designed to handle missing values, which are, however, common in both data sources. In addition to coupling the assumptions for causal effect identifiability and the mechanism of the missing value and to defining appropriate estimation strategies, one difficulty to consider is the specific structure of the multi-source data with only partial information on treatment and outcome. We propose multiple imputation strategies to handle missing values when generalizing treatment effects, each handling the multi-source structure of the problem differently. As an alternative, we also propose a machine learning-based estimation approach that treats incomplete covariates as semi-discrete variables. The proposed strategies rely on different sets of assumptions concerning the impact of missing values on identifiability. We discuss these assumptions and assess the methods through an extensive simulation study, as well as on a large major trauma registry and an RCT to study the effect of the drug tranexamic acid on mortality in major trauma patients admitted to ICU. This analysis illustrates how the handling of the missing value can impact the conclusion about the effect generalized from the RCT to the target population.

E1335: **Causal inference and dynamic treatment rule estimation based on contrast-approximating linear models**
Presenter: **David Whitney**, London School of Hygiene and Tropical Medicine, United Kingdom
Co-authors: Stijn Vansteelandt, Karla DiazOrdaz

Optimal treatment rules, which assign treatment based on subject characteristics in a way that optimizes expected outcomes, are of widespread interest in statistics, economics, engineering, and other fields. Approaches to estimating optimal treatment rules include both data-adaptive machine learning and model-based methods. Machine learning approaches, while powerful, have been criticized for being less accessible to non-specialist audiences. Additionally, many machine learning approaches are tailored to specific types of treatment (e.g. binary or continuous). Model-based strategies can mitigate these criticisms but do so at the expense of strong modelling assumptions that are unlikely to hold in practice. To address these concerns, we propose a flexible framework that treats modelling the data-generating process as distinct from defining a parsimonious and interpretable rule. Each treatment rule in our proposal corresponds to a working structural mean model. Our estimator of the model coefficients allows for machine learning of nuisance parameters and accommodates any type of treatment. If the working model is correctly specified, then the corresponding treatment rule is optimal. We illustrate the finite sample performance of our proposal relative to other methods for estimating optimal treatment regimes in simulation and real-world data applications.

E1435: **The projected covariance measure for assumption-lean variable significance testing**
Presenter: **Rajen D Shah**, University of Cambridge, United Kingdom
Co-authors: Anton Rask Lundborg, Ilmun Kim, Richard Samworth

Testing the significance of a variable or group of variables X for predicting a response Y given additional covariates Z, is a ubiquitous task in statistics. A simple but common approach is to specify a linear model and test whether the X regression coefficient is non-zero. However, when the model is misspecified, as will invariably be the case, the test may have poor power, for example, when X is involved in complex interactions, or lead to many false rejections. We study the problem of testing the model-free null of conditional mean independence, i.e. that the conditional mean of Y given X and Z does not depend on X. We propose a simple and general framework that can leverage flexible nonparametric or machine learning methods, such as additive models or boosted trees, to yield both robust error control and high power. The procedure involves using these methods to perform regressions, first to estimate a form of projection of Y on X and Z using one-half of the data, and then to estimate the expected conditional covariance between this projection and Y on the remaining half of the data. While the approach is general, we show that a version of our procedure using spline regression achieves what we show is the minimax optimal rate in this nonparametric testing problem.

E1936: **The synthetic instrument**
Presenter: **Linbo Wang**, University of Toronto, Canada
Co-authors: Dingke Tang, Dehan Kong, Linbo Wang

In many observational studies, researchers are interested in studying the effects of multiple treatments on the same outcome. Unmeasured confounding is a key challenge in these studies as it may bias the causal effect estimate. To mitigate this bias, we introduce a novel device, called synthetic instrument, to leverage the information contained in multiple treatments for causal effect identification and estimation. We show that under linear structural equation models, the problem of causal effect estimation can be formulated as an ℓ_0 penalization problem, and hence can be solved efficiently using off-the-shelf software. Simulations show that our approach outperforms state-of-art methods in both low-dimensional and high-dimensional settings. We further illustrate our method using a mouse obesity dataset.

EO590 Room K0.19 ADVANCES IN DESIGN-BASED CAUSAL INFERENCE Chair: Nicole Pashley

E0233: **Sensitivity analysis for null results: Implications for studies of racially biased policing**
Presenter: **Jake Bowers**, University of Illinois @ Urbana-Champaign, United States
Co-authors: Thomas Leavitt, Luke Miratrix

A method of formal sensitivity analysis is proposed for causal inference that addresses the problem of null results: a null result in an observational study is no more or less likely to emerge because of hidden confounding than a strong result. The motivation comes from the problem of null results in the study of the causal effects of race of civilians on police use of force and shows how it adds to existing critiques of null results. We show how in a small simulated dataset, a pattern of hidden confounding and a pattern of post-treatment missingness like that seen in datasets used to study race and police can combine to produce a misleading null effect. And we show how our method of sensitivity analysis for null effects reveals that the null result is, in fact, sensitive to this kind of bias. We compare both an approach for tests of the weak null of no average effects and an approach for the strong null of no effects.

E1080: **Power and sample size calculations for rerandomized experiments**
Presenter: **Zach Branson**, Carnegie Mellon University, United States
Co-authors: Xinran Li, Peng Ding

Power analyses are an important aspect of experimental design, because they help determine how experiments are implemented. It is common to specify a desired level of power and compute the sample size necessary to obtain that power. Such calculations are well-known for completely randomized experiments, but there can be many benefits to using other experimental designs. For example, it has recently been established that rerandomization, where subjects are randomized until covariate balance is obtained, increases the precision of causal effect estimators. This work establishes the power of rerandomized treatment-control experiments, thereby allowing for sample size calculators. We find the surprising result that, while power is often greater under rerandomization than complete randomization, the opposite can occur for very small treatment effects. The reason is that inference under rerandomization can be relatively more conservative than complete randomization, in the sense that it can have a lower Type-I error, and this additional conservativeness adversely affects power. This surprising result is due to treatment effect heterogeneity, a quantity often ignored in power analyses. We find that heterogeneity increases power for large effect sizes but decreases power for small effect sizes.

E0984: **A design-based Riesz representation framework for randomized experiments**
Presenter: **Fredrik Savje**, Yale University, United States
Co-authors: Christopher Harshaw, Yitan Wang

A new design-based framework is described for drawing causal inference in randomized experiments. Estimands in the framework are defined as arbitrary linear functionals of the potential outcome functions. This makes the framework expressive, allowing experimenters to formulate and investigate a wide range of causal questions. We describe a class of estimators for estimands defined using the framework and investigate their properties. The construction of the estimators is based on insights from the Riesz representation theorem. We provide necessary and sufficient conditions for unbiasedness and consistency. Finally, we provide conditions under which the estimators are asymptotically normal, and describe a conservative variance estimator to facilitate inference about the estimands.

E0851: **Analyzing randomized experiments subject to outcome misclassification via integer programming**
Presenter: **Siyu Heng**, New York University, United States
Co-authors: Pamela Shaw

Results from randomized experiments (trials) can be severely distorted by outcome misclassification, such as from measurement error or reporting bias in binary outcomes. All existing approaches to outcome misclassification rely on some data-generating (super-population) model and, therefore, may not be applicable to randomized experiments without additional assumptions. We propose a model-free and finite-population-exact framework for randomized experiments subject to outcome misclassification. A central quantity in our framework is "warning accuracy," defined as the threshold such that the causal conclusion drawn from the measured outcomes may differ from that based on the true outcomes if the outcome measurement accuracy did not surpass that threshold. We show how learning the warning accuracy and related concepts can benefit a randomized experiment subject to outcome misclassification. We show that the warning accuracy can be computed efficiently (even for large datasets) by adaptively reformulating an integer program with respect to the randomization design. Our framework covers both Fisher's sharp null and Neyman's weak null, works for a wide range of randomization designs, and can also be applied to observational studies adopting randomization-based inference. We apply our framework to a large randomized clinical trial for the prevention of prostate cancer.

EO338 Room K0.20 ADVANCES IN ANALYZING AND MODELING COMPLEX HIGH DIMENSIONAL DATA Chair: Wenbo Wu

E0389: **Shifting-corrected regularized regression model for NMR metabolomic identification**
Presenter: **Thao Vu**, University of Colorado, United States

The process of identifying metabolites in complex mixtures plays a critical role in metabolomic studies to obtain an informative interpretation

of underlying biological processes. Manual approaches are time-consuming and heavily reliant on the knowledge and assessment of nuclear magnetic resonance (NMR) experts. We propose a shifting-corrected regularized regression method, which identifies metabolites in a mixture automatically. Using a novel weight function, the proposed method is able to detect and correct peak shifting errors caused by fluctuations in experimental procedures. Simulation studies show that the proposed method performs better with regard to the identification of metabolites in a complex mixture. We also demonstrate real data applications of our method using experimental and biological NMR mixtures.

E0406: **Inverse probability weighting-based mediation analysis for microbiome data**
Presenter: **Yuexia Zhang**, The University of Texas at San Antonio, United States
Co-authors: Linbo Wang, Jianhua Hu, Jian Wang, Jiayi Shen, Jessica Galloway-Pena, Samuel Shelburne
Mediation analysis is an important tool for studying causal associations in biomedical and other scientific areas and has recently gained attention in microbiome studies. Using a microbiome study of acute myeloid leukemia (AML) patients, we investigate whether the effect of induction chemotherapy intensity levels on the infection status is mediated by microbial taxa abundance. The unique characteristics of the microbial mediators— high dimensionality, zero inflation, and dependence—call for new methodological developments in mediation analysis. The presence of an exposure-induced mediator-outcome confounder, antibiotic use, further requires a delicate treatment in the analysis. To address these unique challenges in our motivating AML microbiome study, we propose a novel nonparametric identification formula for the interventional indirect effect (IIE), a measure recently developed for studying mediation effects. We develop the corresponding estimation algorithm using the inverse probability weighting method. We also test the presence of mediation effects via constructing the standard normal bootstrap confidence intervals. Simulation studies show that the proposed method has good finite-sample performance in terms of the IIE estimation, and type-I error rate and power of the corresponding test. In the AML microbiome study, our findings suggest that the effect of induction chemotherapy intensity levels on infection is mainly mediated by patients' gut microbiome.

E0957: **Online data selection and sparse estimation for multivariate streaming data**
Presenter: **Rui Xie**, Universify of Central Florida, United States
Co-authors: Shuyang Bai, Yongkai Chen, Ping Ma
Real-time analysis of large-scale streaming multivariate data often faces a trade-off between statistical estimation efficiency and computational cost efficiency. For multivariate data streams, one needs to carefully balance the trade-off, especially for sparse and possibly under-determined regression problems, which require more computational efforts. Data selection enables one to process large-scale streaming data in real time, so one can fit and update the sparse model in seconds instead of hours. We study the online real-time joint data-dependent sample selection and continuous variable selection for a multi-dimensional spare regression problem for streaming data. We propose a class of online data selection methods that simultaneously achieve sampling and sparse estimation to improve the computational efficiency of the online analysis. The online sparse model estimation involves using coordinate descent algorithms for nonconvex penalized regression, and the real-time data selection adapts optimal design-based sequential online sampling. The performance of the sampling-assisted online sparse estimation method is assessed via simulation studies and real data examples.

