PROGRAMME AND ABSTRACTS

12th International Conference on

Computational and Financial Econometrics (CFE 2018)

http://www.cfenetwork.org/CFE2018

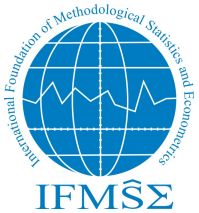
and

11th International Conference of the
ERCIM (European Research Consortium for Informatics and Mathematics) Working Group on

Computational and Methodological Statistics (CMStatistics 2018)

http://www.cmstatistics.org/CMStatistics2018

University of Pisa, Italy
14 – 16 December 2018



ISBN 978-9963-2227-5-9

International Organizing Committee:
Ana Colubi, Erricos Kontoghiorghes, Herman Van Dijk and Caterina Giusti.

CFE 2018 Co-chairs:
Alessandra Amendola, Michael Owyang, Dimitris Politis and Toshiaki Watanabe.

CFE 2018 Programme Committee:

Francesco Audrino, Christopher Baum, Monica Billio, Christian Brownlees, Laura Coroneo, Richard Fairchild, Luca Fanelli, Lola Gadea, Alain Hecq, Benjamin Holcblat, Rustam Ibragimov, Florian Ielpo, Laura Jackson, Robert Kohn, Degui Li, Alessandra Luati, Svetlana Makarova, Claudio Morana, Teruo Nakatsuma, Yasuhiro Omori, Alessia Paccagnini, Sandra Paterlini, Ivan Paya, Christian Proano, Artem Prokhorov, Arvid Raknerud, Joern Sass, Willi Semmler, Etsuro Shioji, Genaro Sucarrat, Robert Taylor, Martin Wagner, Peter Winker and Jean-Michel Zakoian.

CMStatistics 2018 Co-chairs:
Miguel de Carvalho, John Einmahl, Timothy Johnson and Marco Riani.

CMStatistics 2018 Programme Committee:

Ana Maria Aguilera, Alexander Aue, Adelchi Azzalini, Arne Bathke, Jan Beirlant, Veronica Berrocal, Malgorzata Bogdan, Efstathia Bura, Silvia Cagnone, Yu Cheng, Jeng-Min Chiou, Stephane Chretien, Bertrand Clarke, Harry Crane, Claudia Czado, Michael Daniels, Herold Dehling, Fabrizio Durante, Jochen Einbeck, Roland Fried, Pedro Galeano, Yulia Gel, Michele Guindani, Christopher Hans, Hyokyoung Grace Hong, Salvatore Ingrassia, Maria Iorio, Keith Knight, Carlos Lamarche, Thomas Lee, Christophe Ley, Xabier Luna, Marloes Maathuis, Alfio Marazzi, Lola Martinez-Miranda, Geoffrey McLachlan, George Michailidis, Domingo Morales, Jeffrey Morris, Hernando Ombao, Davy Paindaveine, Byeong Park, Taesung Park, Igor Pruenster, Juan Romo, Holger Rootzen, Peter Rousseeuw, Jason Roy, Matteo Ruggiero, Thomas Scheike, Armin Schwartzman, Gilles Stupfler, Ingrid Van Keilegom, Tim Verdonck, Maria-Pia Victoria-Feser, Huixia Judy Wang, Jane-Ling Wang, Nanny Wermuth, Yichao Wu, Tingting Zhang, Chen Zhou and David van Dyk.

Local Organizer:
University of Pisa

Dear Friends and Colleagues,

We welcome you warmly to Pisa, for the 12^{th} International Conference on *Computational and Financial Econometrics* (CFE 2018) and the 11^{th} International Conference of the ERCIM Working Group on *Computational and Methodological Statistics* (CMStatistics 2018). This annual conference has become a leading joint international meeting at the interface of statistics, econometrics, empirical finance and computing.

The conference aims at bringing together researchers and practitioners to discuss recent developments in computational methods for economics, finance, and statistics. The CFE-CMStatistics 2018 programme consists of about 350 sessions, 5 plenary talks and 1440 presentations. There are about 1550 participants. The size and quality of the conference 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 conference venue will provide the appropriate environment to enhance your contacts and to establish new ones. The conference is a collective effort by many individuals and organizations. The Scientific Programme Committee, the Session Organizers, the local hosting university and many volunteers have contributed substantially to the organization of the conference. We acknowledge their work and the support of our hosts and sponsors, and particularly the University of Pisa.

The Elsevier journal, Econometrics and Statistics (EcoSta), has been inaugurated at previous editions of this conference. The EcoSta is the 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, the Annals of Computational and Financial Econometrics.

Looking forward, the CFE-CMStatistics 2019 will be held at the Senate House University of London, UK, from Saturday the 14th to Monday the 16th of December 2019. Tutorials will take place on Friday the 13th of December 2019. You are invited and encouraged to actively participate in these events.

We wish you a productive, stimulating conference and a memorable stay in Pisa.

Ana Colubi, Erricos J. Kontoghiorghes and Herman K. Van Dijk: coordinators of CMStatistics & CFEnetwork.

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 1750 members and the following specialized teams

BM: Bayesian Methodology
CODA: Complex data structures and Object Data Analysis
CPEP: Component-based methods for Predictive and Exploratory Path modeling
DMC: Dependence Models and Copulas
DOE: Design Of Experiments
EF: Econometrics and Finance
GCS: General Computational Statistics WG CMStatistics
GMS: General Methodological Statistics WG CMStatistics
GOF: Goodness-of-Fit and Change-Point Problems
HDS: High-Dimensional Statistics
ISDA: Imprecision in Statistical Data Analysis
LVSEM: Latent Variable and Structural Equation Models
MCS: Matrix Computations and Statistics
MM: Mixture Models
MSW: Multi-Set and multi-Way models
NPS: Non-Parametric Statistics
OHEM: Optimization Heuristics in Estimation and Modelling
RACDS: Robust Analysis of Complex Data Sets
SAE: Small Area Estimation
SAET: Statistical Analysis of Event Times
SAS: Statistical Algorithms and Software
SEA: Statistics of Extremes and Applications
SFD: Statistics for Functional Data
SL: Statistical Learning
SSEF: Statistical Signal Extraction and Filtering
TSMC: Times Series Modelling and Computation

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 by 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 activities of the network 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.

Specialized teams

Currently the CFEnetwork has over 1000 members and the following specialized teams

AE: Applied Econometrics
BE: Bayesian Econometrics
BM: Bootstrap Methods
CE: Computational Econometrics
ET: Econometric Theory
FA: Financial Applications
FE: Financial Econometrics
TSE: Time Series Econometrics

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.

CFE-CMStatistics 2018 - Interactive Programme

2018-12-14

Opening, 08:50 - 09:00

A - Keynote
CFE - CMStatistics
09:00 - 09:50

Coffee Break
09:50 - 10:20

B
CFE - CMStatistics
10:20 - 12:00

C - Keynote
CFE - CMStatistics
12:10 - 13:00

Lunch Break
13:00 - 14:40

D
CFE - CMStatistics
14:40 - 16:20

Coffee Break
16:20 - 16:50

E
CFE - CMStatistics
16:50 - 18:30

Welcome Reception
19:00 - 20:30

2018-12-15

F
CFE - CMStatistics
08:45 - 10:05

Coffee Break
10:05 - 10:35

G
CFE - CMStatistics
10:35 - 12:40

Lunch Break
12:40 - 14:10

H
CFE - CMStatistics
14:10 - 15:50

Coffee Break
15:50 - 16:20

I
CFE - CMStatistics
16:20 - 18:00

J
CFE - CMStatistics
18:10 - 19:25

Conference Dinner
20:30 - 23:30

2018-12-16

K - Keynote
CFE - CMStatistics
08:45 - 09:35

Coffee Break
09:35 - 10:05

L
CFE - CMStatistics
10:05 - 12:10

M - Keynote
CFE - CMStatistics
12:20 - 13:10

Lunch Break
13:10 - 14:40

N
CFE - CMStatistics
14:40 - 16:20

Coffee Break
16:20 - 16:50

O
CFE - CMStatistics
16:50 - 18:05

P - Keynote
CFE - CMStatistics
18:15 - 19:05

Closing, 19:05 - 19:20

TUTORIALS, MEETINGS AND SOCIAL EVENTS

WINTER SCHOOL AND TUTORIALS

The COST Action CRoNoS Winter Course on Time Series takes place from Tuesday the 11th to Thursday the 13th of December 2018 at the Aula Magna, Polo Economia-Polo Didattico (see maps on pages VIII and X). The courses on Thursday are also designated as tutorials of the conference. The first tutorial is given by Prof. Tommaso Proietti (Modelling seasonality in high frequency data) from 9:00 to 13:30. The second tutorial is given by Prof. Dimitris Politis (Model-free prediction for stationary and nonstationary time series) from 15:00 to 19:30.

SPECIAL MEETINGS by invitation to group members

- The *CSDA Editorial Board* meeting will take place on Thursday, 13th December 2018, 18:00 - 19:00, at the room Sala Convegni, Polo Piagge (see maps on pages VIII and IX). The CSDA reception will take place on Thursday, 13th December 2018, 19:00 - 20:30 in front of the Sala Convegni.
- The *Econometrics and Statistics (EcoSta) Editorial Board* meeting will take place on Friday, 14th December 2018, 13:00-13:45, at the room Sala Convegni, Polo Piagge (see maps on pages VIII and IX). The EcoSta reception will take place on Thursday, 13th December 2018, 19:00 - 20:30 in front of the Sala Convegni.
- The *COST Action CRONOS* meeting will take place on Friday 14th December 2018, 18:30-19:00, at the room Sala Convegni of the Polo Piagge (see maps on pages VIII and IX). All the applicants to the ECI Award are invited to participate.

SOCIAL EVENTS

- *The coffee breaks* will take place at the tent located on the backyard of the Polo Piagge (see maps on pages VIII and IX). You must have your conference badge in order to attend the coffee breaks.
- *Welcome Reception, Friday, 14th December 2018, 19:00-20:30.* The Welcome Reception is open to all registrants and accompanying persons who have purchased a reception ticket. It will take place at the the tent located on the backyard of the Polo Piagge (see maps on pages VIII and IX). Conference registrants must bring their conference badge. Information about the welcome reception booking is embedded in the QR code on the conference badge. Preregistration is required due to health and safety reasons, and limited capacity of the venue. Entrance to the reception venue will be strictly allowed only to those who have prebooked.
- *Conference Dinner, Saturday 15th of December, 20:30 to 23:30.* The conference dinner is optional and registration is required. It will take place at Officine Garibaldi, via Gioberti, 39 - 56124 Pisa (see map on page XI). Participants must bring their conference badge in order to to attend the conference dinner. Information about the purchased conference dinner ticket is embedded in the QR code on the conference badge.

GENERAL INFORMATION

Address of venue

- Polo didattico delle Piagge dell'Universita di Pisa (Polo Piagge), Via Giacomo Matteotti 3, 56124 Pisa. Some sessions will take place at the Polo Economia-Polo Didattico, and the Conference Center (Palazzo dei Congressi di Pisa, Via Giacomo Matteotti, 1 56124 Pisa). All the buildings are in close proximity to each other (see maps on page VIII).

Registration

The registration will be open from Thursday, 13th December 2018, to Sunday, 16th December 2018, from 8:15 to 18:15. The registration desk will be located in front of the Sala Convegni, at the ground floor of the Polo Piagge (see maps on pages VIII and IX).

Lecture rooms

The paper presentations will take place at the Polo Piagge and the Polo Economia-Polo Didattico of the University of Pisa. The list of rooms, location and their capacity is available in the interactive programme (see also floor maps at pages IX and X). Due to health and safety regulations the maximum capacity of the rooms should be respected.

The opening and the first four keynote talks will take place at the Palazzo dei Congressi (see map on page VIII). The closing keynote talk will take place at the Sala Convegni of the Polo Piagge (see floor map at page IX).

Presentation instructions

The lecture rooms will be equipped with a mini-pc and a computer projector. The session chairs should obtain copies of the talks on a USB stick before the session starts (use the lecture room as the meeting place), or obtain the talks by email prior to the start of the conference. Presenters must provide the session chair with the files for the presentation in PDF (Acrobat) format on a USB memory stick. This must be done at least ten minutes before each session. Chairs are requested to keep the sessions on schedule. Papers should be presented in the order they are listed in the programme for the convenience of attendees who may wish to go to other rooms mid-session to hear particular papers. In the case of a presenter not attending, please use the extra time for a break or a discussion so that the remaining papers stay on schedule. The session chairs are kindly requested to have a laptop for backup. An IT technician will be available during the conference and should be contacted in case of problems.

Posters

The poster sessions will take place at the Ground Level Hall of the Polo Piagge (see floor map on page IX). The posters should be displayed only during their assigned session. The authors will be responsible for placing the posters in the poster panel displays and removing them after the session. The maximum size of the poster is A0 portrait.

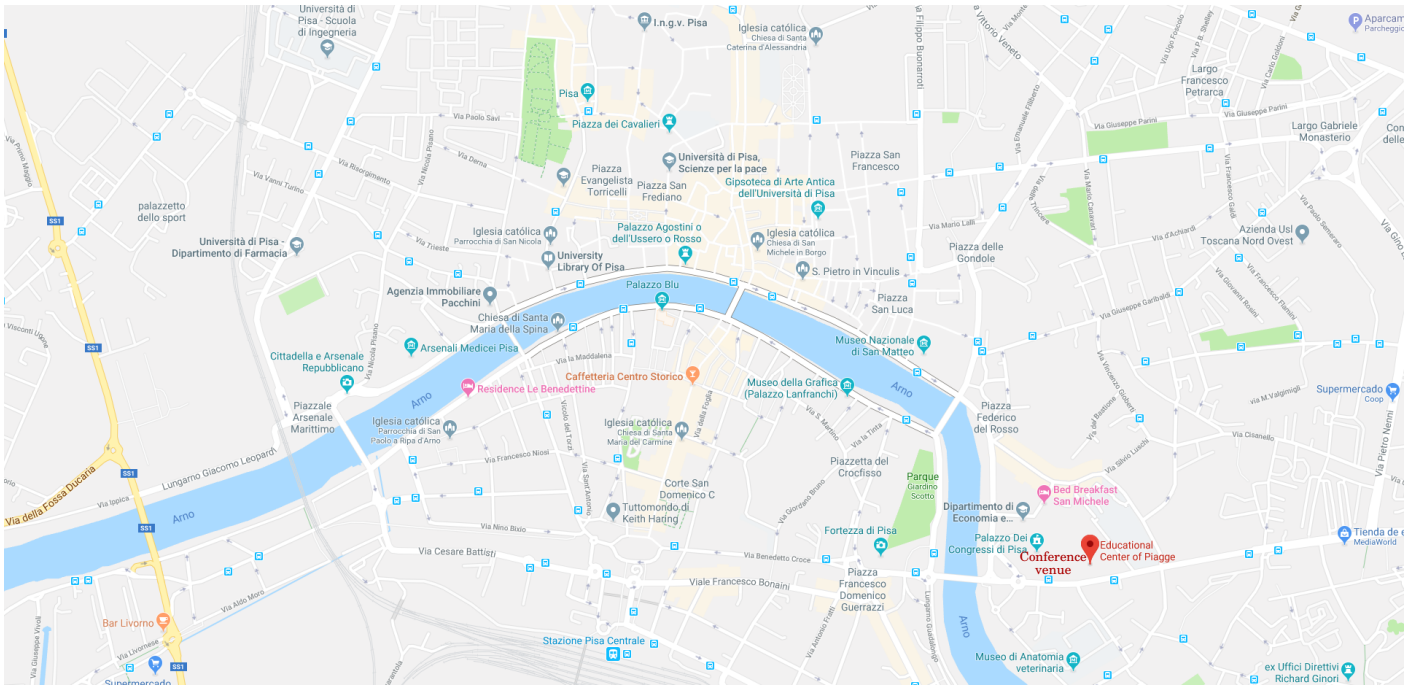
Internet Connection

Participants from any Eduroam-enabled institution should use the Eduroam service in order to obtain access to Internet. For participants without Eduroam access, there will be wireless Internet connection. You will need to have your own laptop in order to connect to the Internet. The login and password will be displayed on the announcement board by the registration desk.

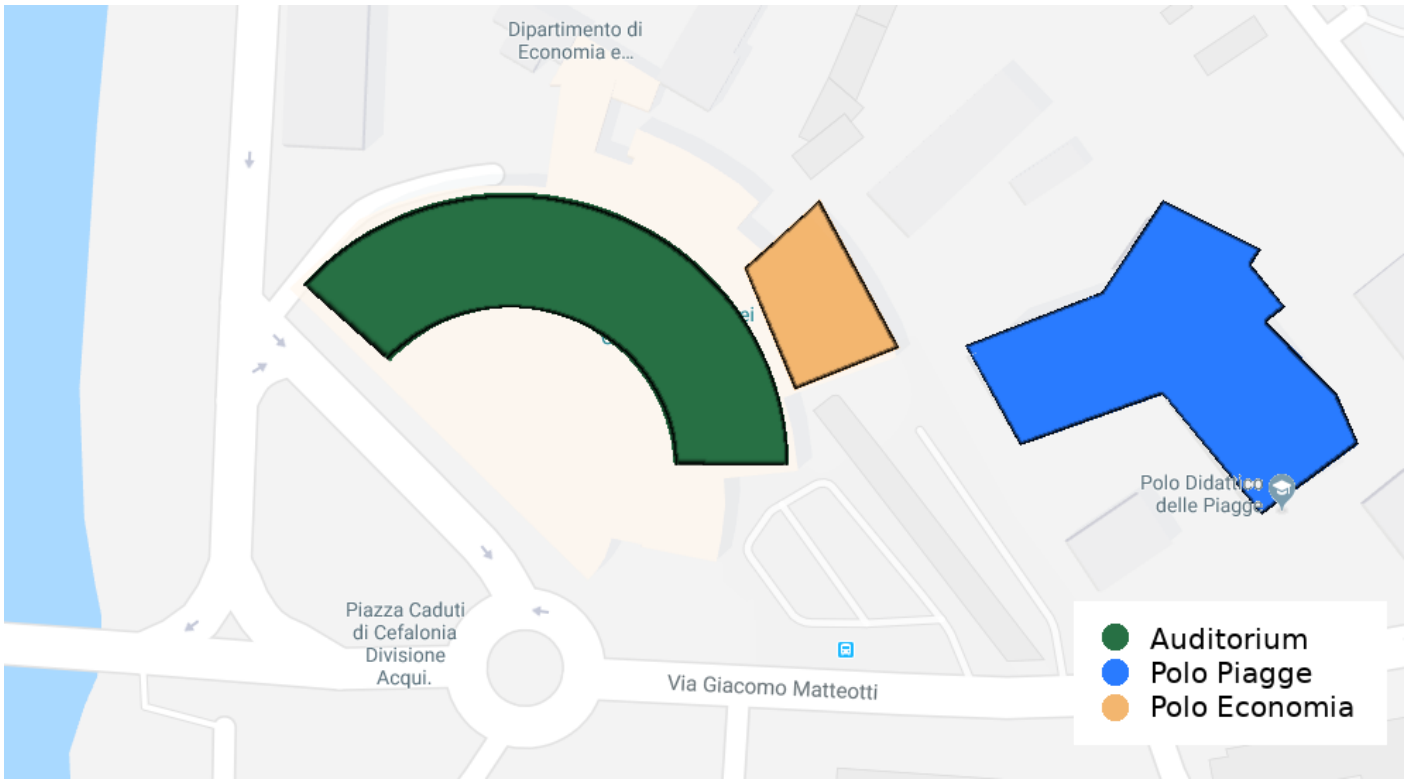
Exhibitors

Elsevier and Springer.

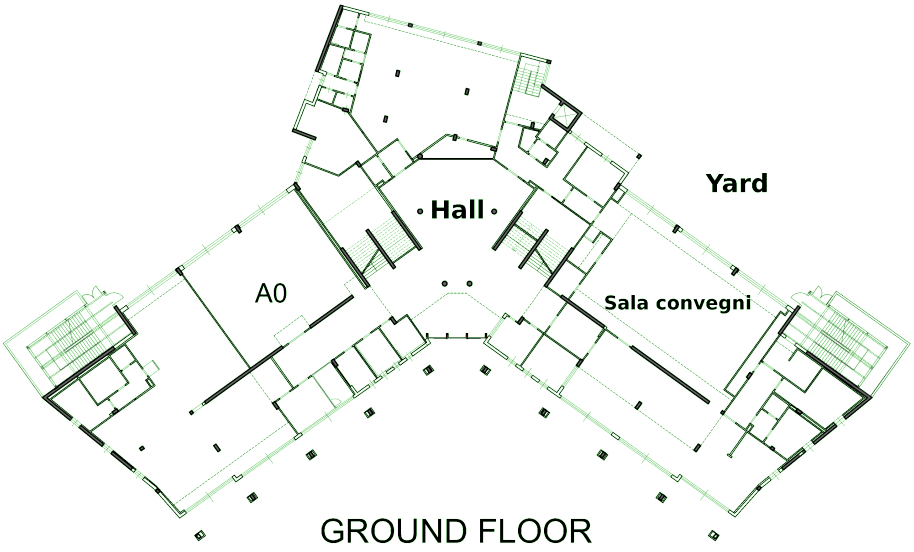
Map of the venue and nearby area

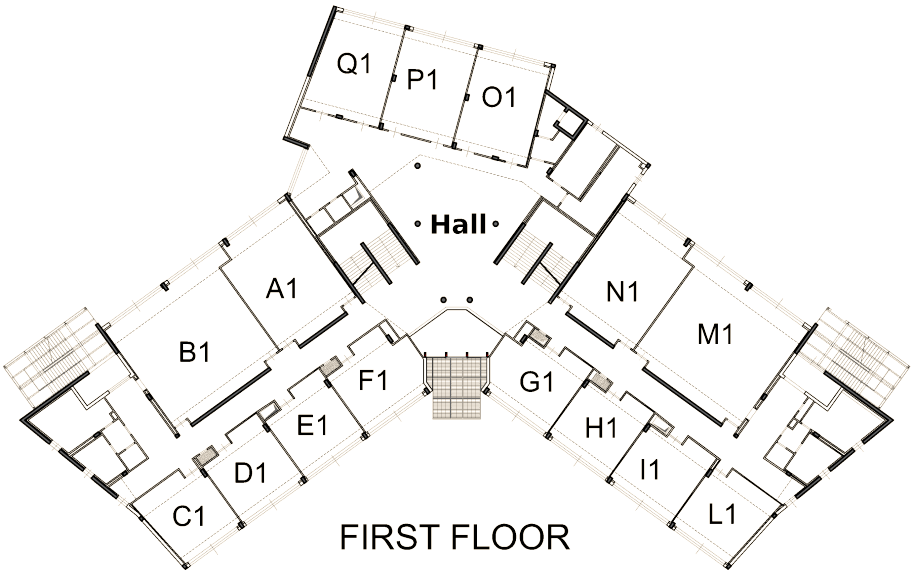


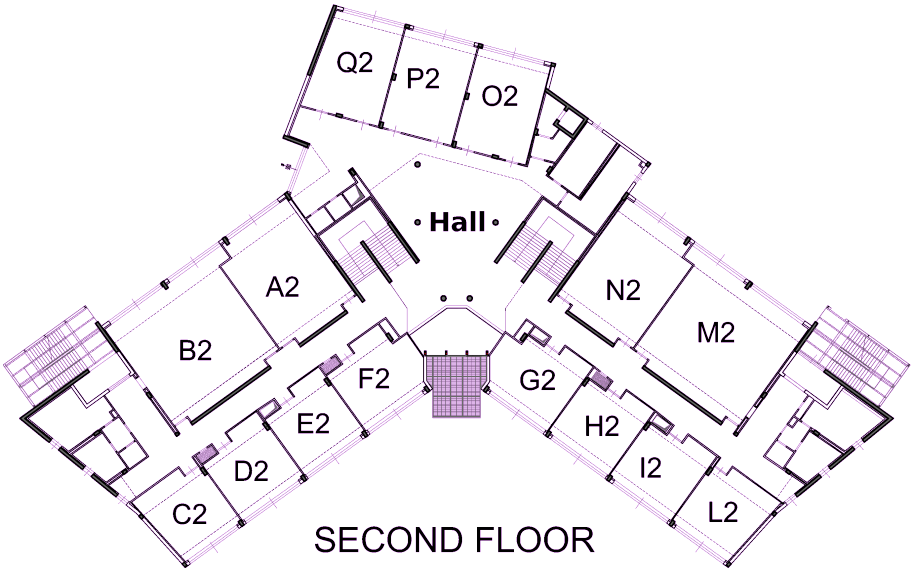
Map of the buildings



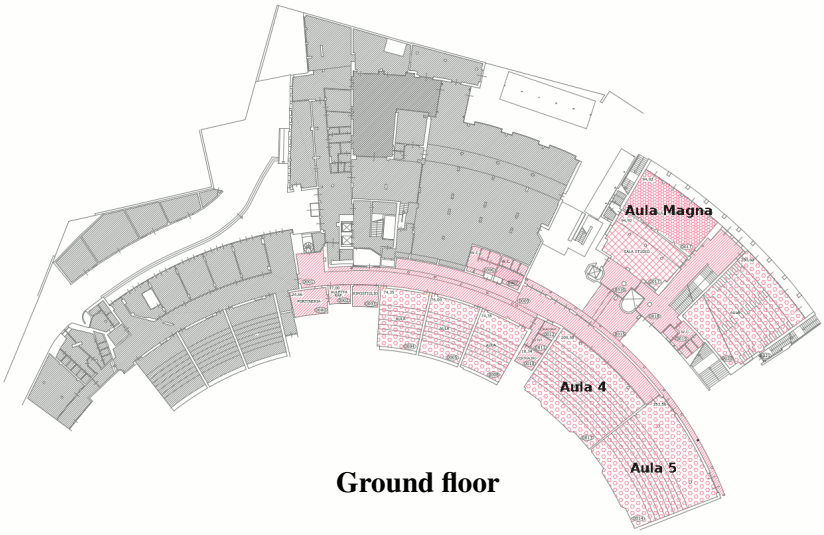
Polo Piagge floor maps



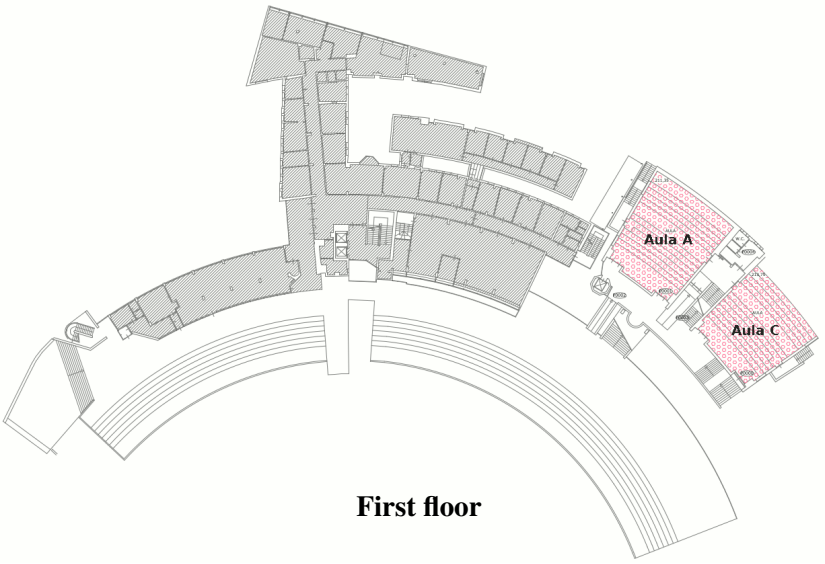




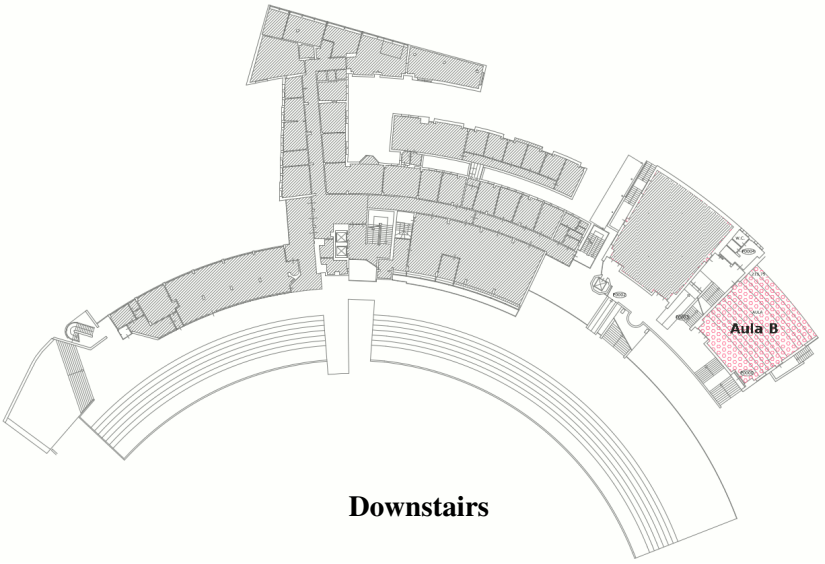
Polo Economia-Polo Didattico floor maps



Ground floor

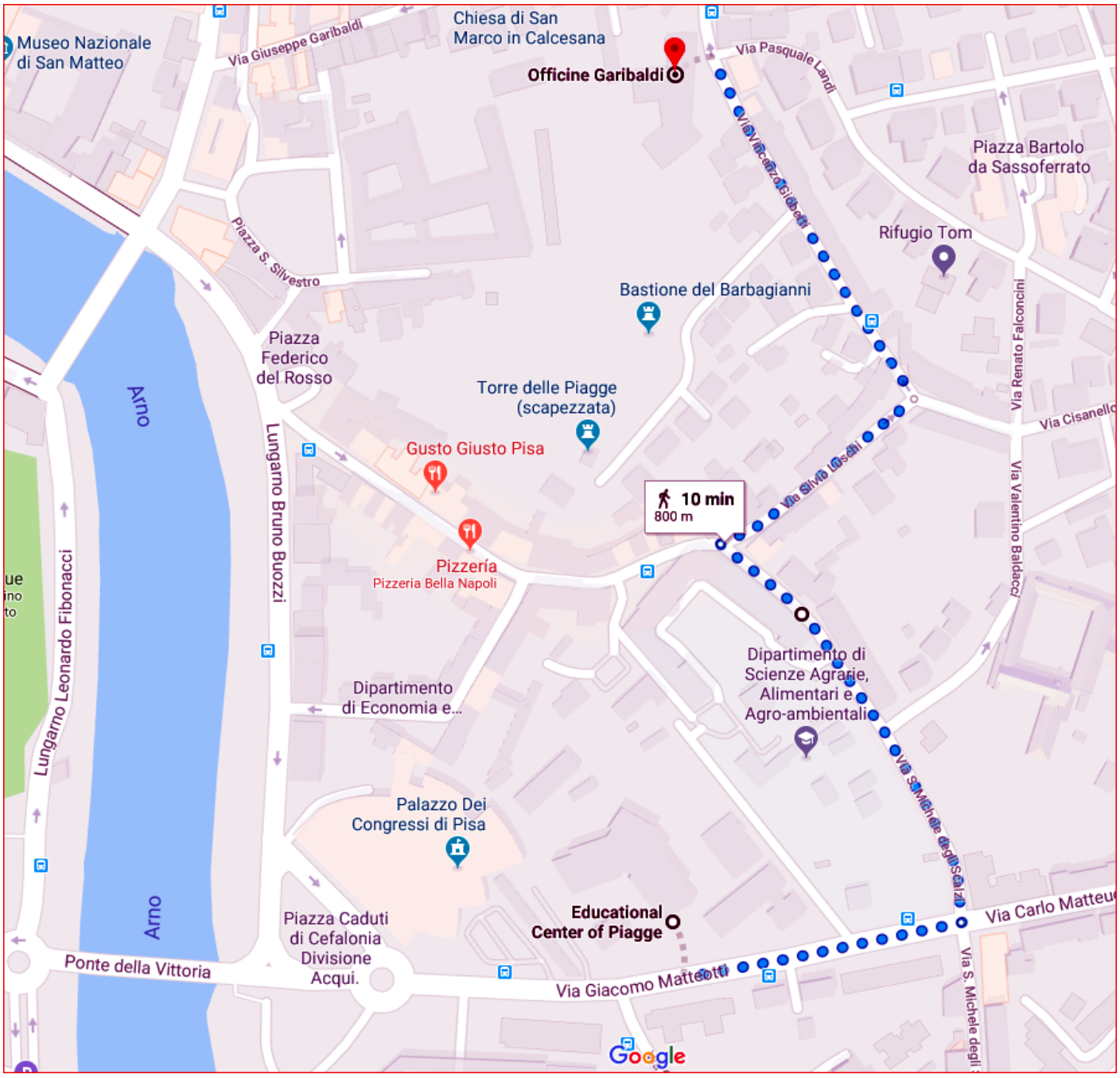


First floor



Downstairs

Conference dinner venue google map



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. In general, the interaction of mathematical methods and numerical implementations for the analysis of large and/or complex datasets arising in areas such as medicine, epidemiology, biology, psychology, climatology and communication is considered. 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 containing novel components in econometrics and statistics are encouraged to be submitted for publication in special peer-reviewed, or regular issues of the new Elsevier journal Econometrics and Statistics (EcoSta) and its supplement Annals of Computational and Financial Econometrics.

Papers should be submitted using the Elsevier Electronic Submission tool EES: http://ees.elsevier.com/ecosta (in the EES please select the appropriate special issue). For further information please consult http://www.cfenetwork.org or http://www.cmstatistics.org.

Call For Papers Computational Statistics & Data Analysis (CSDA)

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

Papers containing strong computational statistics, or substantive data-analytic elements, can also be submitted to special peer-reviewed, or regular issues of the journal Computational Statistics & Data Analysis (CSDA). Papers should be submitted using the Elsevier Electronic Submission tool EES: http://ees.elsevier.com/csda (in the EES please select the appropriate special issue). Any questions may be directed via email to: csda@dcs.bbk.ac.uk.

Contents

Friday 14.12.2018 09:00 - 09:50 Room: Auditorium Chair: Manfred Deistler Keynote talk 1

Group transformation models

Speaker: **Christian Gourieroux, University of Toronto and CREST, Canada** Alain Monfort, Jean-Michel Zakoian

Semi-parametric transformation models relate the endogenous variables to the errors through a parametric transformation depending on observed explanatory variables without specifying the error distribution. It is shown that in such model, called group transformation model (GTM), any pseudo maximum likelihood (PML) estimation approach provides consistent estimators of the sensitivity parameters of the explanatory variables, whenever artificial intercept parameters are introduced at appropriate places. This modelling principle with the associated PML estimation method is illustrated by several examples of application to multivariate ARCH models, qualitative models, Loss-Given-Default, omitted heterogeneity, seasonal adjustment, peer effect, or directional statistics.

Friday 14.12.2018 12:10 - 13:00 Room: Auditorium Chair: Ana Colubi Keynote talk 2

Varying-coefficient additive models: Two birds with one stone?

Speaker: **Jane-Ling Wang, University of California Davis, United States** Xiaoke Zhang

Both varying-coefficient and additive models have been widely adopted as non-parametric modeling approaches that enjoy flexibility and parsimony. An intriguing question is how to choose between these two models in practice. Recently, it was shown that this dichotomy can be altogether bypassed by embedding both models into a larger model, the varying-coefficient additive model (VCAM), which includes both models as special cases. However, that work was specifically designed for densely observed functional response with vector covariates. We show how to extend the VCAM model to more general settings that allow for sparsely observed functional responses, a.k.a. longitudinal data, and longitudinal covariates, in addition to vector covariates. A new algorithm is proposed and its performance is demonstrated through simulations and data applications. The algorithm involves non-convex maximization so the choice of the initial estimates plays a crucial role. We discuss several options and their empirical performance. Theoretical results are established for the nonparametric component functions of the model, including rates of convergence, and future directions will be discussed.

Sunday 16.12.2018 08:45 - 09:35 Room: Auditorium Chair: Miguel de Carvalho Keynote talk 3

The ups and downs of communicating statistics in an age of fragmented media and contested science

Speaker: **David Spiegelhalter, University of Cambridge, United Kingdom**

Those who value quantitative and scientific evidence are faced with claims both of a reproducibility crisis in scientific publication, and of a post-truth society abounding in fake news and alternative facts. In addition, scientists and institutions often exaggerate the importance of their work in order to gain publicity or advance an agenda. These issues are of vital importance to statisticians, and all are deeply concerned with trust in expertise. By considering the pipelines through which scientific and political evidence is propagated through the media, we will consider possible ways of improving both the trustworthiness of the statistical evidence being communicated, and the ability of audiences to assess the quality and reliability of what they are being told. The examples will include stories about the 'risks' of burnt toast and coffee, and whether there is 'no safe level of alcohol'.

Sunday 16.12.2018 12:20 - 13:10 Room: Auditorium Chair: Michele Guindani Keynote talk 4

Bayesian nonparametric updating of parametric models with Monte Carlo sampling

Speaker: **Chris Holmes, University of Oxford, United Kingdom**

Bayesian nonparametric learning of parametric models through the use of suitably randomized objective functions is discussed. Bayesian nonparametric credible regions, analogous to bootstrap confidence intervals, on parameters of objective functions such as likelihoods can exhibit better properties than their parametric counterparts, particularly when the models are wrong. Inference is achieved via parallelizable independent Monte Carlo posterior sampling of parameters, avoiding MCMC and issues such as burn in and chain dependence, and is highly scalable on modern computer architectures. We demonstrate the approach on a number of examples including nonparametric inference for Bayesian logistic regression, variational Bayes (VB), and Bayesian random forests.

Sunday 16.12.2018 18:15 - 19:05 Room: Sala Convegni Chair: Alessandra Luati Keynote talk 5

Regularized estimation of high dimensional auto- and cross-covariance matrices

Speaker: **Tommaso Proietti, University of Roma Tor Vergata, Italy**

The estimation of the (auto- and) cross-covariance matrices of a stationary random process plays a central role in prediction theory and time series analysis. When the dimension of the matrix is of the same order of magnitude as the number of observations and/or the number of time series, the sample cross-covariance matrix provides an inconsistent estimator. In the univariate framework, we proposed an estimator based on regularizing the sample partial autocorrelation function, via a modified Durbin-Levinson algorithm that receives as an input the banded and tapered sample partial autocorrelations and returns a consistent and positive definite estimator of the autocovariance matrix; also, we established the convergence rate of the regularized autocovariance matrix estimator and characterised the properties of the corresponding optimal linear predictor. The multivariate generalization is based on a regularized Whittle algorithm, shrinking the lag structure towards a finite order vector autoregressive system (by penalizing the partial canonical correlations), on the one hand, and shrinking the cross-sectional covariance towards a diagonal target, on the other.

Friday 14.12.2018 10:20 - 12:00 Parallel Session B – CFE-CMStatistics

EI003 Room A0 ADVANCES IN ROBUST STATISTICS **Chair: Mia Hubert**

E0171: **MacroPCA: An all-in-one PCA method allowing for missing values as well as cellwise and rowwise outliers**
Presenter: **Peter Rousseeuw**, KU Leuven, Belgium
Co-authors: Mia Hubert, Wannes Van den Bossche

Multivariate data are typically represented by a rectangular matrix (table) in which the rows are the objects (cases) and the columns are the variables (measurements). When there are many variables one often reduces the dimension by principal component analysis (PCA), which in its basic form is not robust to outliers. Much research has focused on handling rowwise outliers, i.e. rows that deviate from the majority of the rows in the data (for instance, they might belong to a different population). In recent years also cellwise outliers are receiving attention. These are suspicious cells (entries) that can occur anywhere in the table. Even a relatively small proportion of outlying cells can contaminate over half the rows, which causes rowwise robust methods to break down. A new PCA method is constructed which combines the strengths of two existing robust methods in order to be robust against both cellwise and rowwise outliers. At the same time, the algorithm can cope with missing values. It is the only PCA method that can deal with all three problems simultaneously. Its name MacroPCA stands for PCA allowing for Missing And Cellwise & Rowwise Outliers. Several simulations and real data sets illustrate its robustness. New residual maps are introduced, which help to determine which variables are responsible for the outlying behavior. The method is well-suited for online process control.

E0172: **Exploring compositional data through monitoring robust estimates and dynamic graphics in R**
Presenter: **Valentin Todorov**, UNIDO, Austria

A technique for monitoring robust estimates computed over a range of key parameter values have been proposed recently. Through this approach the diagnostic tools of choice can be tuned in such a way that highly robust estimators which are as efficient as possible are obtained. Key tool for detection of multivariate outliers and for monitoring of robust estimates are the scaled Mahalanobis distances and statistics related to these distances. However, the results obtained with this tool in case of compositional data might be unrealistic. Compositional data are closed data, i.e. they sum up to a constant value (1 if expressed as proportions or 100 if expressed as percentages). This constraint makes it necessary to find a transformation of the data from the so called simplex sample space to the usual real space. To illustrate the problem of monitoring compositional data, we start with a simple example and then, we analyze a real life data set presenting the technological structure of manufactured exports which, as an indicator of their quality, is an important criterion for understanding the relative position of countries measured by their industrial competitiveness. The analysis is conducted with the R package fsdaR, which makes the analytical and graphical tools provided in the MATLAB FSDA library available for R users.

E0617: **The power of monitoring: How to make the most of a contaminated multivariate sample**
Presenter: **Marco Riani**, University of Parma, Italy
Co-authors: Anthony Atkinson, Andrea Cerioli, Aldo Corbellini

Diagnostic tools must rely on robust high-breakdown methodologies to avoid distortion in the presence of contamination by outliers. However, a disadvantage of having a single, even if robust, summary of the data is that important choices concerning parameters of the robust method, such as breakdown point, have to be made prior to the analysis. The effect of such choices may be difficult to evaluate. We argue that an effective solution is to look at several pictures, and possibly to a whole movie, of the available data. This can be achieved by monitoring, over a range of parameter values, the results computed through the robust methodology of choice. We show the information gain that monitoring provides in the study of complex data structures through the analysis of multivariate datasets and using different high-breakdown techniques. Our findings support the claim that the principle of monitoring is very flexible and that it can lead to robust estimators that are as efficient as possible. We also address through simulation some of the tricky inferential issues that arise from monitoring.