E1206: **Dimension reduction with expectation of a conditional difference measure**
Presenter: **Wenhui Sheng**, Marquette University, United States
A flexible model-free approach is introduced to sufficient dimension reduction analysis using the expectation of a conditional difference measure. Without any strict conditions, such as linearity condition or constant covariance condition, the method estimates the central subspace effectively under linear or nonlinear relationships between response and predictors. The method is especially useful when the response is categorical. We also studied the root-n consistency and asymptotic normality properties of the estimates. The efficacy of our method is demonstrated through both simulations and real data analysis.

EO040 Room K0.50 HIGHLIGHTS OF CONTEMPORARY RESULTS IN DESIGN OF EXPERIMENTS Chair: Stefanie Biedermann

E1082: **Scale invariant optimal subsampling**
Presenter: **HaiYing Wang**, University of Connecticut, United States
Subsampling is an effective method to alleviate the computational cost when faced with massive data, and optimal subsampling algorithms aim to achieve a higher estimation efficiency. Existing optimal subsampling probabilities focus on minimizing the asymptotic mean squared error of the subsample parameter estimator. They are scale variant, and their performance changes if the data is scale transformed. We recommend focusing on minimizing the squared prediction error, which results in scale-invariant optimal subsampling probabilities. In addition, the resulting probabilities are invariant to model constraints in softmax regression, and they provide a better subsampling strategy than existing methods in terms of balancing the responses among all categories.

E1135: **Optimal design of experiments on Riemannian manifolds**
Presenter: **Hang Li**, Astrazeneca, United States
The theory of optimal design of experiments has been traditionally developed in Euclidean spaces. New theoretical results and an algorithm for finding the optimal design of an experiment located on a Riemannian manifold are provided. It is shown that analogously to the results in Euclidean spaces, *D*-optimal and *G*-optimal designs are equivalent on manifolds, and we provide a lower bound for the maximum prediction variance of the response evaluated over the manifold. In addition, a converging algorithm that finds the optimal experimental design on manifold data is proposed. Numerical experiments demonstrate the importance of considering the manifold structure in a designed experiment when present, and the superiority of the proposed algorithm.

E1444: **An integrated approach to test for missingness not at random**
Presenter: **Robin Mitra**, University College London, United Kingdom
Co-authors: Jack Noonan, Stefanie Biedermann
Missing data is known to be an inherent and pervasive problem in the process of data collection. The effects are wide-ranging and the loss of data can lead to inefficiencies and introduce bias into analyses. The specific problem of data missing not at random (MNAR) is known to be one of the most complex and challenging problems to handle in this area and testing its prevalence is of great importance. The presence of MNAR missingness can only be tested using a follow-up sample of the missing observations and therefore recovering a proportion of missing values in an efficient way could be crucial in saving the experimenter's costs and time and may result in new treatments/technology reaching the public faster. We develop a strategy to allow researchers to be in a position to be well informed about whether MNAR is a credible issue. Within a multiple regression setting, we demonstrate a proof of concept example and provide recommendations for how the follow-up sample of missing observations should be designed.

E1490: **Replication of partial-profile choice designs: Factor permutation as an alternative to simple repetition**
Presenter: **Heiko Grossmann**, Otto-von-Guericke-University Magdeburg, Germany
For design problems in linear models where a finite group of transformations acts transitively on a finite design space, it is well known that for convex optimality criteria which are invariant under the group, optimal approximate designs can be constructed by symmetrizing a given design. The underlying ideas can also be used to address some practical issues which arise in the area of partial-profile discrete choice experiments. In these experiments, there exist potentially many qualitative factors, of which only a subset is used in each question of a choice questionnaire. Certain

exact designs for these experiments possess a high efficiency but are rigid in the sense that only relatively few of all possible subsets of the factors are used. When using such a design as the basis for a survey, where the number of potential respondents would allow several replications of the design, simply repeating the rigid base design does not seem to be advisable. Instead, we propose to use replications where each replication of the design uses a different permutation of the factors. For the rigid base designs we consider, this approach leads to replicated designs with better coverage of the design space and higher statistical efficiency. Moreover, the replicated designs appear to be robust against efficiency losses due to non-response. We illustrate the general ideas by referring to a design from an actual choice experiment.

EO418 Room S0.03 SPATIO(-TEMPORAL) MODELING FOR BIOMEDICAL AND ENVIRONMENTAL DATA Chair: Rajarshi Guhaniyogi

E0777: **BAMDT: Bayesian additive semi-multivariate decision trees for spatial nonparametric regression**
Presenter: **Huiyan Sang**, Texas A&M University, United States
Co-authors: Zhao Tang Luo, Bani Mallick

Bayesian additive regression trees have gained great popularity as flexible nonparametric function estimation and modeling tools. Nearly all existing BART models rely on decision tree-weak learners with axis-parallel univariate split rules to partition the Euclidean feature space into rectangular regions. In practice, however, many regression problems involve features with multivariate structures (e.g., spatial locations) possibly lying in a manifold, where rectangular partitions may fail to respect irregular intrinsic geometry and boundary constraints of the structured feature space. We develop a new class of Bayesian additive multivariate decision tree models that combine univariate split rules for handling possibly high dimensional features without known multivariate structures and novel multivariate split rules for features with multivariate structures in each weak learner. The proposed multivariate split rules are built upon stochastic predictive spanning tree bipartition models on reference knots, which are capable of achieving highly flexible nonlinear decision boundaries on manifold feature spaces while enabling efficient dimension reduction computations. We demonstrate the superior performance of the proposed method using simulation data and a Sacramento housing price data set.

E0865: **General correlated statistical count structures**
Presenter: **Robert Lund**, The University of California, Santa Cruz, United States

Methods are considered capable of generating a count-valued time series, a spatial random field, or a spatio-temporal random process having any prescribed marginal distribution. A Gaussian copula is used to transform a correlated Gaussian process into the desired count structure. The methods are shown to have the most general autocovariance structure achievable, permit any marginal distribution whatsoever, and can easily accommodate covariates. Hermite expansions are used to relate the autocovariance of the Gaussian process to that of the count process. Particle filtering methods of likelihood evaluation are explored.

E0986: **Land-use Filtering for Nonstationary Spatial Prediction of Collective Efficacy in An Urban Environment**
Presenter: **Brandon Carter**, University of Texas at Austin, United States
Co-authors: Catherine Calder, Christopher Browning, Bethany Boettner, Nicolo Pinchak

Collective efficacy - the capacity of communities to exert social control toward the realization of their shared goals - is a foundational concept in urban sociology and neighborhood effects literature. Traditionally, empirical studies of collective efficacy use large sample surveys to estimate the collective efficacy of different neighborhoods within an urban setting. Such studies have demonstrated an association between collective efficacy and local variation in community violence, educational achievement, and health. Unlike traditional collective efficacy measurement strategies, the Adolescent Health and Development in Context (AHDC) Study implemented a new approach, obtaining spatially-referenced, place-based ratings of collective efficacy from a representative sample of individuals residing in Columbus, OH. We introduce a novel nonstationary spatial model for interpolation of the AHDC collective efficacy ratings across the study area, which leverages administrative data on land use. Our constructive model specification strategy involves dimension expansion of a latent spatial process and the use of a filter defined by the land-use partition of the study region to connect the latent multivariate spatial process to the observed ordinal ratings of collective efficacy. Careful consideration is given to the issues of parameter identifiability, computational efficiency of an MCMC algorithm for model fitting, and fine-scale spatial prediction of collective efficacy.

E1990: **Quantifying uncertainty for spatial predictions: Fast algorithms for large data sets**
Presenter: **Douglas Nychka**, Colorado School of Mines, United States

A benefit of a Gaussian process (GP) model for surface fitting is the companion estimates of the functions' uncertainty. The standard method for assessing the uncertainty of a GP estimate is through conditional simulation, a Monte Carlo sampling algorithm of the multivariate Gaussian distribution. Conditional simulation is a powerful tool, for example allowing for Monte Carlo based uncertainty on surface contours (level sets), a difficult and nonlinear inference problem. Thus, it serves as the basic strategy for nearly all applications to spatial and spatial-temporal inference. This algorithm, however, is limited for large data sets. Accurate approximations are proposed that allow for fast computation. The computational efficiency is achieved by relying on the fast Fourier transform for 2D convolution and also sparse matrix multiplication. Under common spatial applications, a speedup by a factor of 10 to 100 or more is obtained and makes it possible to determine the uncertainty of GP estimates on a laptop and often in an interactive session. Besides the practical benefits of this speedup, the two approximations are examples of interesting features of GP analysis. Namely exploiting the screening effect for spatial prediction and the error bounds in interpolation when the GP is related to an element in a reproducing kernel Hilbert space.

EO328 Room S0.11 NONPARAMETRIC HIGH-DIMENSIONAL STATISTICAL LEARNING Chair: Chenlu Ke

E0410: **Conditional multidimensional scaling**
Presenter: **Anh Bui**, Virginia Commonwealth University, United States

The focus is on the problem of mapping high-dimensional data to a low-dimensional space, in the presence of other known features. This problem is ubiquitous in science and engineering as there are often controllable/measurable features in most applications. Furthermore, the discovered features in previous analyses can become known features in subsequent analyses, repeatedly. To solve this problem, a broad class of methods, which is referred to as conditional multidimensional scaling, is proposed. An algorithm for optimizing the objective function of conditional multidimensional scaling is also developed. The proposed framework is illustrated with kinship terms, facial expressions, and simulated car-brand perception examples. These examples demonstrate the benefits of the framework for being able to marginalize out the known features to uncover unknown, unanticipated features in the reduced-dimension space and for enabling a repeated, more straightforward knowledge discovery process. Computer codesare available in the open-source cml R package.

E0975: **Nonparametric mixture model: Application in contaminated trials**
Presenter: **Zi Ye**, Lehigh University, United States

In personalized medicine, investigating the differential effect of treatments in groups defined by patient characteristics is of paramount importance. In a randomized clinical trial, participants are first classified using diagnostic tools, but such classifiers may not be perfectly accurate. The issue of diagnostic misclassification has recently become prominent and has produced severely biased estimations of treatment effects. The focus is on this problem in a pre-stratified randomized placebo-controlled repeated measures design. We develop a fully nonparametric method for estimating and testing the treatment effect for ordinal, discrete, or skewed outcomes. Consistent estimators and asymptotic distributions are provided for the

misclassification error rates as well as the treatment effect. Simulation studies are conducted to compare the new method with traditional methods. The results show significant advantages of the proposed methods regarding bias reduction, coverage probability, and power.

E1457: **Nonconvex-regularized integrative sufficient dimension reduction for multi-source Data**
Presenter: **Wei Qian**, University of Delaware, United States
Co-authors: Shanshan Ding

As advances in high-throughput technology significantly expand data availability, integrative analysis of multiple data sources has become an increasingly important tool for biomedical studies. We propose an integrative and nonconvex-regularized sufficient dimension reduction method to achieve simultaneous dimension reduction and variable selection for multi-source data analysis in high dimensions. The proposed method aims to extract sufficient information in a supervised fashion, and our asymptotic results establish new theory for integrative sufficient dimension reduction and allow the number of predictors in each data source to increase exponentially fast with sample size. The promising performance of the integrative estimator and efficient numerical algorithms is demonstrated through simulation and multi-omics breast cancer data analysis.

E1141: **Dimension reduction for spatial regression: The spatial predictor envelope**
Presenter: **Hossein Moradi Rekabdarkolaee**, South Dakota State University, United States

Natural sciences such as geology and forestry often utilize regression models for spatial data with high-dimensional predictors and moderate sample sizes. In this case, efficient estimation of the regression parameters is crucial for both model interpretation and prediction. The predictor envelope is a method of dimension reduction for linear regression with multivariate predictors that assumes certain linear combinations of the predictors are immaterial to the regression. The method can result in substantial gains in estimation efficiency and prediction accuracy over traditional maximum likelihood and least squares estimates. While predictor envelopes have been developed and studied for independent data, no work has been done adapting predictor envelopes to spatial data. The predictor envelope is adapted to a popular spatial model to form the spatial predictor envelope. Maximum likelihood estimates for the SPE are derived, along with asymptotic distributions for the estimates given certain assumptions, showing the SPE estimates to be asymptotically more efficient than generalized least squares, the typical spatial regression estimates. Further, we study the SPE in the context of spatial prediction, or universal kriging, discussing the contexts in which the SPE can provide gains over the typical universal kriging predictions. The effectiveness of the proposed model is illustrated through simulation studies and real data analysis.

EO046 Room S0.12 STATISTICS IN NEUROSCIENCE II Chair: Jeff Goldsmith

E1449: **AI for organoids and organoids for AI**
Presenter: **Brian Caffo**, Johns Hopkins University, United States

Unsupervised methods are discussed for studying brain organoids and functional neural systems. Particularly, we consider the important role that parsimony can play in non-parsimonious decompositions and non-linear embeddings. We consider a study of neurogenesis including in vivo, in vitro, single cell and bulk RNA sequencing. We contrast several methods for joint decompositions that share information across experiment and tissue types and contrast results in novel experiments not used in model training. We discuss the role that in vitro neural systems can play in performing AI tasks. In this, we use multi-electrode arrays (MEAs) to study functioning brain organoids. Such use of organoids for biocomputing is a nascent and exciting field that we refer to as organoid intelligence.

E1496: **Spectral Granger causality using neural networks for biological signals**
Presenter: **Malik Shahid Sultan**, King Abdullah University of Science and Technology, Saudi Arabia
Co-authors: Samuel Horvath, Hernando Ombao

Granger Causality (GC) between channels of electroencephalograms (EEG) will be investigated. Since brain signals are complex, we expect a non-linear dependence structure; therefore, vector autoregressive (VAR) models may not completely characterize GC in the brain networks. To address this limitation, we shall apply deep learning (DL) tools which can learn the non-linear dependence structure in the data. However, these models are inherently black boxes and difficult to interpret. We shall use the learned kernel vector autoregressive (LeKVAR) model, component-wise multi-layer perceptron (cMLPwF), and component-wise long short-term memory (cLSTMwF) proposed in Horvath et al. We demonstrate that these models can learn the non-linear frequency band-specific dependence structure in the time series data and give an estimate of the GC through the filter layer and decoupling lags and time series. We identify GC between and across different signals based on the decomposed spectrum, which gives an insight into the GC between oscillations. We estimate GC using LeKVAR, cLSTMwF, and cMLPwF for the spectral decomposed data of an epilepsy patient and study the evolution of the GC between EEG electrodes pre, during, and post-seizure.

E1669: **Spatial distribution of white matter hyperintensities is related to cognition in MCI**
Presenter: **Jordan Dworkin**, Columbia University, United States
Co-authors: Elizabeth Sweeney

The detection and characterization of white matter hyperintensities (WMH) on structural MRI is important across several neurological contexts. In multiple sclerosis, aging, and Alzheimer's disease, quantification has typically focused on whole-brain summaries of the total volume of the affected tissue. Recently, efforts to uncover more informative clinico-radiological relationships have motivated consideration of hyperintensities spatial distribution throughout the brain. Prior studies in stroke and traumatic brain injury have shown that connectome-based lesion-symptom mapping - a framework for quantifying lesions' spatial impacts by measuring the structural connections they are likely to impact - provides clinically relevant information. In aging and Alzheimer's disease, however, it is currently unknown (A) whether the spatial distribution of white matter hyperintensities is relevant for clinical outcomes, and (B) what level of spatial granularity best captures the relevant signal. Here, we conduct a comparison study to determine whether, and to what extent, spatial information improves prediction of cognitive performance. Cross-validated elastic net regression reveals that in healthy aging, spatial information is unrelated to cognitive performance when accounting for total WMH volume. Notably, spatial information significantly improves prediction in participants with mild cognitive impairment, with network-level dysconnectivity performing best among the tested models.

E1692: **Video segmentation and functional data pipeline for assessing pupil changes due to cannabis consumption**
Presenter: **Julia Wrobel**, Colorado School of Public Health, United States

Research findings on the impact of acute cannabis use on pupillary size have been inconsistent. We developed a video processing and analysis pipeline that extracts pupil sizes from videos obtained during a light stimulus test administered with goggles utilizing infrared videography. The light stimulus test was administered for those with occasional ($N = 36$), daily ($N = 33$), and no cannabis use ($N = 32$), and before and after acute cannabis smoking or a waiting period for the no-use group. Pupils were segmented using a combination of image pre-processing techniques and segmentation algorithms and linear least-squares ellipse fitting estimated pupil size. The segmentation pipeline achieved 99 percent precision on a validating set of images ($N = 517$). Pupil size trajectories in response to the light stimulus were then analyzed across marijuana use groups using techniques from functional data analysis (FDA). Our results show that FDA-based regression models are more sensitive to differences across marijuana use groups than using scalar features extracted from the pupil trajectories. In addition, we find that acute cannabis use from both occasional and daily use groups results in less pupil constriction and slower rebound dilation in the light stimulus test.

EO210 Room S0.13 MODERN STATISTICAL METHODS WITH APPLICATIONS TO COMPLEX DATA ANALYSIS Chair: Yichuan Zhao

E1916: **Disease prediction by detecting and integrating connectomic networks and marginally weak signals**
Presenter: **Yanming Li**, University of Kansas Medical Center, United States

Many contemporary studies use individual genomic or imaging profiles for early prediction of cancer or neuropsychological outcomes, such as cancer subtypes and Alzheimer's disease stages. Current approaches ignore the connection structures of the genome and the brain (e.g. gene pathways or brain networks). Despite having marginally weak effects, many genetic and imaging markers may exude strong predictive effects once considered together with their connected biomarkers. To find such weak signals, the inter-feature connectomic structure of the genome or brain must be explored first. However, given the ultrahigh-dimensional characteristic of genomic/neuroimaging profiles, identifying the whole genome/brain connectomic features is computationally prohibitive. This is also an impediment to detecting weak signals. We hypothesize that a large portion of the predictiveness of disease outcomes is attributed to inter-marker connections and marginally weak signals. By detecting and integrating them, prediction accuracy can be significantly improved. We develop novel statistical/machine-learning algorithms for detecting network-based biomarkers for cancer or AD-related outcome prediction. The identified network signatures and weak signals will also enhance our understanding of the underlying mechanisms of disease development and progression.

E1698: **Nonsmooth low-rank matrix recovery: Methodology, theory and algorithm**
Presenter: **Peng Liu**, University of Kent, United Kingdom

Many interesting problems in statistics and machine learning can be written as $\min_x F(x) = f(x) + g(x)$, where x is the model parameter, f is the loss and g is the regularizer. Examples include regularized regression in high-dimensional feature selection and low-rank matrix/tensor factorization. Sometimes the loss function and/or the regularizer is nonsmooth due to the nature of the problem; for example, $f(x)$ could be quantile loss to induce some robustness or to put more focus on different parts of the distribution other than the mean. We propose a general framework to deal with situations when you have nonsmooth loss or regularizer. Specifically, we use low-rank matrix recovery as an example to demonstrate the main idea. The framework involves two main steps: the optimal smoothing of the loss function or regularizer and then a gradient-based algorithm to solve the smoothed loss. The proposed smoothing pipeline is highly flexible, computationally efficient, easy to implement and well-suited for problems with high-dimensional data. A strong theoretical convergence guarantee has also been established. In the numerical studies, we used L_1 loss as an example to illustrate the practicability of the proposed pipeline. Various state-of-the-art algorithms such as Adam, NAG and YellowFin all show promising results for the smoothed loss.

E1831: **Joint penalized spline modeling of multivariate longitudinal data**
Presenter: **Lihui Zhao**, Northwestern University, United States

A joint penalized spline modeling approach is proposed that can be used to model the repeated measurements from multiple biomarkers of various types (eg, continuous, binary) simultaneously. This approach allows for flexible trajectories for each marker, accounts for the potentially time-varying correlation between markers, and is robust to misspecification of knots. Despite its advantages, the application of multivariate penalized spline models, especially when biomarkers may be of different data types, has been limited in part due to its seemingly complexity in implementation. To overcome this, we describe a procedure that transforms the multivariate setting to the univariate one, and then makes use of the generalized linear mixed effect model representation of a penalized spline model to facilitate its implementation with standard statistical software. We performed simulation studies to evaluate the validity and efficiency through joint modeling of correlated biomarkers measured longitudinally compared to the univariate modeling approach. We applied this modeling approach to real data from a longitudinal study.

E1650: **Jackknife empirical likelihood for the mean difference of two zero-inflated skewed populations**
Presenter: **Yichuan Zhao**, Georgia State University, United States
Co-authors: Faysal Satter

In constructing a confidence interval for the mean difference of two independent populations, the problem of having a low coverage probability when there are many zeros in the data, and the non-zero values are highly positively skewed, may be encountered. The violation of the normality assumption makes parametric methods inefficient in such cases. Jackknife empirical likelihood (JEL) and adjusted jackknife empirical likelihood (AJEL) methods are proposed to construct a nonparametric confidence interval for the mean difference of two independent zero-inflated skewed populations. The JEL and AJEL confidence intervals are compared with the confidence intervals by normal approximation and empirical likelihood. Simulation studies are performed to assess the new methods. Two real-life datasets are also used as an illustration of the proposed methodologies.