EO102 Room A1 SPATIO-TEMPORAL VARIATIONS IN SOCIAL AND EPIDEMIOLOGICAL DATA **Chair: Veronica Berrocal**

E0343: **A stratified age-period-cohort model for spatial heterogeneity in all-cause mortality**
Presenter: **Theresa Smith**, University of Bath, United Kingdom

A common goal in modelling demographic rates is to compare two or more groups. For example comparing mortality rates between men and women or between geographic regions may reveal health inequalities. A popular class of models for all-cause mortality as well as incidence of specific diseases like cancer is the age-period-cohort (APC) model. Extending this model to the multivariate setting is not straightforward, because the univariate APC model suffers from well-known identifiability problems. Often APC models are fit separately for each stratum, and then comparisons are made post hoc. A stratified APC model is introduced to directly assess the sources of heterogeneity in mortality rates using a Bayesian hierarchical model with matrix-normal priors that share information on linear and nonlinear aspects of the APC effects across strata. Computing, model selection, and prior specification are addressed and the model is then applied to all-cause mortality data from the European Union.

E0443: **Bayesian disaggregation of spatio-temporal community indicators estimated via surveys**
Presenter: **Veronica Berrocal**, University of Michigan, United States

The American Community Survey (ACS) is an ongoing survey administered by the US Census Bureau which collects social, economic, and other community data. ACS estimates are released annually, with varying spatial and temporal resolution: 5-year time periods refer to smaller municipal subdivisions, while 1-year time periods refer to larger areas. Although for epidemiological studies, these estimates contain important community information, their varying spatial and temporal resolution pose various challenges: the 5-year ACS estimates might be temporally misaligned with finely resolved health outcome data, conversely, the coarser 1-year estimates are likely spatially misaligned with finely resolved health data. We present a Bayesian hierarchical model that leverages both 1-year and 5-year ACS data and accounts for the survey sampling design to obtain estimates of community health indicators at any given spatial and temporal resolution. The disaggregation is achieved by introducing a latent, point-referenced process, in turn modeled using a multi-resolution basis function expansion, which is linked to the ACS data via a stochastic model that accounts also for the survey design used to collect the data.

E0514: **Spatial clustering of average risks and risk trends in Bayesian disease mapping**
Presenter: **Craig Anderson**, University of Glasgow, United Kingdom
Co-authors: Nema Dean

Spatio-temporal disease mapping focuses on estimating the spatial pattern in disease risk across a set of non-overlapping areal units over a fixed period of time. The key aim is to identify areas which have a high average level of disease risk or where disease risk is increasing over time, thus allowing public health interventions to be focused on these areas. Such aims are well suited to the statistical approach of clustering, and while much research has been done in this area in a purely spatial setting, only a handful of approaches have focused on spatio-temporal clustering of disease risk. We outline a new modelling approach for clustering spatio-temporal disease risk data, by clustering areas based on both their mean

risk levels and the behaviour of their temporal trends. The efficacy of the methodology is established by a simulation study, and is illustrated by a study of respiratory disease risk in Glasgow, Scotland.

E0745: **Time-varying step change detection and forecasting in spatio-temporal areal data models**
Presenter: **Gavino Puggioni**, University of Rhode Island, United States
New methods are proposed that address some common issues in epidemiological studies: time varying step change detection in spatial autocorrelation, and short and medium term forecast stability. The goal is to provide a flexible framework to identify and target areas with increased risk, and to inform early warning systems for risk surveillance. We present a flexible two stage space-time CAR model with Bayesian model averaging on the set of predictors, and a stochastic process that models variations in step-change boundaries. After testing the method in a simulation study, we apply it to real data, and compare its forecasting performance with other commonly used methods.

EO388 Room B1 GRAPHICAL MARKOV MODELS I: MULTIVARIATE DEPENDENCE STRUCTURES Chair: Monia Lupparelli

E0881: **On the interpretation of path weights in undirected Markov random fields**
Presenter: **Alberto Roverato**, University of Bologna, Italy
Co-authors: Robert Castelo
In graphical Gaussian models an undirected graph is used to represent the association structure of variables as a network, and if a pair of variables is not joined by an edge in the graph, then the corresponding partial correlation is equal to zero. Although in graphical Gaussian models the structure of the network can be inferred from the zero pattern of the inverse covariance matrix, if the probability distribution of the variables is faithful to the network, then paths along the network connect random variables with non-zero entries in the covariance matrix. In the analysis of graphical Gaussian models, it has been associated a weight with every path in the network and showed that the covariance between two variables can be computed as the sum of the weights of all the paths joining the two variables. Path weights allow one to identify the relative contribution of a path to the value of the corresponding covariance. However, it is not clear either how to interpret the value of a single path or how to compare two paths with different endpoints. We provide an interpretation of the value taken by the weight of a path by decomposing it into a partial weight and an inflation factor. Furthermore, we identify a class of paths, called chordless paths, whose weights have a remarkably straightforward interpretation.

E0959: **Regression modelling with I-priors**
Presenter: **Wicher Bergsma**, London School of Economics, United Kingdom
The I-prior modelling approach for regression with multiple, possibly multidimensional covariates, and with possible interaction effects, is introduced. The I-prior is a maximum entropy Gaussian prior for the regression function, with covariance function proportional to the Fisher information on the regression function. The proposed approach is a general, practical, methodology unifying a variety of models, including multilevel, varying coefficient, longitudinal, and multidimensional or functional response models. In contrast to Gaussian process regression, a simple EM algorithm can be constructed for I-prior models. This is especially important when there are many hyperparameters, when direct optimization of the marginal likelihood may be difficult. The approach has high model parsimony, in particular for models involving many interaction effects. As a consequence of this model parsimony, we obtain a simple semi-Bayes methodology for selecting interaction effects. Whereas in previous approaches the reproducing kernel Hilbert space framework was adequate, in the I-prior approach it is necessary to consider regression functions in a reproducing kernel Krein space.

E1198: **Multivariate dependence structures for ordinal data: A ϕ-divergence based approach**
Presenter: **Maria Kateri**, RWTH Aachen University, Germany
Dependence structures among ordinal variables will be studied in connection to ϕ-divergence measures. Log-linear models for ordinal classification variables will be redefined through the Kullback-Leibler divergence and embedded in generalized families of models derived by replacing the Kullback-Leibler by the ϕ-divergence. The scaling role of the ϕ-divergence in constructing models for ordinal data and its effect on describing the underlying dependence structure will be discussed. The focus will be on high-dimensional contingency tables. Representative applications for members of the ϕ-divergence based model families will be presented.

E0696: **Some issues on Bayesian analysis of binary bidirected graphs**
Presenter: **Claudia Tarantola**, University of Pavia, Italy
Bayesian analysis of binary bidirected graphs has not been developed as much as traditional methods. No conjugate analysis is available and MCMC methods must be employed. The likelihood of the model cannot be analytically expressed as a function of the marginal log-linear interactions, but only in terms of the probability parameters. Hence, at each step of the MCMC an iterative procedure needs to be applied in order to calculate the cell probabilities and consequently the model likelihood. Finally, in order to have a well-defined model of marginal independence, the considered MCMC algorithm should generate parameter values leading to a joint probability distribution with compatible marginals. We will present a novel MCMC strategies that handles the previously discussed problems. A simulation study will be discussed.

EO426 Room D1 INSTRUMENTAL VARIABLES: THEORY AND APPLICATIONS Chair: Federico Crudu

E0218: **Errors-in-variables models with many proxies**
Presenter: **Federico Crudu**, University of Siena, Italy
A novel method is introduced to estimate linear models when explanatory variables are observed with error and many proxies are available. The empirical Euclidean likelihood principle is used to combine the information that comes from the various mismeasured variables. We show that the proposed estimator is consistent and asymptotically normal. In a Monte Carlo study we show that our method is able to efficiently use the information in the available proxies, both in terms of precision of the estimator and in terms of statistical power. An application to the effect of police on crime suggests that measurement errors in the police variable induce substantial attenuation bias. Our approach, on the other hand, yields large estimates in absolute value with high precision, in accordance with the results put forward by the recent literature.

E0285: **Nonparametric instrumental estimation of additive models**
Presenter: **Samuele Centorrino**, Stony Brook University, United States
Co-authors: Sorawoot Srisuma
A two-step estimator is proposed for nonparametric additive regression functions with multiple endogenous and exogenous conditioning variables. In the first step we construct a sieve nonparametric instrumental variable estimator that achieves the optimal rate of convergence in a minimax sense. We smooth this over in the second step using kernel methods. The subsequent estimator has an asymptotic normal distribution and has an oracle property. In particular, the asymptotic distribution of each additive component is the same as it would be if all the other components were known.

E0305: **Inference in instrumental variables models with heteroskedasticity and many instruments**
Presenter: **Giovanni Mellace**, University of Southern Denmark, Denmark
Co-authors: Federico Crudu, Zsolt Zandor

A specification test is proposed for instrumental variable models that is robust to the presence of heteroskedasticity. The test can be seen as a generalization the Anderson-Rubin test. Our approach is based on the jackknife principle. We are able to show that under the null the proposed statistic has a Gaussian limiting distribution. Moreover, a simulation study shows its competitive finite sample properties in terms of size and power. Finally, we provide an empirical application using college proximity instruments to estimate the returns of education.

E0352: **A strategy to reduce the count of moment conditions in panel data GMM**
Presenter: **Irene Mammi**, Ca' Foscari University of Venice, Italy
Co-authors: Maria Elena Bontempi

The problem of instrument proliferation and its consequences are well known. The literature provides little guidance on how many instruments is too many. Commonly employed strategies to alleviate the instrument proliferation problem, such as lag-depth truncation and/or collapsing of the instrument set, involve either a certain degree of arbitrariness or untested restrictions on the instrument matrix. A statistically-grounded and data-driven strategy to reduce the instrument count is introduced. It applies the principal component analysis (PCA) on the instrument matrix and exploits the PCA scores as a new instrument set for generalized method-of-moments (GMM) estimation of dynamic panel data models. Through extensive Monte Carlo simulations, the performances of the difference GMM, level and system GMM estimators are assessed, when lag truncation, collapsing and principal component-based IV reduction are performed on the instrument set. Empirical applications complement the analysis. Results show that principal component-based IV reduction is a promising strategy to lower the instrument count.

EO584 Room E1 STATISTICAL MODELS AND INFERENCE WITH NETWORK DATA Chair: Donglin Zeng

E0316: **Estimating heterogeneous biomarker networks and effects on disease outcomes**
Presenter: **Yuanjia Wang**, Columbia University, United States

Biomarkers are often organized into networks, in which the strengths of network connections vary across subjects depending on subject-specific covariates (e.g., genetic variants). Variation of brain network connections, as subject-specific feature variables, has been found to predict disease clinical outcomes. We develop a two-stage statistical method to estimate covariate-dependent brain networks to account for heterogeneity among network measures and evaluate their association with disease clinical manifestation. In the first stage, we propose a conditional Gaussian graphical model with mean and precision matrix depending on covariates to obtain subject- specific networks. In the second stage, we associate a subject biomarker network measure (connection strengths) estimated from the first step along with the biomarkers and covariates to identify important features of a clinical outcome. The second stage allows us to evaluate the improvement in predictiveness of adding network measures compared to using biomarkers and covariates alone. We assess the performance of the proposed method by extensive simulation studies and apply the method to a Huntington's disease (HD) study to investigate the effect of the HD causal gene on the rate of change in motor symptom as mediated through brain subcortical and cortical gray matter atrophy connections.

E0317: **Learning directed acyclic graphs with mixed effects structural equation models from observational data**
Presenter: **Donglin Zeng**, University of North Carolina at Chapel Hill, United States
Co-authors: Yuanjia Wang

The identification of causal relationships between random variables from large-scale observational data using directed acyclic graphs (DAG) is highly challenging. We propose a new mixed-effects structural equation model (mSEM) framework to estimate subject-specific DAGs, where we represent joint distribution of random variables in the DAG as a set of structural causal equations with mixed effects. The directed edges between nodes depend on observed exogenous covariates on each of the individual and unobserved latent variables. The strength of the connection is decomposed into a fixed-effect term representing the average causal effect given the covariates and a random effect term representing the latent causal effect due to unobserved pathways. We propose a penalized likelihood-based approach to handle high dimensionality of the DAG model and a fast computational algorithm to achieve desirable sparsity by hard-thresholding the edges. We theoretically prove the identifiability of mSEM. Using simulations and an application to protein signaling data, we show substantially improved performance when compared to existing methods and consistent results with a network estimated from interventional data. Lastly, we identify gray matter atrophy networks in regions of brain from patients with Huntington's disease and corroborate our findings using white matter connectivity data collected from an independent study.

E0578: **Toward valid causal and statistical inference with social network data**
Presenter: **Elizabeth Ogburn**, Johns Hopkins University, United States

Interest in and availability of social network data has led to increasing attempts to make causal and statistical inferences using data collected from subjects linked by social network ties. When social relations can engender dependence in the variables of interest, treating such observations as independent results in invalid, anticonservative statistical inference, but there is a dearth of methods that can account for this kind of dependence. We develop a test for network dependence that can be used to screen for the appropriateness of i.i.d. statistical methods and apply it to data from the Framingham Heart Study (FHS). Our results suggest that some of the many decades worth of research on coronary heart disease and other health outcomes using FHS data could be invalid due to unacknowledged network dependence. We also extend recent work on causal inference for causally connected units to more general social network settings: we describe estimation and inference for causal effects that are specifically of interest in social network settings, and our asymptotic results allow for dependence of each observation on a growing number of other units as sample size increases.

E1563: **Estimating the effects of match-object covariates in a generalized Bradley–Terry model**
Presenter: **Muhammad Hilmi Bin Abdul Majid**, University of Warwick, United Kingdom
Co-authors: Chenlei Leng

In paired comparison data, pairs of object within a set are compared and for each comparison, there will be a winner and a loser. This is a special case of network data whereby it is known that there is a directed edge (match) between two objects but the direction of the edge (match outcome) is unknown. A generalized Bradley–Terry model that includes match-object specific covariates will be introduced. Due to the incidental parameter problem, the maximum likelihood estimate for the covariates are biased and statistical inference will be invalid. A conditional likelihood estimate using simple cycles of length three can be used to remove the bias and is proven to be consistent and asymptotically normal.

EO392 Room F1 LABEL NOISE ISSUES IN STATISTICS AND MACHINE LEARNING Chair: Efoevi Angelo Koudou

E0717: **Weighted performance evaluation of classifiers with a noisy ground truth**
Presenter: **Ilias Benjelloun**, Universita de Lorraine, France

In classification, one important task is to evaluate accurately what has been learned. The goal is to obtain an estimate of how well the classifier produced by the learning algorithm performs on unseen data, or if it is better than another already established learning algorithm. The less the bias and the lower the variance, the better the estimate. Over the past few decades, many performance measures have been proposed, and different evaluation procedures and statistical tests have been developed and used. For example, with limited test data, to obtain a good estimate (with low variance) of the performance of a learning algorithm, the standard way is to perform a 10-fold cross-validation procedure. However, it is rare when the quality of the test data is taken into account. Indeed, mislabelling errors affect the estimates in terms of bias, leading to overestimate the performance of some classifiers while underestimating those of others. We are interested in taking advantage of existing denoising ensembling methods to design an evaluation procedure that is less affected by noisy data. The procedure is then used to evaluate the performance of classifiers in a setting where noise is artificially introduced in the data, and is compared with traditional evaluation procedures. The effect of different types of noise on the evaluation methods are also studied.

E0845: **Consequences and assessment of label noise**
Presenter: **Benoit Frenay**, University of Namur, Belgium

When label noise pollutes a dataset, most machine learning algorithms will be affected. The consequences of label noise are diverse and well documented in the literature. They are reviewed in details, including changes in classification performances, learning requirements, complexity of learned models, observed frequencies and feature relevance. Then, we will show how a simple, generic probabilistic framework can be used to mitigate the impact of label noise in tasks such as classification, segmentation and feature selection. In practice, experimental validation of label noise tolerant algorithms is not trivial. There exist only a few datasets with clearly identified label noise and most works in the literature use simple random, uniform label noise that may not be realistic. Finally, methods that deal with label noise will be assessed.

E1011: **Some issues on ranking of classifiers in the presence of label noise**
Presenter: **Efoevi Angelo Koudou**, IECL CNRS /Universite de Lorraine, France
Co-authors: Bart Lamiroy

Few instances are reviewed on how the problem of evaluating the performance of classifiers in the presence of noisy labels has been addressed in the recent literature. Some interesting attempts to solve this problem has been carried out under assumptions that could appear to be unrealistic. For instance, some authors used an approach based on the fact that the classifier error is independent of that of the labeller. We discuss how this independence assumption could be relaxed.

E1156: **Distribution dependent learning with asymmetric label noise**
Presenter: **Henry Reeve**, University of Birmingham, United Kingdom
Co-authors: Ata Kaban

Finite sample bounds are presented for a nearest neighbour based algorithm in the presence of unknown asymmetric label noise. Our first result shows that minimax optimal rates are attained whenever the regression function is Lipschitz continuous and the marginal density is uniformly bounded away from zero. In particular, fast rates may be attained whenever Tysbakov's noise condition holds. We then consider the more general non-compact setting in which the density may be arbitrarily close to zero within its support. In this setting, learning with label noise becomes more challenging and depends heavily upon the behaviour of the marginal distribution in neighbourhoods of the regression functions extrema.

EO262 Room G1 SURVIVAL ANALYSIS Chair: Ingrid Van Keilegom

E0267: **Flexible parametric generalised additive survival models with informative censoring**
Presenter: **Robinson Dettoni**, University College London, United Kingdom
Co-authors: Giampiero Marra, Rosalba Radice

Most estimation methodologies for censored time to event data assume that censoring is non-informative. In many applications, the censoring scheme may be in effect informative. The aim is to introduce a survival model with informative censoring which is flexible and easy to apply. The estimation of such models poses several challenges, and we propose a penalized maximum likelihood approach to this end. This framework allows us to incorporate the information provided by the censoring times to improve the efficiency of the proposed estimator. We also estimate the baseline functions flexibly via means of monotonic P-splines. Covariate effects are flexibly determined using additive predictors. Such a framework allows one to calculate easily several quantities of interest and their variances, such as time-dependent hazard or odds ratios, which would otherwise be difficult to obtain with a non-parametric approach. Confidence intervals for linear and non-linear functions of the model's coefficients, with good finite sample properties, are also provided. Information criteria and cross-validation can be employed to detect informative censoring in applications. The performance of the proposed method is investigated theoretically and via simulation studies. Both theory and simulation highlight the usefulness of the proposal. The proposed framework has been implemented in the R package GJRM, and applied to data on infants hospitalized for pneumonia.

E1126: **Flexible parametric model for survival data subject to dependent censoring**
Presenter: **Negera Wakgari Deresa**, KU Leuven, Belgium
Co-authors: Ingrid Van Keilegom

When modeling survival data, it is common to assume that the (log-transformed) survival time (T) is conditionally independent of the (log-transformed) censoring time (C) given a set of covariates. There are numerous situations in which this assumption is in doubt, and a number of correction procedures have been developed for different models. However, in most cases, some prior knowledge about the association between T and C is required. When neither prior knowledge nor auxiliary information is available, the application of many existing methods turns out to be limited. We develop a flexible parametric model to estimate the association between T and C, without any additional information. We show that the association between T and C is identifiable. The performance of the proposed method is investigated both asymptotically and through finite sample simulations. We also develop a diagnostic plot approach to assess the quality of the fitted model. Finally, the approach is illustrated on real data coming from a study on liver transplantations.

E0778: **Joint modelling of longitudinal and survival data**
Presenter: **Eleni-Rosalina Andrinopoulou**, Erasmus Medical Center, Netherlands

In epidemiological follow-up studies different types of outcomes are typically collected for each individual. These include longitudinally measured responses (e.g., biomarkers), and the time until an event of interest occurs (e.g., death, intervention). Often these outcomes are separately analysed, but in many occasions, it is of scientific interest to study their association. This type of interest has given rise in the class of joint models for longitudinal and time-to-event data. Joint models can be used when focusing either on the survival outcome when we wish to account for the effect of an endogenous time-dependent covariate or on the longitudinal outcome, and we wish to correct for non-random dropout. The idea behind these

models is to couple a survival model for the time-to-event process with a mixed-effects model for the longitudinal outcome. Several extensions of the standard joint model that consists of one longitudinal and one survival outcome have been proposed including among others the use of multiple longitudinal outcomes and the investigation of different manners to associate the longitudinal and the survival process. Several applications of these type of models will be discussed.

E0604: **Joint modeling of floral transition time and leaf appearance process in maize plant**
Presenter: **Sandra Plancade**, INRA / Catholic University of Louvain-la-Neuve, France
Co-authors: Anouar El Ghouch, Sylvie Huet, Christine Dillmann
Plant growth is usually modeled through phases which affect simultaneously various phenotypes, including external traits that can be repeatedly observed, as well as internal traits that require destructive observations, and a question of interest is to get information on the latter based on observations of the former. The focus will be on the estimation of two traits of maize plant development: the floral transition time, a key-step occurring inside the stem, and the process of leaf appearance or phyllochrone, based on repeated measurements of the number of leaves for a set of plants, as well as dissections indicating whether floral transition already occurred; thus, floral transition and leaf appearance times are subject to current status and interval censoring mechanisms respectively. In literature, censoring is usually circumvented by aggregating measurements from a plot of plants selected as homogeneous, in order to create a "pseudo-plant". We propose an alternative approach based on an accelerated failure time model estimated via an expectation maximization algorithm, which allows us to account for plant-level variability.

EO144 Room H1 SAMPLING: PLANNING, DESIGN, MODELING, INFERENCE AND APPLICATIONS Chair: Subir Ghosh

E0558: **Pseudo-population bootstrap for design-based inference on spatial phenomena**
Presenter: **Lorenzo Fattorini**, University of Siena, Italy
Spatial prediction on continuous surfaces or for finite populations of points or spatial units are usually performed in a model-dependent framework, assuming spatial processes generating the values of the survey variable for depicting his map on the whole study area. Recently, the properties of inverse distance weighting interpolation were derived in a completely design-based framework. Asymptotic scenarios and conditions ensuring design-based asymptotic unbiasedness and consistency were derived. They mainly require smoothness of the variable under study and the use of spatially balanced sampling schemes. As the resulting map constitutes a pseudo-population converging to the true spatial population, it can be adopted in a resampling procedure, selecting bootstrap samples from it by means of the spatial scheme actually adopted to select the sample from which interpolation has been performed. The procedure can be used to make inference on the distribution of complex estimators of the spatial population parameters as well as to make inference on the estimated map.

E1055: **Near optimum and average optimum variance estimation and prediction in small area estimation**
Presenter: **Subir Ghosh**, University of California, United States
Optimum variance component estimation methods are used in small area estimation for finding the best linear unbiased predictors. Near optimum estimation and average optimum estimation methods are proposed when the exact optimum estimators do not exist or difficult to obtain. Two estimation methods provide the closed form expressions for their estimators. The robustness properties of the proposed estimators, in comparison with the other estimators, are also investigated, using data simulated from a skew normal distribution.

E1072: **Small area model-based estimation using big data: Applications**
Presenter: **Stefano Marchetti**, Dipartimento di Economia e Management, Universita di Pisa, Italy
National statistical offices aim to produce statistics for citizen and policy-makers. Survey sampling has been recognized to be an effective method to obtain timely and reliable estimates for a specific area in socio-economic fields. Usually, it is important to infer population parameters at a finer area level, where the sample size is small and does not allow for reliable direct estimates. Small area estimation (SAE) methods by means of auxiliary variables allow as to obtain reliable estimates when the direct o are unreliable. SAE methods can be classified into unit-level and area-level models: Unit-level models require a common set of auxiliary variables between survey and census/registers known for all the population units; area-level models are based on direct estimates and aggregated auxiliary variables. Privacy policies and high census costs make difficult the use of unit-level data, particularly out of the statistical offices. Aggregated auxiliary variables from different sources are more easily available and can be used in are-level models. Big data a collection of data that contains greater variety arriving in increasing volumes and with ever-higher velocity adequately processed can be used as auxiliary variables in area-level models. We show two applications of SAE: the use of mobility data to estimates poverty incidence at local level in Tuscany, Italy and the use of twitter data to estimate the share of food consumption expenditure at the province level in Italy.

E1153: **Poststratifying on variables measured with error**
Presenter: **Daniel Oberski**, Utrecht University, Netherlands
Samples, by design or by accident, often do not reproduce known population totals on average. To solve this potential problem, a common approach is reweighting, with weights calculated based on the disparity between the totals that are known and those that are measured. But when measured variables are error-prone, weights cannot be calculated. For example, when weighting a social media population based on US State, how should we account for the fact that social media users may give incorrect information about their state of residence? A novel method is introduced to postratify based on variables measured with error, the "Mixture of States with Poststratification" (Ms. P). The method is applied to a nonprobability sample of 73,000 Facebook users.

EO030 Room I1 STATISTICAL DISTRIBUTIONS IN OUR MODERN TIMES: ROLE MODELS OR NOT Chair: Christophe Ley

E0291: **Parametric assumptions in extreme value theory**
Presenter: **Jenny Wadsworth**, Lancaster University, United Kingdom
In univariate extreme value theory, parametric limiting distributions for the tail arise under weak conditions on the underlying distribution. There is no such luck for the multivariate or spatial case: although non-degenerate limit distributions often arise, and certain conditions must be satisfied, there is no finite-dimensional parameterization of all possible dependence structures. However, parametric assumptions are a common simplification and often necessary for tackling high-dimensional problems. We will review some of the different approaches to multivariate extremes and aim to discuss the relative merits of (non) parametric inference methods.

E0611: **Modelling multivariate skew and heavy-tailed data: A comparison of the main models**
Presenter: **Sladjana Babic**, Ghent University, Belgium
Co-authors: Christophe Ley, David Veredas
The most popular flexible classes of multivariate distributions are presented and their advantages and drawbacks are discussed. By flexible distribution we mean that, besides the usual location and scale parameters, the distribution has also both skewness and tail parameters. The following flexible families of multivariate distributions are presented: elliptical distributions, skew-elliptical distributions, multiple scaled mixtures of multinormal distributions, multivariate distributions based on the transformation approach, copula-based multivariate distributions and meta-elliptical

distributions. A theoretical comparison based on the properties of each model is done, while we conduct a Monte Carlo simulation study to check the fitting abilities of every model. To this end we generate data from every model and compare the competitor models in terms of their fitting qualities. This allows us to draw general conclusions concerning the flexibility of the various distributions.

E0663: **On the choice of size distributions for earthquakes modelling**
Presenter: **Rosaria Simone**, University of Naples Federico II, Italy
Co-authors: Christophe Ley
Power-law models are very popular in several environmental and geo-physical applications: in particular, empirical evidence with major implications stems from earthquake modelling. Departing from the Gutenberg-Richter law for magnitude and seismic moment, many research efforts are continuously addressed to the proposal of candidate size distributions for seismic events. Indeed, the accurate modelling of earthquake sizes is the first priority before a step forward in the joint modelling of time and space could be taken. A critical overview of the topic is provided by advancing a model able to fit also the temporal dimension of the phenomenon. For illustrative purposes, data concerning the Pacific Ring of Fire have been considered.

E0985: **Parametric hidden Markov fields for segmenting environmental spatial series with circular components**
Presenter: **Francesco Lagona**, University Roma Tre, Italy
Hidden Markov random fields are convenient tools for segmenting environmental spatial data according to a finite number of regimes that represent the conditional distributions of the data under specific environmental conditions. Under this setting, the data are modelled by a finite mixture of parametric densities, whose parameters vary across space according to a latent Markov random field. Motivated by environmental studies that require the segmentation of angular data, we describe two hidden Markov random fields for the analysis of a spatial series of angular measurements and, respectively, for the analysis of a cylindrical spatial series, i.e. a bivariate spatial series of directions and intensities. Both models are estimated by composite-likelihood methods, because of the numerical intractability of the likelihood function. These proposals are illustrated on two cases studies of wildfire seasonality and sea current circulation. In the first case, the model indicates the most likely places where fires could occur in specific periods of the year and captures the association between fire occurrences and land cover within each season of the year. In the second case, the model offers a clear-cut segmentation of sea current dynamics, which reflects the orography of the study area and captures regime-specific, non-linear relationships between the speed and the direction of the currents.

EO322 Room L1 APPROACHES FOR COMPLEXITY IN DATA ANALYSIS Chair: Efstathia Bura

E0584: **Global testing under the sparse alternatives for single index models**
Presenter: **Zhigen Zhao**, Temple University, United States
For the single index model $y = f(\beta^\tau x, \varepsilon)$ with Gaussian design, where f is unknown and β is a sparse p-dimensional unit vector with at most s nonzero entries, the aim is to test the null hypothesis that β, when viewed as a whole vector, is zero against the alternative that some entries of β is nonzero. Assuming that $var(E[x|y])$ is non-vanishing, we define the generalized signal-to-noise ratio (gSNR) λ of the model as the unique non-zero eigenvalue of $var(E[x|y])$. We show that if $s^2\log^2(p) \wedge p$ is of a smaller order of n, denoted as $s^2\log^2(p) \wedge p \prec n$, where n is the sample size, one can detect the existence of signals if and only if gSNR$\succ \frac{p^{1/2}}{n} \wedge \frac{s\log(p)}{n}$. Furthermore, if the noise is additive (i.e., $y = f(\beta^\tau x) + \varepsilon$), one can detect the existence of the signal if and only if gSNR$\succ \frac{p^{1/2}}{n} \wedge \frac{s\log(p)}{n} \wedge \frac{1}{\sqrt{n}}$. It is rather surprising that the detection boundary for the single index model with additive noise matches that for linear regression models. These results pave the road for thorough theoretical analysis of single and multiple index models in high dimensions.

E0917: **Novel model-free estimation approaches of linear dimension reduction**
Presenter: **Lukas Fertl**, TU Vienna, Austria
The purpose is to introduce a new way of estimating the dimension reduction matrix B in the dimension reduction model $y = f(B'x) + \varepsilon$, where B is a $p \times d$ ($d < p$) unknown matrix of parameters and ε is a random error independent of x. The idea is based on considering the variance of y conditional on x being in the span of a direction vector v as an estimating equation. This estimator falls in the class of semi-parametric methods, and we will denote it as conditional variance estimators. The performance of the estimator is competitive compared to currently used ones. Its main advantage is that it is more robust against a wide range of distributions for x and nonlinear $f()$. Extensions to other estimation or testing problems will be also presented. Furthermore, it can also be used when the size of the sample is smaller than the number of covariates ($n < p$).

E0943: **Real time dimension reduction and outlier detection**
Presenter: **Andreas Artemiou**, Cardiff University, United Kingdom
Co-authors: Yuexiao Dong, Seung Jun Shin
The aim is to discuss how a newly proposed algorithm for real time dimension reduction can be used for outlier detection. We first introduce the idea of SVM-based sufficient dimension reduction and then discuss how this can be extended to real time sufficient dimension reduction. Finally, we demonstrate that this algorithm can be used for efficient real time outlier detection. The good performance of the algorithm is demonstrated with simulated and real data.

E0962: **Sparse sufficient dimension reduction by nonconvex ADMM**
Presenter: **Bingyuan Liu**, Penn State University, United States
Co-authors: Amal Agarwal, Lingzhou Xue
Sufficient dimension reduction (SDR) is widely used for dimension reduction and feature extraction in high-dimensional data analysis. With a better interpretability, the sparse SDR provides an appealing alternative. However, the statistical consistency and efficient estimation for sparse SDR in high dimensional setting remains an open question. We first introduce the L_0-constrained inverse moment method and study its asymptotic properties (including convergence rate and feature selection consistency) under the high-dimensional setting where the dimension diverges as the sample size increases. Computationally, we propose the new nonconvex alternating directional method of multipliers (ADMM) to solve the nonconvex and nonsmooth optimization in sparse SDR. We study the computational guarantees of the folded concave penalized estimation to approximate the L_0 penalization and show an explicit iteration complexity bound for the proposed nonconvex ADMM to reach the stationary solution. We demonstrate the numerical properties of our proposed methods in both simulation studies and a real application.

EO340 Room M1 FUNCTIONAL DATA ANALYSIS AND BIOLOGICAL APPLICATIONS Chair: Marzia Cremona

E0902: **Brain structural connectivity mapping: Insights from functional data analysis**
Presenter: **Aymeric Stamm**, Human Technopole - IIT, Italy
Co-authors: Simone Vantini

Brain structural connectivity mapping pertains to reconstructing the axons that connect the different parts of our brain. This is done by tracking diffusion of water within axons using MRI and it is known as the process of tractography. The data provided by tractography consists in a set of curves defined on a three-dimensional domain which can take values in different spaces depending on which features we look at along their path. This information is critical in a number of neurological applications such as, for example, brain tumour removal surgery. Indeed, neurosurgeons must understand which tissue is still alive and which has been damaged by the tumour in order to remove as many tumoral cells as possible without affecting normal, possibly vital, functions. The availability of an atlas of structural connections within the healthy human brain would then be highly relevant as a benchmark to compare patients' brains against. We see tractography data as functional data and take advantage of cutting-edge statistical methods from the functional data analysis literature to provide an atlas of the healthy cortico-spinal tract, which regroups axons that connect the primary motor cortex to the spinal cord and therefore handle voluntary motion of all parts of our body.

E0674: **Human movement data - reliable enough for functional data analysis?**
Presenter: **Lina Schelin**, Umea University, Sweden
Co-authors: Alessia Pini

In movement laboratories, advanced measurement systems are used to capture human motion, forces that cause motion, and muscle activity patterns during motion. A common feature of such systems is that they generate functional data. Human movement has an inherent natural variation and we cannot expect observed movement curve data to be identical when a task is repeated. Still, it is crucial that measurement tools are valid and reliable, i.e., that they consistently measure the quantity that they are supposed to measure. Functional data analysis methods are already being used for the analysis of human movement data. However, reliability studies are mainly performed on reduced data, such as specific events or features extracted from the functional data. A few works on reliability for curve data have been proposed in the literature, but with limitations and no clear recommendations. We present and compare methods identified in the literature for reliability assessment of functional data, both on simulated data and for an application to knee kinematic data.

E0817: **Characterizing protein-DNA binding event subtypes in ChIP-exo data using read distribution shapes and DNA sequences**
Presenter: **Naomi Yamada**, Penn State University, United States
Co-authors: William Lai, Nina Farrell, Franklin Pugh, Shaun Mahony

Regulatory proteins associate with the genome either by directly binding cognate DNA motifs or via protein-protein interactions with other regulators. The ChIP-exo protocol precisely characterizes protein-DNA interactions by combining chromatin immunoprecipitation (ChIP) with 5 to 3 prime end exonuclease digestion. Since different regulatory complexes bind to DNA differently, analysis of ChIP-exo read distributions (curves generated by the read counts along the genome) should enable detection of multiple protein-DNA binding modes for a given regulatory protein. To systematically detect multiple protein-DNA interaction modes in a single ChIP-exo experiment, we introduce the ChIP-exo mixture model (ChExMix). ChExMix defines possible binding event subtypes by both clustering observed ChIP-exo read distribution shapes and performing targeted de novo motif discovery around the predicted binding events. ChExMix then uses an expectation maximization learning scheme to probabilistically model the genomic locations and subtype membership of binding events using both ChIP-exo read distributions and DNA sequence information. We demonstrate that ChExMix achieves accurate detection and classification of binding event subtypes using in silico mixed ChIP-exo data. We further demonstrate that ChExMix identifies cooperative binding interactions of key transcription factors in MCF-7 cells. Thus, ChExMix can effectively stratify ChIP-exo binding events into biologically meaningful subtypes.

E1045: **A novel approach to joint sparse functional clustering and alignment**
Presenter: **Valeria Vitelli**, University of Oslo, Norway

When performing functional clustering, the problem of selecting the portions of the domain which are most relevant to the classification purposes has already been considered. When misalignment is also present, the only possible approach is to first align the curves, and then use a sparse functional clustering method to estimate the groups and select the domain. However, it has been already proved that aligning and clustering the curves jointly is beneficial for the analysis. We thus propose a novel algorithm which jointly performs all these tasks: clustering, alignment, and domain selection. We prove the well-posedness of the problem, and test the method on simulated data. We also perform the analysis of the Berkeley Growth Study data, as a benchmark for functional data, and propose the use of the method for genomic data.

EO402 Room N1 MULTIVARIATE AND SPATIAL EXTREMES Chair: Marco Oesting

E0759: **Cluster-based extremal inference for multivariate time series**
Presenter: **Anja Janssen**, KTH Royal Institue of Technology, Sweden
Co-authors: Holger Drees

Statistical procedures for inference on extremal properties of a multivariate time series are affected by the underlying extremal dependence structures. Many common time series models exhibit a clustering of extreme values and this will typically affect the variance of estimators which were built for i.i.d. observations. On the other hand, the behavior of quantities of interest, for example marginal distributions of the spectral tail process, is closely related to the overall dependence structure which we see in extremal clusters. We explore how this connection can be exploited to derive new estimators for extremal quantities.

E1238: **Hierarchical space-time modeling of exceedances**
Presenter: **Gwladys Toulemonde**, Universita de Montpellier, France
Co-authors: Jean-Noel Bacro, Carlo Gaetan, Thomas Opitz

The statistical modeling of space-time extremes in environmental applications is a valuable approach to understand complex dependences in observed data and to generate realistic scenarios for impact models. Motivated by hourly rainfall data in Southern France presenting asymptotic independence, we propose a novel hierarchical model for high threshold exceedances leading to asymptotic independence in space and time. The approach is based on representing a generalized Pareto distribution as a Gamma mixture of an exponential distribution, enabling us to keep marginal distributions which are coherent with univariate extreme value theory. The key idea is to use a kernel convolution of a space-time Gamma random process based on influence zones defined as cylinders with an ellipsoidal basis to generate anisotropic spatio-temporal dependence in exceedances. Statistical inference is based on a composite likelihood for the observed censored excesses. The practical usefulness of our model is illustrated on the previously mentioned hourly precipitation data set from a region in Southern France.

E0791: **Understanding and communicating widespread flood risk**
Presenter: **Ross Towe**, Lancaster University, United Kingdom
Co-authors: Jonathan Tawn, Rob Lamb
During the winter of 2015/2016, the UK was hit by a sequence of storms that resulted in widespread flooding. As a result of these storms, the UK government formed the National Flood Resilience Review (NFRR) to better understand the drivers of flooding. Some of the questions raised by the NFRR included: what is the chance of an extreme river flow occurring at one or more gauges, somewhere within the national river gauge network in any one year? In order to address questions of this nature it is vital to understand the dependence of large values of river flow and determine which combinations of locations are likely to simultaneously observe high flows. We use a multivariate extreme value model to summarise the changing behaviour of floods across the UK. In order to account for the likelihood of missing data at a number of river flow gauges, an adaption to the current methodology is proposed. A constraint of the current methodology is that events are restricted to the locations where measurements of river flow are made. A number of algorithms can be used to interpolate river flow to ungauged parts of the river network, however we propose to also use information about rainfall.

E1086: **Semi-parametric estimation for max-mixture spatial processes**
Presenter: **Pierre Ribereau**, Universita Lyon 1, France
A semi-parametric estimation procedure is proposed in order to estimate the parameters of a max-mixture model as an alternative to composite likelihood estimation. This procedure uses the F-madogram. We propose to minimize the square difference between the theoretical F-madogram and an empirical one. We evaluate the performance of this estimator through a simulation study and we compare our method to composite likelihood estimation. We apply our estimation procedure to daily rainfall data from East Australia.

EO274 Room P1 ADVANCES IN STATISTICAL ANALYSIS OF MICROBIOME DATA Chair: Li Ma

E0312: **Testing statistical interactions between microbiome community profiles and covariates**
Presenter: **Michael Wu**, Fred Hutchinson Cancer Research Center, United States
Microbiome profiling studies are being conducted to find associations between bacterial taxa and a wide range of different outcomes. However, the dimensionality, compositionality, inherent biological structure, and limited availability of samples pose significant challenges. Community level analysis, wherein the entire profile is assessed for association with outcomes, can resolve some of these difficulties but does not easily generalize to analyzing effect modification due to bias incurred in estimating main effects. Thus, under the semi-parametric kernel machine testing framework, we propose a new framework for interaction testing at the community level that incorporates bias reduction approaches in estimating main effects while flexibly capturing interaction terms. Simulations and real data analyses show that our approach correctly controls type I error while maintaining power under a range of situations.

E0334: **Conditional regression based on a multivariate zero-inflated logistic model for human microbiome data**
Presenter: **Zhigang Li**, Department of Biostatistics, University of Florida, United States
Co-authors: James OMalley, Hongzhe Li
Massive high dimensional human microbiome data is commonly seen in molecular epidemiology research and have substantially increased in complexity to address critical health concerns due to complex data structure. Analysis challenges arise from compositional, phylogenetically hierarchical, sparse and high dimensional structure of microbiome data. Compositional structure could induce spurious relationships due to the linear dependence between compositional components. In addition, the hierarchical structure of microbiome data from the phylogenetic tree generates dependence at the hierarchical levels which poses a further modeling challenge. Furthermore, the sparsity of microbiome data due to excessive zero sequencing reads for microbial taxa remains an unresolved issue in the literature. Coupled with the high dimensional feature, microbiome data raises great challenging problems in the field of mediation data analysis. We will develop a zero-inflated logistic normal model to address these issues. A simulation study will show the performance of the approach and a real study example will be included as well.