EO659 Room Safra Lecture Theatre RECENT ADVANCEMENTS IN CAUSAL INFERENCE Chair: Trinetri Ghosh

E1455: **Deconfounding causal inference using latent multiple mediator pathways**
Presenter: **Yubai Yuan**, Penn State University, United States

Causal effect estimation from observational data is one of the essential problems in causal inference. However, most estimation methods rely on the strong assumption that all confounders are observed, which is impractical and untestable in the real world. We develop a mediation analysis framework inferring the latent confounder for debiasing both direct and indirect causal effects. Specifically, we introduce generalized structural equation modeling that incorporates structured latent factors to improve the goodness-of-fit of the model to observed data, and deconfound the mediators and outcome simultaneously. One major advantage of the proposed framework is that it utilizes the causal pathway structure from cause to outcome via multiple mediators to debias the causal effect without requiring external information on latent confounders. In addition, the proposed framework is flexible in terms of integrating powerful nonparametric prediction algorithms while retaining interpretable mediation effects. In theory, we establish the nonparametric identification of both causal and mediation effects based on the proposed deconfounding method. Numerical experiments on both simulation settings and a normative aging study indicate that the proposed approach reduces the estimation bias of both causal and mediation effects.

E1016: **Independence weights for causal inference with continuous treatments**
Presenter: **Jared Huling**, University of Minnesota, United States

Studying causal effects of continuous treatments is important for gaining a deeper understanding of many interventions, policies, or medications, yet researchers are often left with observational studies for doing so. In the observational setting, confounding is a barrier to the estimation of causal effects. Weighting approaches seek to control for confounding by reweighting samples so that confounders are comparable across different values of the treatment. Treatments with continuous or otherwise non-categorical values are often present in medical studies involving Electronic Health Record Data. For example, in studying the causal effect of mechanical power of ventilation in those with acute respiratory disease, mechanical power is continuous. Yet, for continuous treatments, weighting methods are highly sensitive to model misspecification. We elucidate the key property that makes weights effective in estimating causal quantities involving continuous treatments. We show that to eliminate confounding, weights should make treatment and confounders independent on the weighted scale. We develop a measure that characterizes the degree to which a set of weights induces such independence and propose a new model-free method for weight estimation by optimizing our measure. The empirical effectiveness of our approach is demonstrated in a suite of challenging numerical experiments, where we find that our weights are quite robust and work well under a broad range of settings.

E1184: **Causal effect estimation in graphical models with unmeasured confounders**
Presenter: **Rohit Bhattacharya**, Williams College, United States

Co-authors: Razieh Nabi, Ilya Shpitser

Recent developments will be discussed in (i) semiparametric estimation of causal effects in graphical models with unmeasured confounders, and (ii) the design of semiparametric tests for verifying the key identifying assumptions in such models. In particular, we discuss doubly robust estimation strategies for a class of causal graphical models defined by a simple graphical criterion on the treatment variable (this class includes the popular conditionally ignorable model and front-door model as special cases.) We then discuss two newly proposed goodness-of-fit tests, which under mild assumptions, can be used to verify the key identifying assumptions in this class of models. These tests rely on variationally independent pieces of a natural parameterization of the observed data likelihood, and have the appealing property that they require no additional modeling than what is used in the downstream semiparametric estimators. That is, the same models used to perform the pre-test can be re-used for downstream causal effect estimation. We end with a short discussion on theoretical and empirical comparisons of this approach to instrumental variable approaches to handling unmeasured confounding.

E0846: **Efficient estimation of average treatment effect on the treated under endogenous treatment assignment**

Presenter: **Trinetri Ghosh**, University of Wisconsin-Madison, United States

Co-authors: Menggang Yu, Jiwei Zhao

When evaluating a complex intervention, instead of the average treatment effect (ATE), researchers are more interested in the average treatment effect on the treated (ATT), a quantity that is more relevant and more interpretable to policy-makers. We consider the ATT estimation motivated by a case study, where the treatment assignment might depend on the potential untreated outcome and hence is endogenous. We focus on the efficient estimation of ATT by characterizing the geometric structure of the model, deriving the semiparametric efficiency bound for ATT estimation, and proposing an estimator that can achieve this bound. We rigorously establish the theoretical results of the proposed estimator. The finite-sample performance of the proposed estimator is studied through comprehensive simulation studies and an application to our motivating study.

EO604 Room Virtual R01 RECENT ADVANCES IN HIGH DIMENSIONAL TIME SERIES ANALYSIS Chair: Danna Zhang

E0422: **Identification and estimation of change points in factor models for high-dimensional time series data**

Presenter: **Xialu Liu**, San Diego State University, United States

The focus is on estimating and identifying a factor model for high-dimensional time series that contains structural breaks in the factor loading space at unknown time points. We first study the case when there is one change point in factor loadings, and propose a consistent estimator for the structural break location. We show that the proposed estimators for change-point location and loading spaces are consistent when the number of factors is correctly estimated or overestimated. The algorithm for multiple change-point detection is also developed. A distinguishing feature of the proposed method is that it is specifically designed for the changes in the factor loading space and the stationarity assumption is not imposed on either the factor or noise process, while most existing methods for change-point detection of high-dimensional time series with/without a factor structure require the data to be stationary or close to a stationary process between two change points, which is rather restrictive. Numerical experiments, including a Monte Carlo simulation and a real data application, are presented to illustrate the proposed estimators perform well.

E0654: l_2 **inference for change points in high-dimensional time series via a two-way MOSUM**

Presenter: **Jiaqi Li**, Washington University in Saint Louis, United States

Co-authors: Likai Chen, Weining Wang, Wei Biao Wu

A new inference method is proposed for multiple change-point detection in high-dimensional time series, targeting dense or spatially clustered signals. Specifically, we aggregate MOSUM (moving sum) statistics cross-sectionally by an l2-norm and maximize them over time. To account for breaks only occurring in a few clusters, we also introduce a novel Two-Way MOSUM statistic, aggregated within each cluster and maximized over clusters and time. Such an aggregation scheme substantially improves the performance of change-point inference. We contribute to both theory and methodology. Theoretically, we develop an asymptotic theory concerning the limit distribution of an l_2-aggregated statistic to test the existence of breaks. The core of our theory is to extend a high-dimensional Gaussian approximation theorem fitting to non-stationary, spatial-temporally dependent data-generating processes. We provide consistency results of estimated break numbers, time stamps and sizes of breaks. Furthermore, our theory facilitates novel change-point detection algorithms involving newly proposed Two-Way MOSUM statistics. We show that our test enjoys power enhancement in the presence of spatially clustered breaks. A simulation study presents favorable performance of our testing method for non-sparse signals. Two applications concerning equity returns and COVID-19 cases in the United States demonstrate the applicability of our proposed algorithms.

E0669: **CP factor model for dynamic tensors**

Presenter: **Yuefeng Han**, University of Notre Dame, United States

Observations in various applications are frequently represented as a time series of multidimensional arrays, called tensor time series, preserving the inherent multidimensional structure. We present a factor model approach, in a form similar to tensor CP decomposition, to the analysis of high-dimensional dynamic tensor time series. As the loading vectors are uniquely defined but not necessarily orthogonal, it is significantly different from the existing tensor factor models based on Tucker-type tensor decomposition. The model structure allows for a set of uncorrelated one-dimensional latent dynamic factor processes, making it much more convenient to study the underlying dynamics of the time series. A new high-order projection estimator is proposed for such a factor model, utilizing the special structure and the idea of the higher-order orthogonal iteration procedures commonly used in the Tucker-type tensor factor model and general tensor CP decomposition procedures. Theoretic al investigation provides statistical error bounds for the proposed methods, which shows the significant advantage of utilizing the special model structure.

E0568: **Spectral inference for high dimensional time series**

Presenter: **Danna Zhang**, University of California, San Diego, United States

High dimensional non-Gaussian time series data are increasingly encountered in a wide range of applications. We consider the problem of spectral inference of high dimensional time series using the framework of functional dependence measure. In particular, we establish a distributional theory on high dimensional spectra estimates by Gaussian approximation, which can be applied to address various testing problems for time series. We also develop two different resampling methods to implement spectral inference in practice and show the theoretical validity in the high dimensional setting.

EO200 Room Virtual R02 RECENT ADVANCES IN TAILORED DECISION MAKING Chair: Muxuan Liang

E0226: **Penalized doubly robust regression-based estimation of adaptive treatment strategies**

Presenter: **Erica Moodie**, McGill University, Canada

Co-authors: Zeyu Bian, Sahir Bhatnagar, Susan Shortreed, Sylvie Lambert

Adaptive treatment strategies (ATSs) are often estimated from data sources with many covariates measured, only a subset of which are useful for tailoring treatment or control of confounding. Including all available covariates in the analytic model could yield a needlessly complicated treatment decision, with poor statistical efficiency. Hence, we aimed to incorporate variable selection techniques into ATSs. Variable selection with the objective of optimizing treatment decisions has been the subject of very little literature. We will present a regression-based estimation method that can naturally incorporate variable selection through a penalization approach that incorporates sparsity while ensuring strong heredity, and show how we can additionally incorporate confounder selection into the approach. We illustrate the methods by analyzing a pilot sequential multiple

assignment randomized trial of a web-based, stress management intervention using a stepped-care method for cardiovascular diseases patients to determine useful tailoring variables while adjusting for chance imbalances in important covariates due to the smaller sample size in the pilot.

E1543: **Deeply debiased off policy interval estimation**
Presenter: **Chengchun Shi**, LSE, United Kingdom
Co-authors: Runzhe Wan, Victor Chernozhukov, Rui Song
Off-policy evaluation learns a target policy value with a historical dataset generated by a different behavior policy. In addition to a point estimate, many applications would benefit significantly from having a confidence interval (CI) that quantifies the uncertainty of the point estimate. We propose a novel deeply-debiasing procedure to construct an efficient, robust, and flexible CI on a target policy's value. Our method is justified by theoretical results and numerical experiments. A Python implementation of the proposed procedure is available on GitHub.

E1587: **Inference with non-differentiable surrogate loss in a general high-dimensional classification framework**
Presenter: **Muxuan Liang**, University of Florida, United States
Co-authors: Yingqi Zhao, Yang Ning, Maureen Smith
Penalized empirical risk minimization with a surrogate loss function is often used to derive a high-dimensional linear decision rule in classification problems. Although much literature focuses on the generalization error, there is a lack of valid inference procedures to identify the driving factors of the estimated decision rule, especially when the surrogate loss is non-differentiable. We propose a kernel-smoothed de-correlated score to construct hypothesis testings and interval estimations for the linear decision rule estimated using a piece-wise linear surrogate loss, which has a discontinuous gradient and non-regular Hessian. Specifically, we adopt kernel approximations to smooth the discontinuous gradient near discontinuity points and approximate the non-regular Hessian of the surrogate loss. In applications where additional nuisance parameters are involved, we propose a novel cross-fitted version to accommodate flexible nuisance estimates and kernel approximations. We establish the limiting distribution of the kernel-smoothed de-correlated score and its cross-fitted version in a high-dimensional setup. Simulation and real data analysis are conducted to demonstrate the validity and superiority of the proposed method.