E1097: **Clustering for microbiome data with structural zeros**
Presenter: **Julia Fukuyama**, Indiana University, United States
Species abundance tables from microbiome studies are famously sparse, with the zeros coming from a combination of stochastic and structural (true) zeros. Both sources of zeros lead to problems in parameter estimation, but they have different sources and need to be modeled differently. In addition, the structural zeros are often strongly associated with covariates, e.g., the identity of the host in human microbiome data. We describe a new method for clustering species in the presence of structural and stochastic zeros. In addition to reducing the dimensionality of the data, this clustering improves interpretability by identifying groups of bacterial species that perform the same functions. The clusters and the corresponding model for structural zeros can be used to examine ecological theories of microbial community assembly and maintenance. We show results on human data and discuss the ecological implications.

E0687: **Approximate message passing algorithms for de novo reconstruction in metagenomics**
Presenter: **Sergio Bacallado**, Cambridge University, United States
Microbiome studies sequence the DNA in samples containing a mixture of bacterial genomes. DNA sequences must be assigned to different taxa, and when not all taxa have been cultured and characterised previously this problem is known as de novo reconstruction. Mathematically, de novo reconstruction is a deconvolution problem which reduces to a matrix factorisation with highly structured factors. As it is relatively easy to formulate a prior and a probability model, Bayesian approaches to this problem have been proposed, but their computational cost can be high. We present an alternative algorithm based on approximate message passing which is evaluated by simulation.

EO160 Room Q1 DOUBLY STOCHASTIC COUNTING PROCESSES Chair: Paula Bouzas

E0535: **Statistical methods for replicated spatio-temporal point processes**
Presenter: **Daniel Gervini**, University of Wisconsin-Milwaukee, United States
Spatio-temporal point processes are widely used in statistics. However, the literature has mostly focused on the single-realization scenario. When many replications of a temporal point process are available at various spatial points, inferential tools such as kriging can be simplified and questionable assumptions such as isotropy are not necessary. We will introduce these new methods, based on doubly-stochastic Poisson process models, and show their application in the analysis of spatial and temporal bike demand in the Divvy shared-bicycle system of the city of Chicago.

E1165: **Nonparametric tests for Cox processes**
Presenter: **Lionel Truquet**, ENSAI, France
In a functional setting, we elaborate and study two test statistics to highlight the Poisson nature of a Cox process when n copies of the process are available. The approach involves a comparison of the empirical mean and the empirical variance of the functional data and can be seen as an extended version of a classical overdispersion test for count data. The limiting distributions of our statistics are derived using a functional central limit theorem for cadlag martingales. The procedures are easily implementable and do not require any knowledge on the covariate. We address a

theoretical comparison of the asymptotic power of our tests under some local alternatives. A numerical study reveals the good performances of the method. We also present two applications of our tests to real data sets.

E0835: **Goodness-of-fit test for compound Cox process**
Presenter: **Paula Bouzas**, University of Granada, Spain
Co-authors: Nuria Ruiz-Fuentes
A goodness-of-fit test is proposed for a compound Cox process when it is observed in a discrete set of time points; moreover, the available information can be collected as recurrent event data or panel count data. The test determines if a new sample path fits a given model whether it is already known or it has to be estimated previously. In the latter case, functional data analysis is used throughout the procedure. The assessment of the goodness-of-fit test is carried out by means of its application to several simulation cases. In each case, a number of new sample paths with or without provoked perturbations are tested to prove the accuracy of the proposed hypothesis test.

E0940: **Applications of goodness-of-fit test for compound Cox processes**
Presenter: **Nuria Ruiz-Fuentes**, University of Jaen, Spain
Co-authors: Paula Bouzas
The number of extreme values or the turning points of a stochastic process, as different as they may be, can be modeled by compound Cox processes. Having estimated the model by means of functional data analysis and having observed a new sample path, a goodness-of-fit test determines if it follows the same model. The conclusion can be used to answer a number of important questions that arise with real data. Some examples illustrate the application of the test. For example, dealing with extreme meteorological events, the application of the test helps to distinguish between climate zones or whether a certain period of time follows the usual pattern. Studying the turning points in the stock prices, the test assesses a possible similar behaviour of several stocks or markets.

EO446 Room D2 RECENT DEVELOPMENT OF THE DESIGN OF EXPERIMENTS AND INDUSTRIAL STATISTICS Chair: Chang-Yun Lin

E0552: **Characterizations of indicator functions for fractional factorial designs**
Presenter: **Satoshi Aoki**, Kobe University, Japan
A polynomial indicator function of designs is a basic tool to characterize fractional factorial designs in the field of computational algebraic statistics. For the case of two-level designs, the structure of the indicator function is well-known. For example, the coefficients of indicator functions have clear meanings relating to the concept of the aberration and resolution for the two-level cases. The polynomial relation among the coefficients is also derived for the two-level cases, which can be used to classify designs with given sizes. However, for the cases of multi-level designs with rational factors, such relations are complicated and interpretations are difficult. We consider the structure of the indicator function of general designs and its applications.

E1677: **Supersaturated multistratum designs**
Presenter: **Chang-Yun Lin**, National Chung Hsing University, Taiwan
Supersaturated designs have gained much attention in the past two decades. Such designs can reduce substantial cost in the initial stage of experiments to identify the few important factors from many of interest. Most exiting supersaturated designs in the literature are developed for completely randomized experiments, which have single- stratum structures. However, in many cases, complete randomization is infeasible, and hence designs with more complicated structures are needed. The supersaturated designs in multistratum structures are studied. We propose the generalized Df criterion and develop the multistratum columnwise-pairwise exchange algorithm for constructing and selecting efficient supersaturated multistratum designs. The supersaturated split-plot, strip-plot, and staggered-level designs are constructed and their sensitivity, power, and type I error rate are studied. An example with SAS and R codes is provided to demonstrate how to conduct data analysis for supersaturated multistratum designs.

E0844: **Design and analysis of covering arrays using prior information**
Presenter: **Ryan Lekivetz**, SAS Institute Inc., United States
Validating complex engineered systems is proving to be an increasingly difficult task for test engineers given the tight budgetary constraints that they usually face. The behavior of such systems is typically due to many inputs, each of which may have several possible settings. Consequently, the input space for these systems is often so large that only a small fraction of the input space can be used for testing. The challenge for the test engineer is to find a way to sample from the input space in such a way that testing effectiveness is maximized. As it turns out, covering arrays can be used to address this challenge. We introduce the concept of covering arrays and illustrate how they may be used to validate engineered systems, and how prior knowledge of the system can be used in both the design and analysis of a covering array.

E0216: **On the design of experiments with ordered treatments**
Presenter: **Ori Davidov**, University of Haifa, Israel
There are many situations where one expects an ordering among $K \geq 2$ experimental groups or treatments. Although there is a large body of literature dealing with the analysis under order restrictions, surprisingly, very little work has been done in the context of the design of experiments. We provide key observations and fundamental ideas which can be used as a guide for designing experiments with ordered treatments. In particular, we focus on designs that are optimal for testing hypotheses. The theoretical findings are with supplemented thorough numerical illustrations.

EO438 Room P2 BAYESIAN INFERENCE AND DECISION Chair: Eva Lopez Sanjuan

E1252: **The multi armed bandit problem under delayed rewards conditions in digital campaign management**
Presenter: **Miguel Martin-Blanco**, Universidad Politecnica de Madrid, Spain
Co-authors: Antonio Jimenez-Martin, Alfonso Mateos Caballero
The most representative allocation strategies to deal with the multi-armed bandit problem are analyzed in a context with delayed rewards by means of a numerical study based on a discrete event simulation. The scenario that we address is a digital marketing content recommendation system, called campaign management, used by marketers to create specific digital content that can be issued or configured for viewing by certain population segments according to a series of business variables, user profile or behavior. Both batch mode and online update architectures are considered for feedback from the different contents displayed to users. The results show that possibilistic reward (PR) methods outperform other allocation strategies in this scenario with delayed rewards.

E1253: **An improved prior choice for the parameters in the generalized Pareto distribution**
Presenter: **Eva Lopez Sanjuan**, Universidad de Extremadura, Spain
Co-authors: Mario Martinez Pizarro, M Isabel Parra Arevalo, Jacinto Martin Jimenez
In the parameter estimation of limit extreme value distributions, standard methods only use some of the available data. For the generalized Pareto distribution, only the observations above a certain threshold are considered, therefore a big amount of information is wasted. The aim is making the most of the information provided by the observations, in order to improve the accuracy in Bayesian parameter estimation. The strategy consists in taking advantage of the existing relationship between the parameters of baseline and generalized Pareto distributions to obtain informative prior

distributions. Different simulations have been carried out in order to compare the effectiveness of the proposed method to the standard ones. Specifically, simulations for different baseline distributions were studied: normal, exponential and Cauchy distributions, because of the different behavior of the tails. The proposed method can be extended to other baseline distributions.

E1257: **An improved prior choice for Gumbel distribution parameters**
Presenter: **M Isabel Parra Arevalo**, Universidad de Extremadura, Spain
Co-authors: Francisco Javier Acero Diaz, Ruben Gomez Gonzalez, Jacinto Martin Jimenez
The methods for parameter estimation of the extreme-value distributions use only a few observations. When the focus is on modeling the extreme data based on block maxima approach using Gumbel distribution, only one observation from each block is used. A strategy that allows us to take advantage of the information from all available observations is proposed, pursuing the objective of increasing the accuracy of Bayesian parameters estimation. It consists on harnessing the existing relationship between the parameters of baseline and Gumbel distributions to obtain informative prior distributions. Our method shows good performance when dealing with very shortened available data. Different statistical analysis tests are used to compare the performance and the standard algorithm. The empirical effectiveness of the approach is demonstrated through a simulation study and a case study. Reduction in the credible interval width and enhancement in parameter location show that approach based on highly informative prior adapt to very shortened data better than the standard method does.

E1682: **Stochastic decision-making using particle methods**
Presenter: **Maciej Marowka**, Imperial College London, United Kingdom
Co-authors: Nikolas Kantas
A novel numerical, particle based method is proposed to estimate the optimal control inputs for a risk sensitive stochastic decision-making problem where a multiplicative reward is used. The problem is to identify a control sequence such that the resulting observations from the non-linear non Gaussian state-space model match a required deterministic reference sequence with respect to the particular choice of reward. The approach is based on earlier efforts for deterministic systems and is essentially a sequential Monte Carlo (SMC) algorithm for an appropriate dual filtering problem. Extensions using SMC^2 for nonlinear models based on hierarchical (deep) state space models are developed. We will illustrate the performance of the method on particular synthetic data examples and discuss possible applications to the optimal trading problem in the environment with time varying cointegration model, where the cointegration space bases are driven by a latent stochastic process.

EO665 Room Q2 BAYESIAN MODELING FOR HETEROGENEOUS GROUPS Chair: Feng Liang

E0371: **Anchored Bayesian Gaussian mixture models**
Presenter: **Mario Peruggia**, The Ohio State University, United States
Co-authors: Deborah Kunkel
A novel approach to the specification of Bayesian Gaussian mixture models is described which eliminates the label switching problem. Label switching refers to the invariance of the posterior distribution for the component-specific parameters to relabeling of the components when an exchangeable prior is used. There are two common approaches to address this issue. The first breaks the exchangeability assumption by imposing artificial constraints on some model parameters (or specifies some other informative prior). The second approach relabels the MCMC samples generated to estimate the exchangeable model in a way that favors one specific relabeling of the components. Our approach forces few observations, which we call the anchor points, to arise from prespecified components of the mixture. Specifying the anchor points is tantamount to specifying an informative, data-dependent prior, in which some observations are assumed to arise from a given component with probability one. We show that a careful choice of the anchor points can yield marginal posterior distributions for the component-specific parameters that are well separated and interpretable.

E0580: **Bayesian prediction with heterogeneous populations: An application to feature sampling**
Presenter: **Federico Camerlenghi**, University of Milano-Bicocca and Collegio Carlo Alberto, Italy
The prediction of future outcomes of a random phenomenon is typically based on a certain number of analogous observations from the past. When observations come from multiple and heterogeneous populations, a natural notion of analogy is partial exchangeability and the problem of prediction can be effectively addressed in a Bayesian nonparametric setting. We define and investigate new classes of hierarchical processes which are useful for prediction in species sampling and feature models. We concentrate our attention on feature models, which generalize species sampling models by allowing every observation to belong to more than one species, now called features. In this setting we are able to forecast the outcome of additional samples having arbitrary size and to derive distributional properties for many statistics of interest, such as the number of hitherto unseen features that will be observed in an additional sample.

E0890: **Hierarchical species sampling models**
Presenter: **Federico Bassetti**, Univeristy of Pavia, Italy
Co-authors: Roberto Casarin, Luca Rossini
A general class of hierarchical nonparametric prior distributions is introduced. The random probability measures are constructed by a hierarchy of generalized species sampling processes with possibly non-diffuse base measures. The proposed framework provides a general probabilistic foundation for hierarchical random measures with either atomic or mixed base measures and allows for studying their properties, such as the distribution of the marginal and total number of clusters. We show that hierarchical species sampling models have a Chinese restaurants franchise representation and can be used as prior distributions to undertake Bayesian nonparametric inference. We provide a method to sample from the posterior distribution together with some numerical illustrations. Our class of priors includes some new hierarchical mixture priors such as the hierarchical Gnedin measures, and other well-known prior distributions such as the hierarchical Pitman-Yor and the hierarchical normalized random measures.

E1056: **Assessing causal effects in the presence of treatment switching through principal stratification**
Presenter: **Alessandra Mattei**, University of Florence, Italy
Co-authors: Fabrizia Mealli, Peng Ding
Clinical trials, focusing on survival outcomes for patients suffering from AIDS-related illnesses and painful cancers in advanced stages, often allows patients in the control arm to switch to the treatment arm if their physical conditions get worse than certain tolerance levels. The Intention-To-Treat analysis, often used in practice, provides valid causal estimates of the effect of assignment, but it does not give information about the effect of the actual receipt of the treatment and ignores the information on treatment switching. Other existing methods propose to reconstruct the outcome a unit would have had if s/he had not switched. But these methods usually rely on strong assumptions, like that there exists no relation between patients prognosis and switching behavior, or the treatment effect is constant. We propose to re-define the problem of treatment switching using principal stratification, and we focus on principal causal effects for patients belonging to subpopulations defined by the switching behavior under control. Our approach appropriately adjusts for the post-treatment information and characterizes treatment effect heterogeneity. For inference, we use a Bayesian approach to properly take into account that (i) switching happens in continuous time; (ii) switching time is not defined for units who never switch in a particular experiment; and (iii) survival time and switching time are subject to censoring. We illustrate our framework using

simulated data.

EG257 Room C1 CONTRIBUTIONS IN APPLIED STATISTICS I **Chair: Andreas Mayr**

E1218: **Double generally weighted moving average chart for time between events**
Presenter: **Schalk Human**, University of Pretoria, South Africa
Co-authors: Janet Van Niekerk, Hossein Masoumi Karakani
Control charts continue to play a key role in the quality control (QC) environment. However, the Shewhart-type attributes charts are inefficient at detecting small/minor changes. To overcome this shortcoming, an alternative approach is to use time-weighted control charts (also known as memory-based control charts) to monitor the time between events (TBE); these time-weighted control charts use all the information from the start until the most recent sample to decide if a process is in-control (IC) or out-of-control (OOC). To this end, a generalized type of time-weighted control chart is proposed to monitor the TBE. This chart is called the Double Generally Weighted Moving Average Time Between Events (DGWMA-TBE), which includes many of the well-known existing time-weighted control charts as special or limiting cases. Evaluation of the run-length distribution reveals that the proposed DGWMA-TBE chart outperforms the Generally Weighted Moving Average (GWMA), Exponentially Weighted Moving Average (EWMA) and Shewhart charts at detecting small to moderate shifts.

E1659: **Machine learning for predicting default of credit card holders and success of kickstarters**
Presenter: **Huei-Wen Teng**, National Chiao Tung University, Taiwan
Applications of machine learning in finance have got extensive attention in recent years. We demonstrate the flexibility of machine learning through two examples: predicting default of credit card clients and forecasting success of kickstarter projects, by using K-nearest neighbours, decision trees, boosting, support vector machine, and neural networks. Neural networks enable constructing complicated functions to map input features to an output response, but their implementation requires accurate and efficient inference to the associated weights. To compare the numerical efficiency in inferring the weights of neural networks, we find that back propagation outperforms randomized hill climbing, simulated annealing, and genetic algorithm, in terms of accuracy and computation time. Finally, we conduct principal component analysis and independent component analysis for dimensionality reduction for the neural network.

E1324: **Looking for gender bias**
Presenter: **Javier de Vicente Maldonado**, Carlos III University, Spain
Measuring the possible bias in the allocation of intrahousehold resources is a difficult exercise, mainly due to the particular nature of the household expenditure data. However, an accurate assessment could lead to the identification of gender discrimination which, at the same time, would be a key factor for women empowerment and economic development. We represent the latent Engel curves, i.e. the substantial drivers of consumption patterns, as an approximate factor model in which the factors are extracted using an algorithm based on the maximum likelihood method. In addition, we propose a new measure of gender bias called latent outlay equivalence ratio based on the original procedure proposed previously. Finally, we illustrate this new approach using the commonly used data from the 1889/90 US Bureau of Labor report, which consists of 1024 budgets of British families in the textile, coal-mining and metal manufacturing industries. We clearly identify the existence of two latent curves that can be defined as basic necessities (e.g. food) and luxuries (e.g. alcohol) respectively. In contrast to the results obtained in previous analysis, we find strong evidence of gender discrimination.

E1189: **Insurance ratemaking using the exponential-lognormal regression model**
Presenter: **Woo Hee Yik**, London school of economics and political science, United Kingdom
Co-authors: George Tzougas, Muhammad Waqar Mustaqeem
The exponential-lognormal regression model is introduced as an alternative to the Pareto regression model that has been widely used for modelling the cost of claims as a function of their risk characteristics in an abundance of alternative insurance applications. The exponential-lognormal regression model can be considered as a plausible model for approximating moderate claim costs which are more frequent than large claim sizes when dealing with real insurance data sets. However, this is the first time that it is used in a statistical or an actuarial context because its log-likelihood is complicated, and hence its maximization needs a special effort. The main contribution is to illustrate that ML estimation of the exponential-lognormal regression model can be accomplished relatively easily via an expectation maximization (EM) algorithm which can address situations where the mixing distribution, such as the lognormal, is not conjugate to the exponential distribution. A real data application based on motor insurance data is examined in order to illustrate the versatility of the proposed algorithm. Finally, assuming that the number of claims follows the negative binomial model, both the a priori and a posteriori premium rates resulting from the exponential-lognormal model for approximating claim sizes are calculated via the net premium principle and compared to those determined by the Pareto model, that has been traditionally used for modelling losses.

CO238 Room A2 FORECASTING AND TIME SERIES **Chair: Robert Kunst**

C0248: **Exploring the predictive ability of LIKES of posts on the Facebook pages of four major city DMOs in Austria**
Presenter: **Ulrich Gunter**, MODUL University Vienna, Austria
Co-authors: Irem Onder, Stefan Gindl
Using data for the period 2010M06 - 2017M02, the aim is to investigate the possibility of predicting total tourist arrivals to four Austrian cities (Graz, Innsbruck, Salzburg, and Vienna) from LIKES of posts on these destinations DMO Facebook pages. Google Trends data are also incorporated in investigating whether forecast models with LIKES and/or with Google Trends deliver more accurate forecasts. To capture the dynamics in the data, the ADL model class is employed. Taking into account the daily frequency of the original LIKES, the MIDAS model class is also employed. While time-series benchmarks from the naive, ETS, and ARMA model classes perform best for Graz and Innsbruck across horizons and accuracy measures, ADL models incorporating only LIKES or both LIKES and Google Trends generally outperform their competitors for Salzburg. For Vienna, the MIDAS model including both LIKES and Google Trends produces the smallest RMSE, MAE, and MAPE values for most horizons. Therefore, for at least two of the four Austrian cities under scrutiny, incorporating complementary information originating from two different web-based predictors is worthwhile in order to produce more accurate tourism demand forecasts. In addition, forecast encompassing tests relative to the naive-1 benchmark reject their null hypothesis at least at the 10% significance level in 18 out of 24 cases across cities and horizons, thus making the use of more sophisticated forecast models meaningful in the first place.

C0763: **DGSE models with expectations correction: Misspecification, forecasting errors and directional accuracy**
Presenter: **Mauro Costantini**, Brunel University, United Kingdom
Co-authors: Giovanni Angelini
The forecasting performance of small-scale New-Keynesian models with expectations correction is considered. These models are designed to reduce misspecification. A comprehensive comparative forecasting evaluation of different specifications through a Monte Carlo analysis is offered, so to establish whether DSGE models with expectations correction perform better than other macro structural models. Further, an empirical application, based on frequentist estimation of DSGE models with expectations correction for the US data, is provided.