E1827: **Using observational data to estimate the optimal dynamic decision rules to tailor the treatment strategy**
Presenter: **Lu Wang**, University of Michigan, United States
A dynamic treatment regime (DTR) is a sequence of decision rules that provide guidance on how to treat individuals based on their static and time-varying status. Existing observational data are often used to generate hypotheses about effective DTRs. A common challenge with observational data, however, is the need for analysts to consider "restrictions" on the treatment sequences. Such restrictions may be necessary for settings where (i) one or more treatment sequences that were offered to individuals when the data were collected are no longer considered viable in practice; (ii) specific treatment sequences are no longer available; or (iii) the scientific focus of the analysis concerns a specific type of treatment sequences (e.g., "stepped-up" treatments). To address this challenge, we propose a Restricted Tree-based Reinforcement Learning (RT-RL) method that searches for an interpretable DTR with the maximum expected outcome, given a (set of) user-specified restriction(s), which specifies treatment options (at each stage) that ought not to be considered as part of the estimated tree-based DTR. In simulations, we evaluate the performance of RT-RL versus the standard approach of ignoring the partial data for individuals not following the (set of) restriction(s). The method is illustrated using an observational dataset to estimate a two-stage stepped-up DTR for guiding the level of care placement for adolescents with substance use disorder.

EO294 Room Virtual R04 RECENT ADVANCES IN STATISTICAL METHODS FOR BIOMEDICAL DATA INTEGRATION Chair: Rui Duan

E0953: **Heterogeneous causal effects in underrepresented populations: federated and transfer learning approaches**
Presenter: **Larry Han**, Harvard University, United States
Co-authors: Rui Duan
In causal inference, it is challenging to estimate treatment effects for underrepresented populations accurately. Data sharing can improve the power to estimate treatment effects, but privacy concerns often constrain the sharing of patient-level data between sites. We propose a novel causal inference framework to leverage multiple sites and multiple subpopulations to make inferences on the conditional average treatment effect (CATE) for an underrepresented target population of interest. The method leverages transfer learning and federated learning to data-adaptively incorporate summary-level information from source populations and sites to learn about the target population CATE. In extensive simulation studies, we show that the proposed method substantially improves the estimation accuracy of the CATE for underrepresented target populations, lowering RMSE by up to 80%. We illustrate our method through a real-world study of COVID-19 vaccine efficacy on infection, hospitalization, and mortality in underrepresented populations using data from multiple VA study sites. Our method reduces the RMSE of the CATE for underrepresented populations by $35-65\%$.

E0955: **DPQL: A lossless distributed algorithm for generalized linear mixed model with application to hospital profiling**
Presenter: **Chongliang Luo**, Washington University in St Louis, United States
Hospital profiling, the process that determines to what extent patient outcomes depend on the hospital, provides a quantitative comparison of healthcare providers based on their quality of care. To implement hospital profiling, the generalized linear mixed model (GLMM) is used to fit outcome models using clinical or administrative claims data. For better generalizability, data across multiple hospitals, databases, or networks are desired. However, due to privacy regulations and the computational complexity of GLMM, a distributed algorithm for hospital profiling is needed. We develop a novel distributed Penalized Quasi Likelihood (dPQL) algorithm to fit GLMM when only aggregated data, rather than individual patient data, can be shared across hospitals. The proposed algorithm is lossless, i.e., it obtains identical results as if individual patient data were pooled from all hospitals. We apply the dPQL algorithm by ranking 929 hospitals for COVID-19 mortality or referral to a hospice that has been previously studied.

E1146: **Clinical knowledge extraction via sparse embedding regression with EHR data**
Presenter: **Chuan Hong**, Duke University, United States
Traditional data mining of EHR data often requires the use of patient-level data, which hinders the ability to share data across institutions. KESER is a knowledge extraction pipeline via sparse embedding regression, which efficiently summarizes patient-level longitudinal EHR data into hospital-specific embedding data and enables the extraction of clinical knowledge based only on summary-level data. KESER bypasses the need for patient-level data in individual analyses providing a significant advance in enabling multi-center studies using EHR data.

E1454: **PALM: Patient-centered treatment ranking via large-scale multivariate network meta-analysis**
Presenter: **Rui Duan**, Harvard University, United States
The growing number of available treatment options has led to an urgent need for reliable answers when choosing the best course of treatment for a patient. As it is often infeasible to compare a large number of treatments in a single randomized controlled trial, multivariate network meta-analyses (NMAs) are used to synthesize evidence from trials of a subset of the treatments, where both efficacy and safety-related outcomes are considered simultaneously. However, these large-scale multiple-outcome NMAs have created challenges to existing methods due to the increasing complexity of the unknown correlations between outcomes and treatment comparisons. We propose a new framework for PAtient-centered treatment ranking via Large-scale Multivariate network meta-analysis, termed as PALM, which includes a parsimonious modeling approach, a fast algorithm for parameter estimation and inference, a novel visualization tool for presenting multivariate outcomes termed as the origami plot, as well as personalized treatment ranking procedures taking into account the individual's considerations on multiple outcomes. In application to an NMA that compares

14 treatment options for labor induction, we provide a comprehensive illustration of the proposed framework and demonstrate its computational efficiency and practicality, and obtained new insights and evidence to support patient-centered clinical decision-making.

EO559 Room Virtual R05 RECENT ADVANCES IN STATISTICAL LEARNING Chair: Eric Chi

E0502: **Hierarchical tensor decompositions and applications**
Presenter: **Jamie Haddock**, Harvey Mudd College, United States

Nonnegative matrix factorization (NMF) has found many applications, including topic modeling and document analysis. Hierarchical NMF (HNMF) variants are able to learn topics at various levels of granularity and illustrate their hierarchical relationship. Recently, nonnegative tensor factorization (NTF) methods have been applied similarly in order to handle data sets with complex, multi-modal structures. Hierarchical NTF (HNTF) methods have been proposed; however, these methods do not naturally generalize their matrix-based counterparts. We propose a new HNTF model which directly generalizes an HNMF model special case, and provide a supervised extension. Our experimental results show that this model more naturally illuminates the topic hierarchy than previous HNMF and HNTF methods.

E0710: **Sketched Gaussian model linear discriminant analysis via randomized Kaczmarz**
Presenter: **Jocelyn Chi**, UCLA, United States
Co-authors: Deanna Needell

Sketched linear discriminant analysis is presented, which is an iterative randomized approach to binary-class Gaussian model linear discriminant analysis (LDA) for very large data. We harness a least squares formulation and mobilize the stochastic gradient descent framework. Therefore, we obtain a randomized classifier with performance that is very comparable to that of full data LDA while requiring access to only one row of the training data at a time. We present convergence guarantees for the sketched predictions on new data within a fixed number of iterations. These guarantees account for both the Gaussian modeling assumptions on the training data and the algorithmic randomness in the sketching procedure. Finally, we demonstrate its performance with varying step-sizes and numbers of iterations. Our numerical experiments demonstrate that sketched LDA can offer a very viable alternative to full-data LDA when the data may be too large for full-data analysis.

E0709: **Tensor t-distribution and tensor response regression**
Presenter: **Qing Mai**, Florida State University, United States

In recent years, promising statistical modeling approaches to tensor data analysis have been rapidly developed. Traditional multivariate analysis tools, such as multivariate regression and discriminant analysis, are now generalized from modeling random vectors and matrices to higher-order random tensors (a.k.a. array-valued random objects). Equipped with tensor algebra and high-dimensional computation techniques, concise and interpretable statistical models and estimation procedures prevail in a wide range of applications. One of the biggest challenges to statistical tensor models is the non-Gaussian nature of many real-world data. Unfortunately, existing approaches are either restricted to normality or implicitly using least squares type objective functions that are computationally efficient but sensitive to data contamination. Motivated by this, we propose a simple tensor t-distribution that is, unlike existing matrix t-distributions, compatible with tensor operators and reshaping of the data. We then study the tensor response regression with tensor t-error, and develop penalized estimation and hypothesis testing under this t-modeling approach. A novel one-step estimation algorithm is developed for the penalized non-convex optimization, and is proven to converge to the global optimum. We study the asymptotic relative efficiency of various estimators under this model and establish the oracle properties in variable selection and near-optimal asymptotic efficiency.

E0668: **A user-friendly computational framework for robust structured regression with the L_2 criterion**
Presenter: **Eric Chi**, Rice University, United States
Co-authors: Xiaoqian Liu, Jocelyn Chi, Kenneth Lange

A user-friendly computational framework is introduced for implementing robust versions of a wide variety of structured regression methods with the L_2 criterion. In addition to introducing an algorithm for performing L2E regression, our framework enables robust regression with the L_2 criterion for additional structural constraints, works without requiring complex tuning procedures on the precision parameter, can be used to identify heterogeneous subpopulations, and can incorporate readily available non-robust structured regression solvers. We provide convergence guarantees for the framework and demonstrate its flexibility with some examples.

EO565 Room Virtual R06 RECENT DEVELOPMENTS FOR MULTIVARIATE ANALYSIS IN HIGH DIMENSIONS Chair: Aaron Molstad

E0424: **A completely tuning-free and robust approach to sparse precision matrix estimation**
Presenter: **Guo Yu**, University of California Santa Barbara, United States

Despite the vast literature on sparse Gaussian graphical models, current methods either are asymptotically tuning-free (which still require fine-tuning in practice) or hinge on computationally expensive methods (e.g., cross-validation) to determine the proper level of regularization. We propose a completely tuning-free approach for estimating sparse Gaussian graphical models. Our method uses model-agnostic regularization parameters to estimate each column of the target precision matrix and enjoys several desirable properties. Computationally, our estimator can be computed efficiently by linear programming. Theoretically, the proposed estimator achieves minimax optimal convergence rates under various norms. We further propose a second-stage enhancement with non-convex penalties, which possesses the strong oracle property. Through comprehensive numerical studies, our methods demonstrate favorable statistical performance. Remarkably, our methods exhibit strong robustness to the violation of the Gaussian assumption and significantly outperform competing methods in heavy-tailed settings.

E0578: **A convex-nonconvex strategy for grouped variable selection**
Presenter: **Xiaoqian Liu**, University of Texas MD, United States
Co-authors: Aaron Molstad, Eric Chi

The grouped variable selection problem is considered. A widely used strategy is to augment the negative log-likelihood function with a sparsity-promoting penalty. Existing methods include the group Lasso, group SCAD, and group MCP. The group Lasso solves a convex optimization problem but is plagued by underestimation bias. The group SCAD and group MCP avoid this estimation bias but require solving a nonconvex optimization problem that may be plagued by suboptimal local optima. We propose an alternative method based on the generalized minimax concave (GMC) penalty, which is a folded concave penalty that maintains the convexity of the objective function. We develop a new method for grouped variable selection in linear regression, the group GMC, that generalizes the strategy of the original GMC estimator. We present an efficient algorithm for computing the group GMC estimator and also prove properties of the solution path to guide its numerical computation and tuning parameter selection. We establish error bounds for both the group GMC and original GMC estimators. A rich set of simulation studies and a real data application indicate that the proposed group GMC approach outperforms existing methods in several different aspects under a wide array of scenarios.

E0599: **Inference on some (nearly)-singular covariance matrices**
Presenter: **Karl Oskar Ekvall**, University of Florida, United States

Recent theory shows reliable inference on small or vanishing variances is possible using connections to inference with singular Fisher information. We briefly review this theory and outline ongoing work on extending the theory to more general settings, including inference on (nearly-)singular covariance matrices. Some computational and practical issues will also be discussed.

E0699: **Statistical inference for joint factor regression with applications to integrating multi-view microbiome data**
Presenter: **Jing Ma**, Fred Hutchinson Cancer Center, United States

Mechanistic understanding of the microbiome requires identifying co-regulating microbial markers (e.g. taxa, metabolites, etc.) that are associated with host health outcomes. A joint factor regression model for analysis of multi-view microbiome data is discussed (e.g. metagenomics, metabolomics, etc.). Unlike existing methods that focus only on variable selection, our method identifies a multivariate association between the outcome and the latent factors common to all omics layers. It also enables hypothesis testing of specific variables underlying the multivariate association. We will illustrate the merit of the proposed method using an analysis of metagenomic and metabolomic data from the Study of Latinos.

EO608 Room Virtual R07 STATISTICAL METHODS FOR MODERN BUSINESS APPLICATIONS Chair: Trambak Banerjee

E0210: **Corporate probability of default: A single-index hazard model approach**
Presenter: **Shaobo Li**, University of Kansas, United States
Co-authors: Shaonan Tian, Yan Yu, Xiaorui Zhu, Heng Lian

Corporate probability of default (PD) prediction is vitally important for risk management and asset pricing. In search of accurate PD prediction, we propose a default-prediction single-index hazard model (DSI). The proposed semiparametric model is flexible and easy to interpret. It encompasses the linear hazard model and enjoys an appealing memoryless feature. Large sample properties are proved for the penalized spline approximation. By applying it to a comprehensive U.S. corporate bankruptcy database we constructed, we discover an interesting V-shaped relationship between the probability of default and the company's financial characteristics. The common discrete linear hazard specification is clearly violated, also confirmed by a simultaneous confidence band. Most importantly, the single-index hazard model passes the Hosmer-Lemeshow goodness-of-fit calibration test while neither does a state-of-the-art linear hazard model in finance nor a parametric class of Box-Cox transformation survival models. In economic value analysis, we find this may translate to as much as three times the profit compared to the linear hazard model. Furthermore, we reexamine the distress risk anomaly via the popular three- and five-factor asset pricing models. Based on the PDs from the proposed model, we find that the distress risk anomaly has weakened or even disappeared during the extended period, including the 2008 financial crisis.

E0979: **On the use of minimum penalties in statistical learning**
Presenter: **Ben Sherwood**, University of Kansas, United States

Modern multivariate machine learning and statistical methodologies estimate parameters of interest while leveraging prior knowledge of the association between outcome variables. We propose the MinPen framework to simultaneously estimate regression coefficients associated with the multivariate regression model and the relationships between outcome variables using common assumptions. The MinPen framework utilizes a novel penalty based on the minimum function to detect and exploit relationships between responses simultaneously. An iterative algorithm is proposed as a solution to the non-convex optimization. Theoretical results such as high dimensional convergence rates, model selection consistency, and a framework for post-selection inference are provided. We extend the proposed MinPen framework to other exponential family loss functions, with a specific focus on multiple binomial responses.

E1131: **Nonparametric empirical bayes prediction in mixed models**
Presenter: **Trambak Banerjee**, University of Kansas, United States
Co-authors: Padma Sharma

Mixed models are classical tools in statistics for modeling repeated data on subjects, such as data on patients, customers or firms collected over time. These models extend conventional linear models to include latent parameters, called random effects, that capture between-subject variation and accommodate dependence within the repeated measurements of a subject. Traditionally, predictions in mixed models are conducted by assuming that the random effects have a zero mean Normal distribution, which leads to the Best Linear Unbiased Predictor (BLUP) of the random effects in these models. However, such a distributional assumption on the random effects is restrictive and may lead to inefficient predictions, especially when the true random effect distribution is far from Normal. Here, we discuss a novel framework, EBPred, for empirical Bayes prediction in mixed models. The predictions from EBPred rely on the Best Predictor (BP) of the random effects, which are constructed without any parametric assumption on the distribution of the random effects. We develop theory to show that the corresponding predictions from EBPred are asymptotically optimal in terms of mean squared error for prediction. Extensive simulation studies demonstrate that EBPred outperforms existing predictive rules in mixed models and the efficiency gain is substantial in many settings.

E1183: **Estimating heterogeneous treatment effects of marketing promotions in B2B markets**
Presenter: **Wreetabrata Kar**, Purdue University, United States

Sales promotions are a common marketing intervention used to stimulate demand. While the use and effectiveness of sales promotions has been well documented in the business-to-consumer literature, there is relatively little said in business-to-business markets. In addition, due to the nature of business-to-business transactions, marketing managers in this space often combine marketing efforts to drive business, such as the use of sales agent visits that coincide with the use of promotions. We explore the dynamic and heterogenous effects of sales promotions in business-to-business markets while quantifying potential synergies in the use of sales visits as an additional marketing intervention. To estimate these effects, we augment the growing literature of counterfactual estimators with advances in machine learning to quantify the causal effects of two different promotional types: a single, time-invariant percentage off discount and a contest-type promotion where clients receive a payoff based on the amount ordered. Initial evidence suggests that these two types of promotions have different levels of effectiveness moderated by client-level characteristics, such as the type of client and their ties to the focal firm. In addition, we find synergistic effects between sales promotions and visits.

EO729 Room Virtual R08 RECENT ADVANCES IN CAUSAL INFERENCE Chair: Fan Xia

E0773: **Mediation analysis with unmeasured treatment-induced confounding**
Presenter: **Fan Xia**, University of California San Francisco, United States

In causal mediation analysis, covariates affected by the treatment or exposure can be a source of confounding between the mediator and the outcome. Like any type of confounders, the treatment-induced confounders can be mismeasured or unmeasured, which could lead to invalid causal inference. Treatment-induced confounders, even when correctly measured, are especially challenging because it mediates part of the exposure effect while confounds the exposure effect through the mediator. As a result, the identification and estimation of natural direct and indirect effects of the exposure to treatment-induced confounders deviate from those of the well-studied average treatment effect. We use variables associated with the treatment-induced confounders as their proxies to account for confounding for the identification of the natural direct and indirect effects. We develop semiparametric theory for estimation and propose estimators that are robust to different types of model misspecifications. We use simulation studies to evaluate the performance of the proposed method.

E0951: **Doubly robust proximal synthetic controls**
Presenter: **Hongxiang Qiu**, University of Pennsylvania, United States
Co-authors: Edgar Dobriban, Xu Shi, Wang Miao, Eric Tchetgen Tchetgen

To infer the treatment effect for a single treated unit using panel data, one common approach is to search for a linear combination of control units' outcomes that mimics the treated unit's pre-treatment outcome trajectory. This linear combination is subsequently used to impute the counterfactual outcomes of the treated unit had it not been treated in the post-treatment period, and estimation of the treatment effect follows. Approaches following this idea have been called synthetic control methods. Existing synthetic control methods rely on correctly modeling certain

aspects of the counterfactual outcome generating mechanism and may require near-perfect matching of the pre-treatment trajectory. Inspired by proximal causal inference, we obtain two novel nonparametric identifying formulae for the average treatment effect for the treated unit: one is based on weighting, and the other combines models for the counterfactual outcome and the weighting function. We develop two GMM estimators based on these two formulae. One new estimator is doubly robust: it is consistent and asymptotically normal if at least one of the outcome model and the weighting model is correctly specified in a parametric model. We demonstrate the performance of the methods via simulations and apply them to evaluate the effect of a tax cut in Kansas on GDP.

E1003: **Accounting for verification bias in prevalence estimation using verbal autopsies**
Presenter: **Zehang Li**, University of California, Santa Cruz, United States

Monitoring data describing the cause of death is an essential component for understanding the burden of disease and evaluating public health interventions, especially during public health emergencies when new diseases emerge. Verbal autopsy (VA) is a well-established method to gather information about deaths outside of hospitals in many low- and middle-income countries. VAs collect symptoms and covariates of a deceased person through a questionnaire conducted to caregivers or people who are familiar with the death. We propose a novel Bayesian hierarchical model framework for estimating the fraction of deaths due to the emerging disease using VA data with reference causes only available for a potentially biased sample of deaths. We use a latent class model to capture the distribution of symptoms given causes that accounts for symptom dependence in a parsimonious way. We discuss several potential sources of bias due to the informative data selection process of cause-of-death verification and adapt our framework to account for the non-ignorable verification mechanism. We demonstrate the performance of our model using both simulation and a mortality surveillance dataset that includes suspected COVID-19-related deaths in Brazil in 2021.

E1008: **An instrumental variable method for point processes: Generalized Wald estimation based on deconvolution**
Presenter: **Shizhe Chen**, University of California, Davis, United States
Co-authors: Zhichao Jiang, Peng Ding

Point processes are probabilistic tools for modeling event data such as neural spike trains, natural disasters, and crimes. While there exists a fast-growing literature studying the relationships between point processes, it remains unexplored how such relationships connect to causal effects. In the presence of unmeasured confounders, parameters from point process models do not necessarily have causal interpretations. We propose an instrumental variable method for causal inference with point process treatment and outcome. We define causal quantities based on potential outcomes and establish nonparametric identification results with a binary instrumental variable. We extend the traditional Wald estimation for point process treatment and outcome, and show that it should be performed after a Fourier transform and thus takes the form of deconvolution. We term this as the generalized Wald estimation and propose an estimation strategy based on well-established deconvolution methods. The proposed estimation strategy is applicable under many commonly-used models without requiring distributional assumptions on the unmeasured confounders. We apply the proposed methodology to analyze the data from an experiment on mouse neuron activities.

EO164 Room K2.31 (Nash Lec. Theatre) EFFECT ESTIMATION UN VARIOUS CONTEXTS Chair: Mireille Schnitzer

E0602: **Evaluating treatment efficacy with stochastic-interventional causal effects in clinical trials with two-phase designs**
Presenter: **Nima Hejazi**, Harvard T.H. Chan School of Public Health, United States
Co-authors: David Benkeser, Peter Gilbert

In clinical trials randomizing participants to active vs control conditions and following units until the occurrence of a primary clinical endpoint, evaluating the efficacy of a quantitative treatment (e.g., drug dosage) is often difficult. Stochastic-interventional effects, which measure the causal effect of perturbing the treatment's observed value, provide an interpretable solution; yet, their use in vaccine trials requires care, for such trials measure immunologic biomarkers – useful for understanding the mechanisms by which vaccines confer protection or as surrogate endpoints – via outcome-dependent two-phase sampling (e.g., case-cohort) designs. These biased sampling designs have earned their popularity: they circumvent the economic burden of measuring biomarkers on all study units without limiting opportunities to detect mechanistically informative biomarkers. We discuss a semiparametric biased sampling correction allowing for asymptotically efficient inference on a causal vaccine efficacy measure, defined by contrasting assignments of study units to active vs control while also shifting observed biomarker expression in the active condition, yielding a causal dose-response analysis informative of next-generation vaccine efficacy and of transporting efficacy from a source pathogen strain (e.g., SARS-CoV-2 at outbreak) to variants of concern (e.g., Omicron BA.5). We present the results of applying this approach in the Moderna COVE COVID-19 vaccine efficacy trial.