C0789: **Simulation-based selection of prediction models in development-economics panels**
Presenter: **Robert Kunst**, Institute for Advanced Studies, Austria
Co-authors: Adusei Jumah
Basing model selection decisions in a forecasting context on simulations from estimated rival structures can be an attractive albeit time-consuming tool in macro-economic forecasting. The simulations fuse data information and the structure hypothesized by tentative rival models. We explicitly focus on applying the idea to a panel of macro-economic data from African countries. The declared target is the optimization of predictions for economic growth in individual countries, based on their growth history and on sector shares. Data on sector shares are conveniently available for most African countries over extended time spans and are representative of the ongoing process of structural transformation. Our procedure chooses among few tentative forecast models in the presence of data, in this example univariate and multivariate models with a varying degree of panel homogeneity assumptions. From models fitted to the data, pseudo-data are generated. Again, the models are applied to the pseudo-data and their out-of-sample performance is evaluated. The ultimate choice of the forecasting model is based on the relative performance of rival models in predicting their own data and those of the rival model.

C1087: **Two are better than one: Volatility forecasting using multiplicative component GARCH models**
Presenter: **Onno Kleen**, Heidelberg University, Germany
Co-authors: Christian Conrad
The forecast performance of multiplicative volatility models that can be decomposed into a short- and a long-term component is examined. First, we show that in multiplicative models, returns have a higher kurtosis and squared returns have a more persistent autocorrelation function than in the nested GARCH model. Second, we provide theoretical and simulation evidence suggesting that the QLIKE loss should be preferred relative to the squared error loss when comparing volatility forecasts. In a Monte-Carlo simulation, we investigate how the multiplicative structure affects forecast performance both in comparison to the nested GARCH model and the popular HAR model. Finally, we consider an application to S&P 500 returns. Based on the QLIKE loss and forecast horizons of two- to three-months ahead, our results show that multiplicative GARCH models incorporating financial and macroeconomic variables improve upon the HAR model.

CO673 Room B2 RECENT ADVANCES IN ECONOMETRICS Chair: Valentina Corradi

C0858: **Quantile co-movement in financial markets**
Presenter: **Tomohiro Ando**, Melbourne Business School, Australia
Co-authors: Jushan Bai
A new procedure is introduced for analyzing the quantile co-movement of a large number of financial time series based on a large-scale panel data model with factor structures. The proposed method attempts to capture the unobservable heterogeneity of each of the financial time series based on sensitivity to explanatory variables and to the unobservable factor structure. In our model, the dimension of the common factor structure varies across quantiles, and the factor structure is allowed to be correlated with the explanatory variables. The proposed method allows for both cross-sectional and serial dependence, and heteroskedasticity, which are common in financial markets. We propose new estimation procedures for both frequentist and Bayesian frameworks. Consistency and asymptotic normality of the proposed estimator are established. We also propose a new model selection criterion for determining the number of common factors together with theoretical support. We apply the method to analyze the returns for over 6,000 international stocks from over 60 countries during the subprime crisis, European sovereign debt crisis, and subsequent period. The empirical analysis indicates that the common factor structure varies across quantiles. We find that the common factors for the quantiles and the common factors for the mean are different.

C0952: **Geometrically stopped Markovian random growth processes and Pareto tails**
Presenter: **Brendan Beare**, University of California, San Diego, United States
Co-authors: Alexis Akira Toda
Many empirical studies document power law behavior in size distributions of economic interest such as cities, firms, income, and wealth. One mechanism for generating such behavior combines independent and identically distributed Gaussian additive shocks to log-size with a geometric age distribution. We generalize this mechanism by allowing the shocks to be non-Gaussian (but light-tailed) and dependent upon a Markov state variable. Our main results provide sharp bounds on tail probabilities and simple formulas for Pareto exponents. We present two applications: (i) we show that the tails of the wealth distribution in a heterogeneous-agent dynamic general equilibrium model with idiosyncratic endowment risk decay exponentially, unlike models with investment risk where the tails may be Paretian, and (ii) we show that a random growth model for the population dynamics of Japanese prefectures is consistent with the observed Pareto exponent but only after allowing for Markovian dynamics.

C1014: **Portfolio sorting error**
Presenter: **Valentina Corradi**, University of Surrey, United Kingdom
Co-authors: Walter Distaso
Portfolio sorts are commonly used in finance to unveil the relation between returns and the sorting variable which groups stocks. Sorting can be according to an observable variable, such as size or book to market, or to an unobservable variable, such volatility, (co)-skewness. Sorting according to unobservable variable is infeasible. Hence, we need to replace the sorting variable with an estimated counterpart. In addition to estimation error, due to the fact that sorting is done according to order statistics rather than "true" deciles, we need to take into account missclassification. Returns assigned to the bottom decile may instead belong to a higher decile. The objective is to introduce sorting procedures robust to misclassification error. This is accomplished by establishing a rule for eliminating the highest and lowest returns within each deciles. If there is no significant relation between returns and sorting variables, missclassification error does not matter. On the other hand, if there is a significant positive or negative relation then it does matter. Often one wants to test the null that the mean return on top and bottom portfolios is the same, versus the alternative that is smaller (larger). Hence, we want to trim away a sufficient number of observations to affect the outcome of the test under the alternative, but not to affect under the null. In the empirical illustration, we study sorting error for portfolios sorted according to (co)-skewness.

C1205: **Expected jumps and the cross-section of equity returns**
Presenter: **Walter Distaso**, Imperial College London, United Kingdom
Co-authors: Massimiliano Caporin, Nancy Zambon
How individual equity prices react to stock specific expected jump components is investigated. We find that a portfolio buying stocks with negative expected jump component and selling stocks with positive expected jump component earns significant returns, equal to 51 basis points per month. The returns of the spread portfolio cannot be explained by traditional risk factors. Furthermore, the associated risk premium is positive, very close to the average monthly return, and remain significant after controlling for portfolio characteristics.

CO436 Room C2 ECONOMIC VALUE OF VARIANCE RISK Chair: Romeo Tedongap

C0359: **Variance risk: A bird's eye view**
Presenter: **Chardin Wese**, ICMA Centre, University of Reading, United Kingdom
Co-authors: Fabian Hollstein
Prior research using daily returns data documents a significantly negative variance risk premium (VRP) for the S&P 500 index but generally not for constituent stocks. Using high-frequency data to compute the realized variance estimates, we show that the average VRP of individual equities is economically large and significant. We decompose the index VRP into factors related to the VRP of equities and the correlation risk premium. The former drives the variations in the index VRP while the latter captures the level of the index VRP. These factors predict excess stock returns in the time-series and cross-section, but at different horizons.

C0906: **Short-term predictability and the cross-section of stock returns**
Presenter: **Bastien Buchwalter**, ESSEC Business School, France
Co-authors: Romeo Tedongap, Johannes Breckenfelder
A novel decomposition of the realized variance into four components is developed: downside tail, downside core, upside core and upside tail realized variance. We show that these measures of market uncertainty are able to predict the excess market return on a monthly basis. We further show that the realized tail variance and realized tail asymmetry, which are defined as the sum and the difference of the upside and downside tail realized variance, respectively, are priced in the cross-section of stock returns. Our findings suggest that uncertainty averse investors demand extra compensation in form of a higher return to hold the stocks that positively relate to (downside) tail market uncertainty.

C1396: **Macro uncertainty and the term structure of the risk premium**
Presenter: **Jules Tinang**, University of Groningen, Netherlands
Leading frictionless consumption-based asset pricing models (long-run risks and Habit formation) predict that the expected return on assets whose cash flows appear in the distant future are higher than or equal to the expected returns on assets which pay-off in the near future. Contrary to that prediction, some recent empirical studies have found that short-term assets earn a higher expected return than long-term assets. We show that allowing the cash flows to be negatively affected by volatility shocks, as observed in the data (leverage effect), could make the short-term assets riskier than long-term assets. This modification gives more flexibility to those models in capturing various shapes of the term structure of equity returns while still matching the observed level of the equity premium and the risk free rate.

C1703: **Pricing of idiosyncratic equity and variance risks**
Presenter: **Elise Gourier**, ESSEC Business School, France
The risk premia of individual stocks is decomposed into contributions from systematic and idiosyncratic risks. We introduce an affine jump-diffusion model, which accounts for both the factor structure of asset returns and that of the variance of idiosyncratic returns. The estimation is performed on a time series of returns and option prices from 2006 to 2012. We find that investors not only require compensation for the systematic movements in returns and variance, but also for non hedgeable idiosyncratic risks. For the stocks of the Dow Jones, these risks account for an average of 50% and 80% of the equity and variance risk premia, respectively. We provide a categorization of sectors based on the risk profile of their exchange traded funds and highlight the high prices of idiosyncratic risks in the energy, financial and consumer discretionary sectors. Other sectors are found to be appealing alternatives for investors who are not willing to be exposed to non diversifiable risks.

CO458 Room E2 QUANTITATIVE INVESTMENT MANAGEMENT Chair: Gaelle Le Fol

C0458: **Abnormal tone and abnormal returns: An event study analysis**
Presenter: **David Ardia**, University of Neuchatel, Switzerland
Co-authors: Keven Bluteau, Kris Boudt
The Cumulative Abnormal Tone (CAT) framework is proposed for tracking the normal and abnormal dynamics of textual tone around events. In this framework, we use text-mining techniques to derive the normal tone based on market and sector-wide news. We then focus on corporate events and track the abnormal tone dynamics around that event. This leads to a cumulative abnormal tone chart. We apply the CAT framework to the analysis of cumulative abnormal returns across earnings press releases. We find that the analysis of firm's abnormal tone provides investors with relevant predictive information on the firms stock performance.

C0475: **Machine learning models applied in trading and their potential issues**
Presenter: **Rafael Molinero**, Molinero Capital Management, United States
While machine learning models display interesting characteristics and allow the modeling of complex non-linear relationships, which could be useful in finance & trading, they also suffer from various issues such as curve fitting, stability issues, as well as low interpretability. We will go over an application of AI models, such as neural networks, for market price prediction and its potential trading application. We will then review the results and issues behind these models.

C0487: **Illiquid asset and portfolio management**
Presenter: **Gaelle Le Fol**, Universite Paris - Dauphine, and CREST, France
Co-authors: Christian Brownlees, Serge Darolles, Beatrice Sagna
The aim is to promote a better analysis of financial risk focused on one specific dimension: liquidity risk. We introduce a multivariate model to analyze the liquidity structure of a large panel of assets. These could be liquidity measures of different assets on one trading venue as well as liquidity measures of the same assets over different venues. The highlight of the modeling approach is that it disentangles the dynamics of liquidity in a systematic market wide liquidity factor component as well as individual network liquidity spillover effects among the series in the panel. The model allows to study the joint dynamics of the liquidity measures in the panel as well as predicting the future degree of liquidity. Some applications of our methodology include the design of basket VWAP (Volume Weighted Average Price) trading strategies, aiming at lowering the market impact, and the optimal management of collateral.

C0540: **Community detection in large vector autoregressions**
Presenter: **Gudmundur Gudmundsson**, Aarhus University, Denmark
A class of vector autoregressive (VAR) models are introduced in which the time series are partitioned into unknown communities. Spillovers are stronger between series that belong to the same community than otherwise. A natural question which arises in this framework is how to detect the communities from data. To this end, we propose an algorithm that uses the eigenvectors of a function of the estimated autoregressive matrices to consistently recover the communities. The methodology is applied to study clustering in industrial production among a group of major economies. We also introduce a regularised VAR estimator motivated by the algorithm, which performs favourably relative to a number of alternatives in an out-of-sample forecasting exercise.

CO510 Room F2 STATISTICAL MODELING IN ELECTRICITY MARKETS Chair: Jonas Andersson

C0539: **Aggregating or diversifying risk: Transmission flows and prices between two wind power areas**
Presenter: **Johannes Mauritzen**, BI Norwegian Business School, Norway
Denmark was an early leader in building out wind power and could typically export to its neighbours in periods of excess and import in periods of shortage. However, Sweden has dramatically increased its wind power capacity. We use a multilevel time series model estimated by Bayesian MCMC to explore whether the risk and volatility from intermittent generations tends to aggregate or diversifies when it is spread spatially.

C0709: **Causality in quantiles and dynamic relations in energy markets**
Presenter: **Jonas Andersson**, Norwegian School of Economics, Norway
Co-authors: Evangelos Kyritsis
The aim is to investigate the non-linear causal relationships between crude oil price and a set of energy prices, namely diesel, gasoline, heating, and natural gas prices, from the perspective of conditional quantiles, by using monthly data for the United States over the period from January 1997 to December 2017. As opposed to a large part of the literature that tests non-causality in a certain moment, we test the null hypothesis of non-causality in various quantile intervals. By doing so, we provide more insights about the complete causal relationship, since the latter is better described by a portfolio of local measures of causality computed in different locations of the conditional distribution than a one-number measure of causality. We find significant causal relationships between the employed energy prices, especially in the tail quantiles, but also a bi-directional causal relationship between energy prices for which the classical Granger non-causality test suggests otherwise.

C0806: **Predicting dependent electricity price spikes through copula functions**
Presenter: **Enrico Foscolo**, Free University of Bozen-Bolzano, Italy
The aim is to forecast the occurrence of extreme prices (the so-called spikes) in the Australian electricity markets from half-hourly electricity spot prices. Specifically, we are interested in the simultaneous occurrences of such spikes in two or more interconnected markets. In order to do so, we propose a copula-based econometric framework for the prediction of co-spike probabilities in interconnected markets. The model allows for a flexible choice of the marginal link functions, while the dependence structure among marginal occurrences is described by means of a copula function. Compared with benchmarks assuming independence among markets or standard multivariate choice models, our approach reveals itself to be more conservative on predicting simultaneous occurrences of extreme prices.

C1534: **Renewables intermittency versus power system (in)flexibility: New insights from tail-index estimates**
Presenter: **Evangelos Kyritsis**, VATT Institute for Economic Research, Finland
Co-authors: Cristian Stet, Ronald Huisman
Power prices are well known to have a fat-tailed probability distribution function. These fat tails represent the probability of sudden high and low prices that occur as a result of sudden changes in supply, partly due to intermittent renewable power generation, and/or in demand due to inelastic short-term demand and the absence of sufficient storage capacity. In addition to the consensus view in the literature, that the level of intermittent renewable power supply has an effect on power prices and volatility, we show that the level of wind and solar supply has an effect on the fatness of the tail of empirical power price distributions. The more wind and solar supply, the fatter the tails of the left side of the distribution function and the thinner the tails on the right side. From our empirical findings on the tail asymmetry in the distribution of electricity prices, we conclude that the power system has higher need for downward than for upward power system flexibility during times with high share of intermittent renewable energy in the power system and the opposite during times with low share of intermittent renewable energy in the power system. A better understanding of the need for power system flexibility is the first step before actually adopting new measures and designing/updating the energy policy in the market.

CO126 Room G2 TEXT MINING IN ECONOMICS AND FINANCE Chair: Peter Winker

C0686: **Expectations, disagreement and news**
Presenter: **Philipp Adaemmer**, Helmut Schmidt University, Germany
Co-authors: Joscha Beckmann, Rainer Schuessler
An increasing amount of research focuses on the effects of news and uncertainty on macroeconomic aggregates. Although it is widely agreed that uncertainty exhibits various transmission channels with regard to the real economy and financial markets, little is known about the effects of economic news on macroeconomic and financial expectations. Recent advances in natural language processing have made it feasible to quantify vast amounts of written texts without relying on pre-determined keywords or manual compilations. We thus combine a correlated topic model and a dictionary based sentiment analysis to extract economic topics from approx. 500,000 U.S. newspaper articles. The results are used to investigate which type of news is correlated with professional economic forecasts and whether this relationship is varying over time. We use a flexible version of dynamic model averaging for the econometric analysis, which allows us to combine a large set of dynamic logistic regression models, differing with respect to the included explanatory variables and the degree of time variation in the parameters. The model's weight within the combination is based on the data support for each individual model, that is, its likelihood. The newspaper articles are obtained from LexisNexis Group and the survey data from Consensus Economics.

C0474: **Business cycle narratives**
Presenter: **Leif Anders Thorsrud**, Norges Bank, Norway
Research about narratives' role in economics is scarce, while real word experience and research in other sciences suggest they matter a lot. The aim is to propose a view and methodology for quantifying the epidemiology of media narratives relevant to business cycles in the US, Japan, and Europe. We do so by first constructing quantitative measures of narratives based on the news topics the media writes about. We then estimate daily business cycle indexes using this type of data, derive virality indexes capturing the extent to which narratives relevant for growth go viral, and finally use so called "Graphical Granger causality" modeling to cast light on cross-country spillovers and whether or not narratives carry news or noise. Our results highlight the informativeness of narratives for describing economic fluctuations, have a clear practical relevance for high-frequency business cycle monitoring, and suggest that narratives capture more than the market's animal spirits.

C1118: **Classifying patents based on their semantic content**
Presenter: **Antonin Bergeaud**, Banque de France, France
Co-authors: Juste Raimbault, Yoann Potiron
Some usual techniques of classification resulting from a large-scale data-mining and network approach are extended. This new technology, which in particular is designed to be suitable to big data, is used to construct an open consolidated database from raw data on 4 million patents taken from the US patent office from 1976 onward. To build the pattern network, not only do we look at each patent title, but we also examine their full abstract and extract the relevant keywords accordingly. We refer to this classification as semantic approach in contrast with the more common technological approach which consists in taking the topology when considering US Patent office technological classes. Moreover, we document that both approaches have highly different topological measures and strong statistical evidence that they feature a different model. This suggests that our method is a useful tool to extract endogenous information.

C0652: **Comparing the relevance of topics in economic journals**
Presenter: **Peter Winker**, University of Giessen, Germany
Co-authors: David Lenz
The comparison of information content of different sources is relevant. We present an approach for comparing text corpora, specifically articles in economic journals. The focus is on the development of the relevance of topics over time and its correlation across journals. Similar questions arise in many fields and, if at all, are mostly answered qualitatively. We present a quantitative framework for comparing text corpora using text mining techniques. Paragraph Vector Topic Modeling is applied to identify latent topics in text corpora and time information is utilized to track the evolution of these topics. This allows the comparison of corpus compositions over time. Three comparison methods are evaluated: Treat both text corpora as a single corpus, train a model on one corpus and evaluate the other corpus based on this model and vice versa, and train a model for each corpus and use a matching approach for pairing corresponding topics. For the empirical application, we exploit the corpus of articles published in the Journal of Economics and Statistics and the corpus of articles published in the Review of World Economics, both from 1913 to 1940. We present topic dynamics for both corpora and information on how strong the correlation of these dynamics. Furthermore, the analysis indicates which of the methods presented above is most promising for this type of analysis. We find that the matching approach and the combined corpus approach produce very reasonable results.

CO474 Room H2 WEALTH DISTRIBUTIONS AND WEALTH INEQUALITY: THEORY AND EMPIRICS Chair: Marco Maria Sorge

C0424: **Inequality, macroeconomic performance and political polarization: An empirical analysis**
Presenter: **Juan Carlos Pena**, University of Bamberg, Germany
There is an unprecedented wave of populist movements across Western Europe and beyond in the last years. The increasing electoral support of these populist movements was a signal that at any time they can obtain the political power. This is how Italy was the first that will be governed by a populist coalition during the next years. The aim is to investigate the determinants of political polarization with special focus on income inequality. We construct a dataset for 20 advanced countries using annual data ranging from 1970 to 2016 by covering 292 parliamentary elections. We show that a) traditional mainstream parties (center-left, center, and center-right) are punished for bad economic conditions; b) far-left (populist and radical parties) parties benefit under economic distress; c) greater income inequality shift voting behavior and rise the electoral support of far-left parties; d) far- right (populist and radical parties) parties benefit with recession but not from financial crises and finally; e) It seems that social globalization plays also a role, in particular for far-right parties.

C0562: **Inequality and finance in a rent economy**
Presenter: **Alberto Russo**, Universita Politecnica delle Marche, Italy
Co-authors: Alberto Botta, Eugenio Caverzasi, Mauro Gallegati, Joseph Stiglitz
The aim is to contribute to the understanding of the interactions between finance and inequality. We investigate the ways through which income and wealth inequality may have influenced the development of modern financial systems in advanced economies, the US economy first and foremost, and how modern financial systems have then fed back on income and wealth distribution. We focus in particular on securitization and on the production of complex structured financial products. We analyse this topic by elaborating a hybrid Agent-Based Stock-Flow-Consistent (AB-SFC) macroeconomic model, encompassing heterogenous (i.e. households) and aggregate sectors. Our findings suggest that the increase in economic growth, favoured by the higher levels of credit supply coming with securitization, may determine a more unequal and financially unstable economic system. We also find that a lower degree of tax progressiveness and wider wage inequality further polarize income and wage distribution, and reduce economic growth.