E0648: **Causal bounds for outcome-dependent sampling in observational studies**
Presenter: **Erin Gabriel**, University of Copenhagen, Denmark
Co-authors: Michael Sachs, Arvid Sjolander

Outcome-dependent sampling designs are common in many different scientific fields, including epidemiology, ecology, and economics. As with all observational studies, such designs often suffer from unmeasured confounding, which generally precludes the nonparametric identification of causal effects. Nonparametric bounds can provide a way to narrow the range of possible values for a nonidentifiable causal effect without making additional untestable assumptions. The nonparametric bounds literature has almost exclusively focused on settings with random sampling, and the bounds have often been derived with a particular linear programming method. We derive novel bounds for the causal risk difference, often referred to as the average treatment effect, in six settings with outcome-dependent sampling and unmeasured confounding for a binary outcome and exposure. Our derivations of the bounds illustrate two approaches that may be applicable in other settings where the bounding problem cannot be directly stated as a system of linear constraints.

E1628: **Proximal causal inference under confounded outcome-dependent sampling**
Presenter: **Kendrick Li**, University of Michigan, United States
Co-authors: Xu Shi, Eric Tchetgen Tchetgen, Wang Miao

Outcome-dependent sampling is widely used in epidemiology and econometrics to reduce time and effort when studying causal relationships between the exposure and outcome variables. In these types of studies, unmeasured confounding and selection bias are often of concern and may invalidate a causal analysis if not appropriately accounted for. In particular, a latent factor that has causal effects on the treatment, outcome, and sample selection process may cause both unmeasured confounding and selection bias, rendering standard causal parameters unidentifiable without additional assumptions. We introduce the identification and inference of treatment effect under a homogeneous odds ratio model, leveraging a pair of proxies to the source of unmeasured confounding: a negative control exposure (NCE) which is a priori known not to affect the outcome and selection, and a negative control outcome (NCO) which is a priori known not to be affected by the treatment. We introduce three estimators of the odds ratio effect, one of which is doubly robust with respect to the specification of two nuisance functions which restrict the treatment assignment mechanism and outcome distribution, respectively, such that the estimator is consistent and asymptotically normal if either model is correctly specified, without knowing which one is.

E1673: **On the estimation of peer effects for sampled networks: The special case of respondent-driven sampling**
Presenter: **Mamadou Yauck**, UQAM, Canada

Respondent-driven sampling (RDS) is a form of link-tracing sampling, a technique for sampling hard-to-reach populations that aims to leverage individuals' social relationships to reach potential participants. An RDS sample represents a partially observed network of unknown dependence structure. Current analytical approaches for RDS data focus mainly on estimating means and proportions but give little technical consideration to

multivariate modeling. Progress in this area is limited by a missing data problem: the observed RDS network reveals partial information about social connections between individuals in the sample. We show that the parameters of regression models are not, in general, identifiable because different full data distributions may give rise to the same observed data distribution. The lack of identification causes a violation of some model assumptions; in the estimation of peer effects, the conditional expectation of the error term in the linear regression, given the vector of covariates, is not zero. Thus, standard inferential methods such as maximum likelihood estimation will not, in general, be valid. We introduce additional assumptions to characterize the asymptotic biases of the maximum likelihood estimators for the peer effects and propose bias-corrected estimators.

EO064 Room K2.40 ADVANCES FOR FUNCTIONAL AND HIGH-DIMENSIONAL DATA ANALYSIS (VIRTUAL) Chair: Yumou Qiu

E1721: **A distribution-free independence test for high dimension data**
Presenter: **Zhanrui Cai**, Iowa State University, United States

Test of independence is of fundamental importance in modern data analysis, with broad applications in variable selection, graphical models, and causal inference. When the data is high dimensional and the potential dependence signal is sparse, independence testing becomes very challenging without distributional or structural assumptions. We propose a general framework for independence testing by first fitting a classifier that distinguishes the joint and product distributions, and then testing the significance of the fitted classifier. This framework allows us to borrow the strength of the most advanced classification algorithms developed from the modern machine learning community, making it applicable to high-dimensional, complex data. By combining a sample split and a fixed permutation, our test statistic has a universal, fixed Gaussian null distribution that is independent of the underlying data distribution. Extensive simulations demonstrate the advantages of the newly proposed test compared with existing methods. We further apply the new test to a genetic dataset, where the high dimensionality makes existing methods hard to apply.

E1911: **Projected permutation tests for canonical correlation analysis**
Presenter: **Yunhui Qi**, Iowa State University, United States
Co-authors: Yumou Qiu, Peng Liu

Canonical correlation analysis (CCA) and sparse CCA (sCCA) are dimension-reduction tools to explore relationships between two sets of variables. An essential step in applying CCA and sCCA is to determine the number of canonical components. To do this, we sequentially test whether the remaining singular values of the cross-correlation matrix are zero, given the leading components. We call our test a projection-based permutation test (PPT) because we use projection to exclude the information carried by leading components and use permutation to test whether there exists information in the remaining components. In high dimensions, we further propose to use a low-dimensional transformation to amplify signals. Our method overcomes the shadowing problem of parallel analysis by projection, thus having higher power. It is also more robust to the accuracy of sCCA, reducing the requirements of penalty parameters. In addition, we propose a procedure to test if the loadings are zero, which helps select the contributing variables. Simulation studies show our methods have lower error rates and higher power compared to the existing ones. A case study on COVID patients reveals the connection between COVID severity and the functions of proteome and metabolome.

E1914: **Model-based bicluster algorithm for microbiome data**
Presenter: **Peng Liu**, Iowa State University, United States
Co-authors: Zhili Qiao

With the advancement of next-sequencing technologies, huge amounts of microbiome data have become available. Bicluster analysis is a tool to quantitively explore the relationships between microbial samples and between features simultaneously, and aims to reveal the interactions between microbial sub-communities. Due to compositionality and sparsity, it is challenging to conduct bicluster analysis with microbiome data. We propose a Dirichlet-Multinomial (DM) model-based checkerboard biclustering method to cluster microbiome features and samples simultaneously. This method assumes a mixture of DM distributions across microbiome samples, and uses a combination of the Expectation-Maximization algorithm and coordinate descent algorithm to solve for parameter estimates and achieve biclustering results. Simulation studies under a variety of settings show that our proposed method outperforms alternative methods. Application to a real dataset demonstrates the effectiveness of our proposed method and provides interesting biological findings.

E1967: **Bootstrap inference in functional linear regression models with scalar response**
Presenter: **Xiongtao Dai**, University of California, Berkeley, United States
Co-authors: Hyemin Yeon, Daniel Nordman

A fundamental problem of bootstrap inference is investigated for functional linear regression models (FLRMs). The work further develops the central limit theorem (CLT) and residual bootstrap in FLRMs by generalizing, refining, and correcting existing results. The theoretical results on bootstrap and the CLT are novel and apply either conditionally or unconditionally on data regressors. We also develop a new method for obtaining simultaneous inference based on residual bootstrap and provide theoretical supports. Numerical studies verify our theory, showing that the bootstrap performs better than normal approximations, and also suggest a rule of thumb for setting the truncation levels. The bootstrap method is illustrated with an application to wheat spectrum data.

EO344 Room K2.41 RECENT ADVANCES IN STATISTICAL MODELING OF COMPLEX DATA STRUCTURES Chair: Tianxi Li

E0616: **Bootstrapping network data: Conditional and marginal approaches**
Presenter: **Keith Levin**, University of Wisconsin, United States

In network analysis, one frequently must perform inference based upon only one sampled network. This poses a challenge for bootstrap-based approaches, which typically require an iid sample. A class of network models called latent space models overcome these difficulties by generating a network based on unobserved geometric structure, but this raises the question of whether inference in such models should be conducted by conditioning on this latent structure or by marginalizing over it. We develop bootstrap schemes for both cases, i.e., conditional and marginal bootstrap methods for network data. We establish bootstrap validity for both schemes for a broad class of network statistics, including modularity, which has not previously been addressed within the network bootstrap literature. Our experiments include simulated data as well as a thorough exploration of a data set arising from Microsoft Bing search data.

E1087: **L-2 regularized maximum likelihood for beta-model estimation in large and sparse networks**
Presenter: **Yuan Zhang**, Ohio State University, United States

The beta-model is a powerful tool for modeling networks driven by degree heterogeneity. Its simple yet expressive nature particularly well-suits large and sparse networks, where most models are infeasible due to computational challenge and observation scarcity. However, existing algorithms for beta-model do not scale up; and theoretical understandings remain limited to dense networks. Several major improvements to the method and theory of -model are brought to address the urgent needs of practical applications. The contributions include: 1. Method: we propose a new L-2 penalized MLE scheme; we design a novel algorithm that can comfortably handle sparse networks of millions of nodes, much faster and more memory-parsimonious than any existing algorithm; 2. Theory: we present new error bounds on beta-models under much weaker assumptions; we also establish new lower-bounds and new asymptotic normality results; distinct from existing literature, our results cover both small and large regularization scenarios and reveal their distinct asymptotic dependency structures; 3. Application: we apply our method to large COVID-19 network data sets and discover meaningful results.

E1127: **Network autoregressive processes and their applications**
Presenter: **George Michailidis**, University of Florida, United States

Models are developed for Network Autoregressive Processes (NAR), wherein the response of each node linearly depends on its past values, a prespecified linear combination of neighboring nodes and a set of node-specific covariates. The corresponding coefficients are node-specific, while the framework can accommodate heavier than Gaussian errors with both spatial-autorgressive and factor-based covariance structures. We consider the stability (stationarity) of the underlying NAR and develop estimators for both a fixed, as well as a diverging number of network nodes. We also address the issue of selecting the network connectivity and the impact of misspecifying it on inference. The framework is illustrated on both synthetic and real data sets.

E1445: **Computing pseudolikelihood estimators for exponential-family random graph models**
Presenter: **David Hunter**, Pennsylvania State University, United States
Co-authors: Christian Schmid

The reputation of the maximum pseudolikelihood estimator (MPLE) for Exponential Random Graph Models (ERGM) has undergone a drastic change over the past 30 years. While first receiving broad support, mainly due to its computational feasibility and the lack of alternatives, general opinions started to change with the introduction of approximate maximum likelihood estimator (MLE) methods that became practicable due to increasing computing power and the introduction of MCMC methods. Previous comparison studies appear to yield contradicting results regarding the preference of these two point estimators; however, there is consensus that the prevailing method to obtain an MPLE's standard error by the inverse Hessian matrix generally underestimates standard errors. We propose replacing the inverse Hessian matrix with an approximation of the Godambe matrix that results in confidence intervals with appropriate coverage rates and that, in addition, enables examining for model degeneracy. Our results also provide empirical evidence for the asymptotic normality of the MPLE under certain conditions.