C0672: **The rich and the rest: A distributional approach to wealth inequality regimes in Germany**
Presenter: **Jan Schulz**, Otto-Friedrich-University of Bamberg, Germany
The German wealth distribution is analyzed by using microdata from the *Socio-Economic Panel* and the list of the 500 richest Germans by the *manager magazin*. It replicates the finding of two decoupled distributional regimes in wealth distributions already established in the literature for other countries and sample periods as a thermal distribution with a superthermal knee. A theoretical explanation grounded in maximum entropy considerations is provided based on the different asset types the population of both regimes holds that is able to fully qualitatively explain the emergence of this structure without resorting to questionable assumptions about differences in human characteristics. The estimation reveals a comparatively high degree of inequality in both regimes at odds with conventional wisdom about wealth inequality in Germany. An analysis of mobility patterns shows furthermore that in-sample mobility is associated with the inequality of the respective distributional regime implying that inequality cannot be outweighed by a notion of equality of opportunity. Finally, a novel method is presented to cope with the problem of the missing rich prevalent in wealth data, that is, underreporting in the highest wealth regions, through extrapolating from the data not significantly affected by differential non-response using a simple scaling factor.

C0548: **Left and right: A tale of two tails of the wealth distribution**
Presenter: **Marco Maria Sorge**, University of Salerno, Italy
Co-authors: Christian Di Pietro, Marcello DAmato
The aim is to develop a simple framework of wealth transmission where credit market imperfections, indivisibilities in investment in education and the ensuing occupational mobility patterns all shape the stationary (full) distribution of wealth. We investigate the properties of the limit distribution of wealth in a simple OLG model with a bequest motive and occupational choice via educational investment. In equilibrium, wealth dynamics are described by a Kesten process, which, under mild conditions, induces a power-tailed time invariant distribution. Upward wealth mobility only occurs in the form of human capital accumulation and occupational upgrading, while for higher levels the wealth mobility is driven by the accumulation of financial wealth at the lineage level. A heavy right tail of the limit distribution emerges, even in the presence of credit market imperfections, whereas at the bottom of the support the credit constraint induces existence of a mass point (under particular circumstances, a poverty trap. Implications of multidimensional (wealth and ability) heterogeneity for policy design are also studied, with a focus on the scope for temporary (one shot) large scale interventions, as typically advocated in the presence of poverty traps, vis-á-vis persistent redistributive policies.

CO082 Room M2 EMPIRICAL MACRO Chair: Laura Jackson Young

C0550: **Extracting factors from large datasets**
Presenter: **Alessia Paccagnini**, University College Dublin, Ireland
Factor models have become very popular in both shock identification and forecasting analysis. One crucial aspect is to identify the number of factors which summarizes information of the disposable dataset. Several studies propose different methods to estimate the number of factors. In addition, the recent literature about big data has stressed the importance of the increase of the sample of datasets which could create new challenges in this topic. An analysis of different methodologies to estimate the number of factors is proposed, focusing on the increase of the dataset in both time and cross-sectional dimensions. The time variation is taken into consideration. Empirical exercises on the US economy are presented to show how the number of factors matters in shock identification and forecasting analysis.

C0619: **The nonlinear effects of uncertainty shocks**
Presenter: **Laura Jackson Young**, Bentley University, United States
Co-authors: Michael Owyang, Kevin Kliesen
It is widely believed that a rise in uncertainty can have detrimental effects on macro, micro, and financial market outcomes, as well as effects on monetary, fiscal, and regulatory policy. The majority of the evidence on the effect of uncertainty shocks on key economic variables has been produced in a linear environment. These models do not account for the fact that the level of uncertainty can affect how shocks propagate. We develop a time-varying threshold VAR in which shocks that lower uncertainty have limited linear effects but shocks that raise uncertainty above the threshold can have amplified effects. Under these circumstances, we find that uncertainty shocks have effects around three times the magnitude as those found in alternative linear models.

C0642: **Monetary policy across space and time**
Presenter: **Katerina Petrova**, University of St Andrews, United Kingdom
Co-authors: Christian Matthes, Laura Liu
Two questions are asked: (i) is the conduct of monetary policy stable across time and similar across major economies, and (ii) do policy decisions of major central banks have international spillover effects? To answer these questions, we build on recent semi-parametric advances in time-varying parameter models that allow us, in addition to accounting for the parameter drift, to increase the dimension of the VAR and to model jointly three advanced economies (US, UK and the Euroarea). Our main reduced-form finding is an increased connectedness between and within countries during the recent financial crisis. To investigate policy spillovers, we identify three economy-specific monetary policy shocks using a combination of sign and magnitude restrictions. We find that monetary policy shocks were larger in magnitude and more persistent in the early 1980s than in subsequent periods. We also uncover positive spillover effects of policy between countries in the 1980s and diminished, and sometimes negative 'beggar-thy-neighbour' effects in the second half of the sample. Moreover, during the 1980s, we find evidence for policy coordination between the three central banks.

C0847: **News vs. noise: What information is contained in the revisions of the JOLTS data**
Presenter: **Amy Guisinger**, Lafayette College, United States
Co-authors: Julie K Smith
The purpose is to examine the revisions to the major categories (openings, hires, quits and discharges) of the Job Openings and Labor Turnover Survey (JOLTS) and to test if those revisions are mostly due to news (i.e. new information that was missing from the original release) or noise (i.e. structural changes in the economy). The JOLTS data are of particular interest as they provide information about labor demand, which is less researched than labor supply yet vitally important to understand equilibrium in the labor market. We find that the revisions in the JOLTS and CES data match recent results that find that most of the revisions to the initial data appear not to be either news or noise; however, there are two interesting exceptions in our analysis. In the job openings data (which is the measure of labor demand), changes arise from better collected data (i.e. news) not from measurement error (i.e. noise). In addition, in the payroll employment data, changes tend to arise from better collected data (i.e. news) and not measurement error (i.e. noise).

CO484 Room N2 TIME SERIES ANALYSIS: SOME RECENT DEVELOPMENTS Chair: Hideatsu Tsukahara

C0779: **Modeling interval financial time series**
Presenter: **Liang-Ching Lin**, National Cheng Kung University, Taiwan
Co-authors: Li-Hsien Sun
In financial economics, a large number of analysis and models are developed based on the daily closing price, or even at lower frequencies such as weekly or monthly. However, some valuable intra-daily information such as maximum and minimum prices may be discarded. We propose an interval time series model, including the maximum, minimum and closing prices, and then apply the proposed model to forecast the interval. The likelihood function and the corresponding maximum likelihood estimates (MLEs) are obtained by using the stochastic differential equation and the Girsanov theorem. The efficiency of the proposed estimators is illustrated by the simulation study. Finally, in the real data analysis for S&P 500 index, we show that the forecast of proposed method outperforms than several alternatives.

C0921: **Backtesting, prequential analysis and prediction processes**
Presenter: **Hideatsu Tsukahara**, Seijo University, Japan
In the prequential framework, data are observed sequentially in time, and at each time it is required to make a prediction about the distribution of the next observation conditional on the past, based on the data up to that point. Sometime later, we need to check whether the sequence of our forecasts are consistent with that of observations. This ex-post examination of predictive performance is called a backtesting in finance. Through the recent controversy on backtestability issue, we now know that, depending on which aspect of the conditional distribution we try to predict, we are faced with varying difficulty in devising backtesting procedures. Some attempts are made to extend Davis' calibration concept to larger classes of statistical functions (with values in an abstract space). Comparison of two probability forecasting systems under absolute continuity condition may be interpreted in terms of the corresponding prediction processes which always possess Markov property, and we explore its implications. Computation for a few simple examples from time series analysis will be shown to exemplify the theory. Finally, the possibility of extensions to the case with auxiliary random variables (covariates) and to the continuous-time case will be discussed.

C0925: **COGARCH models: A statistical application to real data**
Presenter: **Ilia Negri**, University of Bergamo, Italy
Co-authors: Enrico Bibbona
One of the reason that suggests to use COGARCH models to fit financial log-return data is due to the fact that they are able to capture the so called stylized facts observed in real data: uncorrelated log-returns but correlated absolute log-return, time varying volatility, conditional heteroscedasticity, cluster in volatility, heavy tailed and asymmetric unconditional distributions, leverage effects. The aims are to fit the COGARCH models to a real financial data set, to estimate the parameters of the models via the prediction based estimating functions, and to look at the performance of these estimates comparing them with the estimates obtained via the method of moment estimators. Moreover, as COGARCH models can be seen as an extension of the GARCH idea to continuous in time process a comparison with the last model is also performed.

C0735: **The mixture transition distribution modeling for higher order circular Markov processes**
Presenter: **Hiroaki Ogata**, Tokyo Metropolitan University, Japan
Co-authors: Takayuki Shiohama
The stationary higher order Markov process for circular data is considered. We employ the mixture transition distribution model to express the transition density of the process. The underlying circular transition distribution is based on the Wehrly and Johnson's bivariate circular models. The structure of the circular autocorrelation function is found to be similar to the autocorrelation function of the AR process on the line. The validity of the model is assessed by applying it to a series of real directional data.

CO656 Room O2 ECOSTA JOURNAL PART A: ECONOMETRICS I Chair: Manfred Deistler

C0975: **Microeconometric dynamic panel data methods: Model specification and selection issues**
Presenter: **Jan Kiviet**, University of Amsterdam, Netherlands
A strategy is presented to implement the GMM (generalized method of moments) tools provided by Stata package Xtabond2 such that an adequate model specification and matching set of instrumental variables is extracted from a panel data set to establish a single microeconometric structural behavioral presumably dynamic relationship. In the suggested specification search three comprehensive goals are pursued. Firstly including all the relevant appropriately transformed possibly lagged regressors, as well as any interactions between these if it is required to relax the strict homogeneity restrictions on the dynamic impacts of explanatories in standard linear panel models. Secondly correctly classifying all regressors as either endogenous, predetermined or exogenous, as well as either effect-stationary or effect-nonstationary, together implying the required differencing to obtain valid and relatively strong instruments from lagged internal variables. Thirdly enhancing the inference accuracy by both imposing valid coefficient restrictions and reducing the space spanned by the set of instruments through eliminating weak instruments. For the tests which trigger the decisions to be made considerations are spelled out to interpret their p-values. Complexities to identify the dynamic impact patterns are also unraveled. Finally, the strategy is applied to a classic data set and is shown to yield some new insights.

C1453: **Particle rolling MCMC**
Presenter: **Yasuhiro Omori**, University of Tokyo, Japan
Co-authors: Naoki Awaya
An efficient simulation-based methodology is proposed for the rolling window estimation of state space models. Using the framework of the conditional sequential Monte Carlo update in the particle Markov chain Monte Carlo estimation, weighted particles are updated to learn and forget the information of new and old observations by the forward and backward block sampling with the particle simulation smoother. These particles are also propagated by the MCMC update step. Theoretical justifications are provided for the proposed estimation methodology. The computational performance is evaluated in illustrative examples, showing that the posterior distributions of model parameters and marginal likelihoods are estimated with accuracy. Finally, as a special case, our proposed method can be used as a new sequential MCMC based on particle Gibbs, which is shown to outperform SMC2 that is the promising alternative method based on particle MH in the simulation experiments.

C0229: **WTI crude oil option-implied VaR and CVaR**
Presenter: **Giovanni Barone Adesi**, Lugano University, Switzerland
Co-authors: Giovanni Barone-Adesi, Carlo Sala, Chiara Legnazzi, Marinela Finta
Using option market data, the aim is to derive naturally forward-looking, non-parametric and model-free risk estimates, three desired characteristics hardly obtainable using historical returns. The option-implied measures are only based on the first derivative of the option price with respect to the strike price, bypassing the difficult task of estimating the tail of the return distribution. We estimate and backtest the 1%, 2.5% and 5% WTI crude oil futures option-implied VaR and CVaR for the years 2011-2016 and for both tails of the distribution. Compared with risk estimations based on the Filtered Historical Simulation (FHS) methodology, our results show that the option-implied risk metrics are valid alternatives to the statistically-based historical models.

C1161: **Combining multivariate volatility models**
Presenter: **Alessandra Amendola**, Department of Economics and Statistics - University of Salerno, Italy
Co-authors: Vincenzo Candila, Giuseppe Storti, Manuela Braione
Forecasting conditional covariance matrices of returns involves a variety of modeling options. First, the choice between models based on daily or intradaily returns. Examples of the former are the Multivariate GARCH (MGARCH) models while models fitted to Realized Covariance (RC) matrices are examples of the latter. A second option, strictly related to the RC matrices, is given by the identification of the frequency at which the intradaily returns are observed. A third option concerns the proper estimation method able to guarantee unbiased parameter estimates even for large (MGARCH) models. Thus, dealing with all these modeling options is not always straightforward. A possible solution is the combination of volatility forecasts. The aim is to present a forecast combination strategy in which the combined models are selected by the Model Confidence Set (MCS) procedure, implemented under different loss functions.

CG012 Room I2 CONTRIBUTIONS IN PORTFOLIO OPTIMIZATION I Chair: Nalan Basturk

C0602: **Best subset selection in regularized sparse index tracking**
Presenter: **Nick Koning**, University of Groningen, Netherlands
Co-authors: Paul Bekker
The goal of sparse index tracking is to select a portfolio with a limited number of stocks in order to replicate an index. We propose a novel approach that combines l_1- and l_2-regularization (also known as lasso and ridge regression) with best subset selection. In order to estimate the portfolio, we connect sparse index tracking to a new approach for best subset selection in the standard linear regression model as suggested previously. We provide a numerical comparison to forward stepwise selection on historical data of stock indexes and illustrate the sensitivity of the tracking performance to the values of the parameters.

C0636: **Stochastic bounds for portfolio analysis**
Presenter: **Sofia Anyfantaki**, Bank of Greece, Greece
Co-authors: Stelios Arvanitis, Nikolas Topaloglou, Thierry Post
Concepts and methods are introduced for analyzing whether a given portfolio possibility set contains some element which dominates all portfolios in another possibility set for all risk-averse investors. A general hypothesis structure and assumption framework are employed and feasible approaches to statistical inference and numerical optimization are developed. Various applications are explored in asset pricing and portfolio analysis: (1) testing the efficiency of a latent market portfolio; (2) analyzing investment returns which are hedged for systematic risk; (3) engineering an active enhanced indexing portfolio in the face of sampling error.

C1543: **Testing for parametric orderings efficiency**
Presenter: **Matteo Malavasi**, Macquarie University, Australia
Co-authors: Sergio Ortobelli , Nikolas Topaloglou
Semi-parametric tests to evaluate the efficiency of a benchmark portfolio with respect to different stochastic orderings are developed and empirically compared. Firstly, we classify investors' choices when returns depend on a finite number of parameters: a reward measure, a risk measure and other parameters. We extend stochastic dominance theory under minimal assumptions on reward and risk measures. We prove that, when choices depend on a finite number of parameters and, when the reward measure is isotonic with investors' preference, agents behave as non satiable and risk averse, when the reward measure is lower than the mean, and behave as non satiable and risk seeker when the reward measure is greater than the mean. Then, we introduce a new stochastic ordering that is consistent with the choices of a non satiable, nor risk averse nor risk seeker investor. Secondly, we propose a methodology to semi-parametric tests for the efficiency of a portfolio, when return distribution is uniquely identified by

four parameters, using estimation function theory. Finally, we empirically test whether the Fama and French market portfolio, as well as the NYSE and the Nasdaq indexes are efficient with respect to alternative stochastic orderings.

C1721: **Equity momentum strategies at work**

Presenter: **Nalan Basturk**, Maastricht University, Netherlands

Co-authors: Lennart Hoogerheide, Herman van Dijk, Arco van Oord

Exploiting momentum in a portfolio of stocks is shown to be a profitable strategy empirically. We explore the working of several momentum strategies and propose a set of mean-variance optimized portfolio strategies that incorporate the momentum properties of assets into account. We apply these strategies to monthly NYSE and AMEX common equity returns and evaluate the performance of these strategies based on their return and risk features. The proposed strategies outperform standard momentum particularly in terms of the Sharpe ratios and crash risk. In addition, a combination of the proposed strategies indicates further improvements in return and risk features and these improvements hold under moderate transaction costs.

Friday 14.12.2018 14:40 - 16:20 Parallel Session D – CFE-CMStatistics

EI007 Room Sala Convegni NEW CHALLENGES AND STATISTICAL SOLUTIONS IN NEUROIMAGING Chair: Timothy Johnson

E0153: **A geometric approach towards evaluating fMRI preprocessing pipelines**
Presenter: **Martin Lindquist**, Johns Hopkins University, United States
The preprocessing pipelines typically used in resting-state fMRI (rs-fMRI) analysis are modular in nature, as they are composed of a number of separately developed components performed in a flexible order. We illustrate the shortcomings of this approach, as we introduce a geometrical framework to illustrate how later preprocessing steps can reintroduce artifacts that had previously been removed from the data in a prior step of the pipeline. These issues can arise in practice when any combination of common preprocessing steps such as nuisance regression, scrubbing, CompCor, and temporal filtering are performed in a modular fashion. We illustrate the problem using a few concrete examples and conclude with a general discussion of how different preprocessing steps interact with one another. These results highlight the fact that special care needs to be taken when performing preprocessing on rs-fMRI data, and the need to critically revisit previous work on rs-fMRI data that may not have adequately controlled for these types of effects.

E1707: **High-dimensional mediation methods for a study of major depressive disorder**
Presenter: **Min Qian**, Columbia University, United States
Co-authors: DuBois Bowman, Bin Cheng
Despite steady progress in the study of major depressive disorder, the pathophysiologic mechanisms underlying both successful antidepressant treatment and the persisting vulnerabilities for offspring remain poorly understood. An integral step toward filling these gaps is to investigate biological variables mediating treatment response and familial MDD risk. However, the set of multimodal neuroimaging candidate mediators may be quite large, with possibly correlated elements, posing challenges conventional methods. While statistical and machine-learning methods for handling large-scale problems continue to emerge, there is a paucity of methods for high-dimensional mediation (HDM) analysis. We present a novel statistical methodology to conduct HDM analysis and apply the new method to identify mediators of the impacts of MDD risk on overall functioning, anxiety, and depression outcomes.

E0167: **Reproducibility in functional neuroimaging studies through the lens of multiplicity**
Presenter: **Nicole Lazar**, University of Georgia, United States
Of late much attention has been focused on problems of reproducibility in the scientific literature, with many published studies failing to meet what seems a minimal standard. Functional neuroimaging research has not been immune to these criticisms. As a reaction to the "reproduciblity crisis" in science, a variety of solutions have been proposed, most of which touch on, in one way or another, issues of multiple testing and Type I error control. The reproducibility in functional neuroimaging and other large-scale data settings, via the perspective of multiplicity is discussed.

EO570 Room A1 STATISTICS AND COMPUTING FOR ANALYZING ELECTRONIC HEALTH RECORD DATA Chair: Jin Zhou

E0281: **Easily parallelizable and distributable class of algorithms for structured sparsity, with optimal acceleration**
Presenter: **Joong-Ho Won**, Seoul National University, Korea, South
Many statistical learning problems can be posed as minimization of a sum of two convex functions, one typically a composition of non-smooth and linear functions. Examples include regression under structured sparsity assumptions. Popular algorithms for solving such problems, e.g., ADMM, often involve non-trivial optimization subproblems or smoothing approximation. We consider two classes of primal-dual algorithms that do not incur these difficulties, and unify them from a perspective of monotone operator theory. From this unification we propose a continuum of preconditioned forward-backward operator splitting algorithms amenable to parallel and distributed computing. For the entire region of convergence of the whole continuum of algorithms, we establish its rates of convergence. For some known instances of this continuum, our analysis closes the gap in theory. We further exploit the unification to propose a continuum of accelerated algorithms. We show that the whole continuum attains the theoretically optimal rate of convergence. The scalability of the proposed algorithms, as well as their convergence behavior, is demonstrated up to 1.2 million variables with a distributed implementation.

E0302: **Distributed learning from multiple EHR databases: Contextual embedding models for medical events**
Presenter: **Qi Long**, University of Pennsylvania, United States
Co-authors: Ziyi Li, Xiaoqian Jiang
Electronic health records (EHRs) data offer great promises in personalized medicine. However, EHRs data also present analytical challenges due to their irregularity and complexity. In addition, analyzing EHR data involves privacy issues and sharing such data across multiple institutions/sites may be infeasible. A recent work uses contextual embedding models and successfully builds one predictive model for more than seventy common diagnoses. Although the existing model can achieve a relatively high predictive accuracy, it cannot build global models without sharing data among sites. We proposed a novel distributed method to learn from multiple databases and build predictive models: Distributed Noise Contrastive Estimation (Distributed NCE). We also extend the proposed method with differential privacy to obtain reliable data privacy protections. Our numerical studies demonstrate that the proposed method can build predictive models in a distributed fashion with privacy protection and the resulting models achieve comparable prediction accuracy compared with existing methods that use pooled data across all sites.