EC823 Room S-1.01 COMPUTATIONAL STATISTICS II Chair: Cristian Gatu

E1743: **A flexible bias correction method based on inconsistent estimators**
Presenter: **Yuming Zhang**, University of Geneva, Switzerland
Co-authors: Yanyuan Ma, Samuel Orso, Mucyo Karemera, Maria-Pia Victoria-Feser, Stephane Guerrier

An important challenge in statistical analysis lies in controlling the estimation bias when handling the ever-increasing data size and model complexity. For example, approximate methods are increasingly used to address the analytical and/or computational challenges when implementing standard estimators, but they often lead to inconsistent estimators. So consistent estimators can be difficult to obtain, especially for complex models and/or in settings where the number of parameters diverges with the sample size. We propose a general simulation-based estimation framework that allows us to construct consistent and bias-corrected estimators for parameters of increasing dimensions. The key advantage of the proposed framework is that it only requires to compute a simple inconsistent estimator multiple times. The resulting Just Identified iNdirect Inference estimator (JINI) enjoys nice properties, including consistency, asymptotic normality, and finite sample bias correction, better than alternative methods. We further provide a simple algorithm to construct the JINI in a computationally efficient manner. Therefore, the JINI is especially useful in settings where standard methods may be challenging to apply, for example, in the presence of misclassification and rounding.

E1908: **Generalized linear models for massive data via sketching**
Presenter: **Jason Hou-Liu**, University of Waterloo, Canada
Co-authors: Ryan Browne

Generalized linear models are a popular analytics tool with interpretable results and broad applicability, but they require iterative estimation procedures that impose data transfer and computational costs that can be problematic under some infrastructure constraints. We propose a doubly-sketched stochastic approximation to the iteratively re-weighted least squares algorithm to estimate a variety of generalized linear models using a sequence of sketched surrogate datasets. A uniform sketch reduces data transfer costs, and a subsequent Clarkson-Woodruff sketch reduces local computation costs, yielding substantial wall-clock time savings. Regression coefficients and standard errors are produced, with comparison against single subsample and literature methods. Some theoretical properties of the proposed procedure are shown, with empirical results from simulated and real-world datasets. The efficacy of the proposed method is investigated across a variety of commodity computational infrastructure configurations accessible to practitioners. A highlight of the present work is the estimation of a Poisson-log generalized linear model across 1.67 billion observations on a personal computer in 25 minutes.

E2013: **Computational strategies for regression model selection in the high-dimensional case**
Presenter: **Marios Demosthenous**, National Technical University of Athens, Greece
Co-authors: Cristian Gatu, Erricos Kontoghiorghes, Ana Colubi

Computational strategies for finding the best-subset regression models are proposed. The case of high-dimensional (HD) data where the number of variables exceeds the number of observations is considered. Within this context, a theoretical combinatorial solution is proposed. It is based on a regression tree structure that generates all possible subset models. An efficient branch-and-bound algorithm that finds the best submodels without generating the entire tree is adapted to the HD case. Furthermore, in the HD case, the R package lmSubsets is employed in order to identify the best submodel based on the AIC family selection criteria. Preliminary experimental results are presented and analyzed. The efficient extension of the lmSelect algorithm to HD is discussed.

E0196: **Lockout: Sparse regularization of neural networks**
Presenter: **Gilmer Valdes**, UCSF, United States
Co-authors: Jerome Friedman, Wilmer Arbelo

Many regression and classification procedures fit a parameterized function $f(x;w)$ of predictor variables x to training data based on some loss criterion $L(y,f)$. Often, regularization is applied to improve accuracy by placing a constraint $P(w) \leq t$ on the values of the parameters w. Although efficient methods exist for finding solutions to these constrained optimization problems for all values of t_0 in the special case when f is a linear function, none are available when f is non-linear (e.g. Neural Networks). We present a fast algorithm that provides all such solutions for any differentiable function f and loss L, and any constraint P that is an increasing monotone function of the absolute value of each parameter. Applications involving sparsity-inducing regularization of arbitrary Neural Networks are discussed (both feature and architecture selection). Empirical results indicate that these sparse solutions are usually superior to their dense counterparts in both accuracy and interpretability. This improvement in accuracy can often make Neural Networks competitive with, and sometimes superior to, state-of-the-art methods in the analysis of tabular data.

CO198 Room S-2.23 MACROECONOMIC POLICY Chair: Michael Owyang

C0268: **Good policy or learning evolution: A Markov-Switching approach to understanding the determinants of Fed policy**
Presenter: **Gabriela Best**, California State University, Fullerton, United States

The aim is to analyze the determinants of the Federal Reserve's monetary policy decisions since the 1960s justified by potentially evolving beliefs through a real-time learning process about the structure of the economy and Markov-Switching shifts in policymakers' preferences between dove and hawk regimes. We argue that although central bank learning plays an important role in the determination of Fed policy, there were several shifts in policy makers preferences in the post-war period that explain movements in the monetary instrument. We find a dovish kind of monetary policy

regime present in the 1970s and early 2000s, and before the onset of the Great Recession. The regime-switching dynamics affect the contribution of the shocks to the variables; in particular, it affects the participation of the MEI, government spending, and technology shocks to the variables; however, the effects are asymmetric across variables and regimes. In addition, we use a recent algorithm to deal with the problem of solving rational expectations models under indeterminacy, and we find that impulse responses take the traditional shape and sign under most hawk regimes (except for earlier in the sample during the 1960s), however, under dovish regimes some impulse responses are non-traditional.

C0251: **Who is afraid of eurobonds?**
Presenter: **Anna Rogantini Picco**, Sveriges Riksbank, Sweden
Co-authors: Leonardo Melosi, Francesco Bianchi

The low-interest rate environment and the growing asymmetry in the size of fiscal imbalances pose a serious challenge to the macroeconomic stability of the Euro Area (EA). We show that the current monetary and fiscal framework weakens economic growth even in low-debt countries because of the zero lower bound (ZLB) constraint. We study a new framework that allows EA policymakers to separate the need for short-run macroeconomic stabilization from the issue of long-run fiscal sustainability. The central bank tolerates the increase in inflation needed to stabilize the amount of Eurobonds issued in response to large EA recessions. National governments remain responsible for backing their country-level debt with fiscal adjustments. The policy acts as an automatic stabilizer that benefits both high-debt and low-debt countries, generating a moderate increase in inflation that mitigates the recession and allows the central bank to move away from the ZLB.

C0269: **An estimated model of household inflation expectations: Information frictions and implications**
Presenter: **Shihan Xie**, University of Illinois, Urbana-Champaign, United States

A dynamic model of household inflation expectations is proposed and estimated. The information flow constraint of the household leads to costly information monitoring. Households use a Bayesian learning model to form and update inflation expectations. The model identifies and corrects for sizable reporting and sampling errors prevalent in household surveys. The estimates show that better-educated households track inflation more closely and report their expectations more accurately. Household inflation expectations are less responsive to changes in the inflation target after the Great Recession. Model-implied household inflation expectations improve the fit of the expectation-augmented Phillips curve. Inattention from households makes it more costly for the Fed to lower inflation than would be the case if everyone were perfectly informed.

C0301: **The distributional effects of stabilization policy**
Presenter: **Michael Owyang**, Federal Reserve Bank of St Louis, United States
Co-authors: Alessia Paccagnini, Laura Jackson Young

Recent papers have studied the effects of monetary policy on the distribution of income. Generally, expansionary monetary policy (i.e., lower interest rates) is thought to increase income inequality via capital gains income. We argue that any expansionary stabilization policy -monetary or fiscal- has the same distributional effects. Even policies that target income inequality, such as distributional tax policy, can raise income inequality if their net effects are expansionary.

CO290 Room Virtual R03 RECENT ADVANCES IN FORECASTING Chair: Ekaterina Smetanina

C0617: **Bootstrapping out-of-sample predictability tests with real-time data**
Presenter: **Michael McCracken**, Federal Reserve Bank of St. Louis, United States
Co-authors: Silvia Goncalves, Yongxu Yao

A block bootstrap approach is developed for out-of-sample inference when real-time data is used to produce forecasts. In particular, we establish its first-order validity for West-type tests of predictive ability in the presence of regular data revisions. This allows the user to conduct asymptotically valid inference without having to estimate the asymptotic variances derived in an extension of the original test when data is subject to revision. Monte Carlo experiments indicate that the bootstrap can provide a satisfactory finite sample size and power even in modest sample sizes.

C1013: **Forecasting with panel data: Estimation uncertainty versus parameter heterogeneity**
Presenter: **Allan Timmermann**, UCSD, United States
Co-authors: M Hashem Pesaran, Andreas Pick

Novel forecasting methods are developed for panel data with heterogeneous parameters and examine them together with existing approaches. We conduct a systematic comparison of their predictive accuracy in settings with different cross-sectional (N) and time (T) dimensions and varying degrees of parameter heterogeneity. We investigate conditions under which panel forecasting methods can perform better than forecasts based on individual estimates and demonstrate how gains in predictive accuracy depend on the degree of parameter heterogeneity, whether heterogeneity is correlated with the regressors, the goodness of fit of the model, and, particularly, the time dimension of the data set. We propose optimal combination weights for forecasts based on pooled and individual estimates and develop a novel forecast poolability test that can be used as a pretesting tool. Through a set of Monte Carlo simulations and three empirical applications to house prices, CPI inflation, and stock returns, we show that no single forecasting approach dominates uniformly. However, forecast combination and shrinkage methods provide better overall forecasting performance and offer more attractive risk profiles compared to individual, pooled, and random effects methods.

C1129: **Euro area monetary policy effects: Does the shape of the yield curve matter?**
Presenter: **Barbara Rossi**, Universitat Pompeu Fabra and ICREA, Spain
Co-authors: Giulia Sestieri, Maria Sole Pagliari, Adrian Penalver, Florens Odendahl

The effects of monetary policy in the euro area are investigated. The novelty is that we use the information from movements in the entire yield curve around monetary policy events to shed light on the efficacy of monetary policy. We find that the way quantitative easing announcements, as well as speeches of ECB Presidents, shape expectations inherent in the term structure matter in shaping the response of key macroeconomic variables.

C1212: **Estimating the U.S. output gap using industry level data**
Presenter: **Gianni Amisano**, Federal Reserve Board, United States

U.S. sectoral data are used to try and improve a model-based measure of the output gap. We devise a panel trend-cycle model in which each sector's output, wage inflation, and a proxy for sectoral unemployment are decomposed into cyclical and trend components. Each of these components has a factor structure, with a common, economy-wide factor and sectoral idiosyncrasies. The cyclical component of each sector is linked across variables by Okuns law and a wage Phillips curve. Our results so far show that sector-specific developments generate diverging measures of sectoral slack and that the common cyclical component has different degrees of relevance across sectors. In addition, the resulting measure of the U.S. output gap is more precisely estimated and smaller than that obtained using aggregate data.

Authors Index