E0347: **Probabilistic phenotyping using diagnosis codes to improve power for genetic association studies**
Presenter: **Jennifer Sinnott**, The Ohio State University, United States
Electronic health records linked to blood samples form a powerful new data resource that can provide much larger, more diverse samples for testing associations between genetic markers and disease. However, algorithms for estimating certain phenotypes, especially those that are complex and/or difficult to diagnose, produce outcomes subject to measurement error. We recently proposed a method for analyzing case-control studies when disease status is estimated by a phenotyping algorithm; our method improves power and eliminates bias when compared to the standard approach of dichotomizing the algorithm prediction and analyzing the data as though case-control status were known perfectly. The method relies on knowing certain qualities of the algorithm, such as its sensitivity, specificity, and positive predictive value, but in practice these may not be known if no "gold-standard" phenotypes are known in the population. A common setting where that occurs is in phenome-wide association studies (PheWASs), in which a wide range of phenotypes are of interest, and all that is available for each phenotype is a surrogate measure, such as the number of billing codes for that disease. We propose a new method to perform genetic association tests in this setting, which improves power over existing methods that typically identify cases based on thresholding the number of billing codes, with applications to studies of rheumatoid arthritis in the Partners Healthcare System.

E0448: **Reliable evidence from health care data: Lessons from the OHDSI collaborative**
Presenter: **Marc Suchard**, University of California, Los Angeles, United States
Concerns over reproducibility in science extend to research using existing healthcare data; many observational studies investigating the same topic produce conflicting results, even when using the same data. To address this problem, we propose a paradigm shift. The current paradigm centers

on generating one estimate at a time using a unique study design with unknown reliability and publishing (or not) one estimate at a time. The new paradigm advocates for high-throughput observational studies using consistent and standardized methods, allowing evaluation, calibration, and unbiased dissemination to generate a more reliable and complete evidence base. We demonstrate this new paradigm by comparing all depression treatments for a set of outcomes, producing 17,718 hazard ratios, each using methodology on par with state-of-the-art studies. We furthermore include control hypotheses to evaluate and calibrate our evidence generation process. Results show good transitivity and consistency between databases, and agree with four out of the five findings from clinical trials. The distribution of effect size estimates reported in literature reveals an absence of small or null effects, with a sharp cutoff at $p = 0.05$. No such phenomena were observed in our results, suggesting more complete and more reliable evidence.

EO681 Room B1 GRAPHICAL MARKOV MODELS II Chair: Giovanni Maria Marchetti

E1480: **Random covariance associated to a weighted graph**
Presenter: **Gerard Letac**, Universite Paul Sabatier, France

Given an undirected graph with n vertices and with given weights $w(e)$ on the edge e, we consider a positive definite matrix M of order n such that the off diagonal entries are 0 or $-w(e)$. We provide the diagonal of M with a family of distributions already considered by a group of physicists around 2010 called 'Multivariate reciprocal inverse Gaussian laws' (MRIG). This family has many interesting properties: stability by marginalization and conditioning, simplicity of the Laplace transform and moments. The one dimensional margins are the familiar reciprocal inverse Gaussian (RIG). If the graph is connected the inverse of M has positive coefficients and this makes MRIG an interesting choice for a priori probabilities on concentration matrices for Gaussian graphical models.

E1694: **Marginalized local independence graphs**
Presenter: **Soeren Wengel Mogensen**, University of Copenhagen, Denmark
Co-authors: Niels Richard Hansen

Local independence is an asymmetric notion of independence which describes how a system of stochastic processes evolves over time. Let A, B, and C be three subsets of the coordinate processes of the stochastic system. Intuitively speaking, B is locally independent of A given C if at every point in time knowing the past of both A and C is not more informative about the present of B than knowing the past of C only. Previous work has used directed graphs equipped with δ-separation for graphical representation of local independence structures. In such local independence graphs each node corresponds to an entire coordinate process rather than to a single random variable. We consider marginalization of local independence graphs and introduce a class of graphs which describe partially observed local independence models. We also introduce μ-separation, a generalization of δ-separation. This class of graphs satisfies a central maximality property which allows one to construct a simple graphical representation of an entire Markov equivalence class of marginalized local independence graphs. This is convenient as the equivalence class can be learned from data and its graphical representation concisely describes what underlying structure could have generated the observed local independencies.

E1701: **Mixed graphical models for metabolic biomarkers**
Presenter: **Sofia Massa**, University of Oxford, United Kingdom

Metabolic profiling using NMR spectroscopy measures a range of circulating metabolites among other biomarkers measures. The association of these metabolic biomarkers with a phenotype of interest can help to understand the biological mechanism through which disease risk is influenced. We present a mixed graphical model for studying the association between metabolic biomarkers and a phenotype of interest. The potential of this approach will be illustrated in both simulated and real data settings.

E1423: **Coordinate-free analysis of multivariate categorical data**
Presenter: **Tamas Rudas**, Hungarian Academy of Sciences Centre for Social Sciences, Hungary
Co-authors: Anna Klimova

Graphical models generalize the notion of conditional independence among variables which, most often, is equivalent to a factorization of the joint distribution. Relational models consider more general factorizations, and generalize conditional independence to situations when the sample space is not a Cartesian product of ranges of variables, and the effects entering the factorization are not related to cylinder sets of the sample space, i.e., to groups of variables. Basic concepts of statistical modeling are introduced, which can be applied in this situation. First, motivating examples are presented, then coordinate free exponential families of probability distributions are introduced, which postulate simple multiplicative structures. Some of the properties of these families are similar to that of log-linear or graphical models, but the maximum likelihood estimates under relational models have a few very surprising characteristics.

EO568 Room C1 ALGEBRAIC STATISTICS Chair: Sonja Kuhnt

E0374: **Exact tests to compare contingency tables under quasi-independence and quasi-symmetry**
Presenter: **Fabio Rapallo**, Universita del Piemonte Orientale, Italy
Co-authors: Cristiano Bocci

In the framework of contingency tables analysis, several square tables are considered under the quasi-independence (qi) or the quasi-symmetry (qs) model. Working within the class of log-linear models, a suitable model is defined and a goodness-of-fit test is introduced in order to verify if two or more tables fit a common qi (or qs) model against the alternative that each table follows a different qi (or qs) model. Such an exact test is based on classical tools of algebraic statistics, i.e., the computation of a Markov basis and a MCMC algorithm. When the comparison is limited to two tables, the Markov bases are characterized theoretically, while for larger comparisons the computations are performed through symbolic software. Applications to social mobility tables and rater agreement problems are discussed.

E0993: **On numerical fans for noisy experimental designs**
Presenter: **Arkadius Kalka**, Dortmund University of Applied Sciences and Arts, Germany

Identifiability of models is an important issue in experimental design. The theory of Groebner basis and algebraic fans has been applied to this subject. In the case of noisy designs, e.g., when the design points themselves are observations, some models are only identifiable due to small deviations of the design. In order to avoid such unstable models, the notion of numerical algebraic fan has been developed which deploys the numerical Buchberger algorithm. We compare the numerical algebraic fan with other possible notions of numerical fans, in particular more straightforward methods which we call the numerical statistical fan. A model lies in the numerical statistical fan if its design matrix is approximatively of full rank, otherwise the columns of its design matrix are approximate linear dependent. Given some small precision parameter $\varepsilon > 0$, n vectors may be defined as approximate (or almost) linear dependent if there exists an $(n-1)$-dimensional subspace such that the sum of distances (squared) of the points to the subspace is smaller than ε. Data from a thermal spraying process is used to compare the goodness of fit of models coming from different approaches.

E1080: **Algebraic-based sampling via permutations**
Presenter: **Roberto Fontana**, Politecnico di Torino, Italy

Consider two samples of size n_1 and n_2 coming from some non-negative discrete exponential family. There exists a uniformly most powerful unbiased procedure to test if the two samples come from the same distribution, performed conditionally on the sum of their entries t. The resulting sample space is the fiber of vectors of size n_1+n_2 and entries adding up to t. Vectors can be sampled from this set by building a Markov chain Monte Carlo (MCMC) procedure which exploits Markov basis to connect all the elements of F. The fiber can be partitioned into orbits of permutations and the probability of sampling a vector y given its orbit is uniform. As a consequence, it is possible to sample orbits which are contained in F via MCMC with the appropriate Markov basis, and then y via standard Monte Carlo. The two sampling procedures above result in two well-known estimators: the indicator function and the permutation cumulative distribution function. Both estimators are unbiased, but the one provided by the permutation approach has the lowest variance and the lowest mean absolute deviation. These theoretical results are verified by a simulation study, showing that the permutation approach grants convergence in the least steps too.

E1160: **Discovery of statistical equivalence classes using computer algebra**
Presenter: **Eva Riccomagno**, Universita degli Studi di Genova, Italy

Representations of certain polynomials in terms of a nested representation are intrinsically linked to labelled event trees. A recursive formula corresponds to a construction of such polynomials from a tree and gives an algorithm on the computer algebra software CoCoA to derive all staged trees from a polynomial. We use the results in applications linked to staged tree models.

EO220 Room D1 CAUSAL PARAMETERS: IDENTIFICATION AND INFERENCE Chair: Xavier de Luna

E0370: **Causal inference accounting for unobserved confounding after outcome regression and doubly robust estimation**
Presenter: **Minna Genback**, Umea University, Sweden
Co-authors: Xavier de Luna

Causal inference with observational data can be performed under an assumption of no unobserved confounders (unconfoundedness assumption). There is, however, seldom clear subject-matter or empirical evidence for such an assumption. We therefore develop uncertainty intervals for average causal effects based on outcome regression estimators and doubly robust estimators, which provide inference taking into account both sampling variability and uncertainty due to unobserved confounders. In contrast to sampling variation, uncertainty due to unobserved confounding does not decrease with increasing sample size. The intervals introduced are obtained by modeling the treatment assignment mechanism and thereby deriving the bias of the estimators due to unobserved confounders. We are thus also able to contrast the size of the bias due to violation of the unconfoundedness assumption, with bias due to misspecification of the models used to explain potential outcomes. This is illustrated through numerical experiments where bias due to moderate unobserved confounding dominates misspecification bias for typical situations in terms of sample size and modeling assumptions. We also study the empirical coverage of the uncertainty intervals introduced and apply the results to a study of the effect of regular food intake on health. An R-package implementing the inference proposed is available.

E0393: **Causal mediation with longitudinal mediator and survival outcome**
Presenter: **Vanessa Didelez**, Leibniz Institute for Prevention Research and Epidemiology - BIPS, University of Bremen, Germany

Notions of direct and indirect causal effects are based on nested counterfactuals, i.e. on the outcome Y under the hypothetical situation that treatment X was set to x, while a mediator M was set to $M(x')$ for a different x'. These may have problems of interpretation and identifiability when the outcome of interest is a survival or time-to-event and the mediator a set of longitudinal measurements, as the mediator may not exist for as long in one world compared to another world due to different survival. This is known as the cross-world nature of nested counterfactuals. We will discuss the problem and propose an alternative approach that does not suffer from those shortcomings and makes no cross-world assumptions. The new suggestion follows an extended graphical approach where mechanisms need to be specified that separate the treatment node formalized based on an augmented DAG. We will demonstrate under what assumptions identification of such alternative mediated effects is possible, resulting in the familiar mediation g-formula. Relations to path-specific effects and edge-interventions will be discussed. The proposed new approach is founded in decision theory and while it constitutes an interesting alternative to the prevailing structural equation models, it can be shown that for the particular case of combining linear structural equations for the mediator with an additive hazard model, the familiar linear effect decomposition can be recovered.

E0865: **On causal parameters in recursive systems for binary random variables**
Presenter: **Elena Stanghellini**, University of Perugia, Italy

The relationship between parameters of marginal and conditional logistic regression models is presented when no conditional independence assumptions can be made. We show that the marginal parameters decompose into the sum of terms that vanish whenever the parameters of the conditional model vanish, in parallel with the Cochran's formula for the linear case. The derivations can be applied to decompose the marginal effect of a binary/continuous treatment on a binary response into a direct effect and an indirect effect through a binary mediator and can be extended to a recursive system of binary random variables, leading to results similar to the well-known path analysis. The interest lies on several research questions. Given a data-generating process, a researcher may wish to quantify how much of the total effect of a covariate on a response is due to intermediate variables and can be removed after conditioning on their values. From a different, though related, point of view, one may wish to quantify the distortion on some regression coefficients of interest due to the omission of relevant unmeasured covariates, and use this information to build reasonable bounds or to conduct sensitivity analysis. We further show how these derivations can used in the context of causal inference, and propose a decomposition of the total effect into terms that are due to direct and indirect effects.

E1273: **On bias reduction when estimating the causal effect of pre-emptive kidney transplantation**
Presenter: **Els Goetghebeur**, Ghent University, Belgium

While causal inference made great progress over past decades, much of the applied literature appears oblivious. We evaluate the effect of Pre-emptive (immediate) Kidney Transplantation (EKT) versus starting on dialysis in patients with end-stage kidney disease (EKD). We show how in disease registers which 1) follow patients from EKD onwards and 2) measure sufficient confounders at the start of EKD, one can not only estimate the total effect of PKD on survival time among the (un)treated but also the survival time lost for each day spent on initial dialysis (relative to starting with transplantation). This is possible through accelerated failure time models even when the switch to delayed transplantation depends on unobserved time-varying covariates. These same methods applied to data from transplant registers however do reap biased estimators due to the truncated nature of data entering conditional on survival up to (delayed) transplantation. We identify various sources of bias in this case, and discuss assumptions and likelihood based methods under which this bias can be resolved. The methods are applied to the Swedish kidney register and supported by simulations.

EO394 Room E1 RECENT ADVANCES IN DURATION TIME ANALYSIS Chair: Andreas Groll

E0301: **Regularized Cox frailty models for time varying covariates**
Presenter: **Maike Hohberg**, University of Goettingen, Germany
Co-authors: Andreas Groll

A method is proposed to regularize Cox frailty models that accommodates time-varying covariates and is based in the full likelihood. In previous simulation studies, it has been shown that using the conventional partial likelihood compared to the full likelihood yields a loss in precision especially in small or moderate samples. Given that in many medical applications for example, the sample size is often rather small, it seems surprising that none of the established `R` routines are based on the full likelihood considering that for small datasets using the full likelihood does not drastically increase computing time. We provide a function `coxlasso` that fits regularized Cox models accommodating varying coefficients effects and changes in the covariates. We show the function's superior performance compared to existing routines and assess situations in which using the full likelihood might be most effective.

E0386: **Time-to-event prediction with neural networks and Cox regression**
Presenter: **Haavard Kvamme**, University of Oslo, Norway

New methods for time-to-event prediction are proposed by extending the Cox proportional hazards model with neural networks. Building on methodology from nested case-control studies, we propose a loss function that scales well to large data sets, and enables fitting of both proportional and non-proportional extensions of the Cox model. The behavior of the proposed methods is assessed in simulation studies, and the methods are shown to perform well. In particular, the non-proportional method is able to estimate a number of different forms of the survival function. A case study on customer churn is conducted, and our non-proportional method is found to perform better than the random survival forests and almost as well as binary classifiers (deep neural networks). By clustering the survival curves obtained with our non-proportional method, it is revealed that the customers may be divided in groups with quite different churn patterns. This information is not available through a binary classifier, which only gives probability estimates for fixed follow-up times.

E0430: **Boosting methods for effects selection in Cox frailty models**
Presenter: **Andreas Groll**, Technical University Dortmund, Germany
Co-authors: Trevor Hastie, Thomas Kneib, Gerhard Tutz

As in many other sorts of regression problems, also in survival analysis it has become more and more relevant to face high-dimensional data with lots of potentially influential covariates. These generally can have time-constant or time-varying effect types, which a priori is often unknown to the modeler. A possible solution is to apply estimation methods that aim at the detection of the relevant effect structure by using regularization methods. 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 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. To address these model selection issues, a likelihood-based component-wise boosting approach is proposed that is able to distinguish between these types of effects and one obtains a sparse representation that includes the relevant effects in a proper form. The method is applied to a real world data set, illustrating that the complexity of the influence structure can be strongly reduced by using the proposed boosting approach.

E0710: **Joint modelling approaches to survival analysis via likelihood-based boosting techniques**
Presenter: **Colin Griesbach**, FAU Erlangen-Nuernberg, Germany
Co-authors: Elisabeth Waldmann, Andreas Groll

When analyzing data where event-times are recorded alongside a longitudinal outcome, one commonly used approach in practice is separate modeling of the two outcomes without considering any interaction effects. Especially in survival analysis one main interest is incorporating time-varying covariates into the model. This however is quite a challenge, since popular methods like the extended cox regression produce biased results. Joint modeling on the other hand combines a longitudinal and a survival submodel in one single joint likelihood and thus accounts for interactions like time-varying covariates measured with error, which can be often found in follow-up studies. Previous works proposed algorithms to fit joint models via component-wise gradient boosting techniques which focus on minimizing the predictive risk, offer advantages like variable selection and also work with high dimensional data. However, gradient boosting leads to problems in the survival part of the model, since time-varying effects can not be estimated so easily. Likelihood-based boosting approaches on the other hand are, as verified in various literature, capable of handling time-dependent covariates in survival analysis, since likelihood-based boosting directly optimizes the likelihood by using newton algorithms with a component-wise updating procedure.

EO422 Room F1 Y-SIS SESSION: LOW-DIMENSIONAL LEARNING OF HIGH-DIMENSIONAL DATA Chair: Daniele Durante

E0426: **Multiple kernel learning for integrative clustering in genomic precision medicine**
Presenter: **Alessandra Cabassi**, University of Cambridge, United Kingdom
Co-authors: Paul Kirk

A method is presented to integrate information from diverse, high-dimensional omics datasets, together with clinical information, in order to define clinically actionable disease subtypes. We show how kernel methods, such as kernel k-means, can be used to perform integrative clustering using multiple kernel learning, demonstrating that any symmetric positive-definite matrix representing the pairwise similarities between patients can be used as kernel matrix. Therefore, it is possible to define kernels using the output of Bayesian clustering models or consensus clustering algorithms, for instance. The use of kernel methods allows the dimension of the problem to be reduced from $N \times P$ to $N \times N$: this is a great advantage in omics applications, where P is usually much larger than N. We further extend to the (semi-) supervised setting, in which additional clinical "side information" is available (e.g. survival data), and demonstrate that this can help to guide the clustering toward more relevant stratifications. We apply these methods to cancer datasets, where we combine multiple omics datasets and use phenotypic traits as side information in order to identify disease sub-types.

E0497: **Low-rank approximations with fairness constraints**
Presenter: **Emanuele Aliverti**, University of Padova, Italy
Co-authors: David Dunson

In many high-dimensional applications, there is considerable interest in developing machine learning algorithms that are designed to achieve some notion of fairness. These considerations are crucial when algorithms aid or replace human judgement; for example, in criminal justice risk assessment or medical imaging. When standard methods fail in characterizing with efficiency and flexibility the explosion in the number of dimensions, common strategies for reducing the number of parameters include sparse modelling, latent structure analysis and low-rank factorization; however, without explicit adjustment, there are no guarantees that such methods will also be fair. The focus is on including fairness constraints in low-dimensional representation of the original data, in order to preserve fairness guarantees with respect to sensitive attributes such as race or sex, while maintaining accuracy of the approximation. Efficient computational algorithms and theoretical support for the approaches will be discussed, along with empirical results in a variety of application.

E0596: **Topological invariants for high-dimensional time series**
Presenter: **Tullia Padellini**, Sapienza University of Rome, Italy
Co-authors: Pierpaolo Brutti
When dealing with increasingly complex data, the need for identifying them through a few, interpretable features grows considerably. Topology has proven to be a useful tool in this quest for "insights on the data", since it characterises objects through their connectivity structure, i.e. connected components, loops and voids. This topological approach to data analysis (TDA) can be exploited in the case of high dimensional time series, where, in addition to investigating the relation between observations at each time, we are also interested in summarizing its time evolution. We introduce a new topological summary that takes into account both dimension of interest. Our method allows for an intuitive visualization of complex dependency structures, and, as it is based on persistent homology, it also allows for a quantitative measure of their relevance in explaining the data, i.e. persistence. We investigate the theoretical properties (such as convergence and stability) of our proposal and finally we show it in action.

E0614: **A new biclustering method for functional data**
Presenter: **Jacopo Di Iorio**, Politecnico di Milano, Italy
In recent years, thanks to the augmented possibilities in storing data, researchers started to deal with problems described by data having a huge number of features. It is the case of functional data, usually represented by a set of functions taking values into an infinite dimensional space. Another fundamental need, typical of data mining, is to study data by grouping them according to some measure of similarity: it is possible thanks to clustering techniques. Due to the fact that a large number of these algorithms cannot perform simultaneous and overlapping clustering on both the dimensions of the data, it has been proposed a new family of technique under the name of Biclustering or Co-Clustering. However, differently from clustering, there is not a large literature dedicated to these approaches suitable for functional data. The first attempts to deal with Biclustering for functional data are shown. A new possible technique to obtain biclusters in a functional setting is shown from both a theoretical and algorithmic points of view. Examples and real problem applications are analyzed and described in their results in order to highlight the importance of the introduction of these methods to the world of FDA.

EO508 Room G1 SEMIPARAMETRIC STATISTICAL METHODS FOR COMPLEX SURVIVAL DATA Chair: Sy Han Chiou

E0931: **Estimating shape and size indices for recurrent events**
Presenter: **Yifei Sun**, Columbia University, United States
Co-authors: Sy Han Chiou, Chiung-Yu Huang
Single and multiple index models are becoming increasingly popular in many scientific applications, because they allow flexibility in regression modeling and interpretability of covariate effects. We propose to model a recurrent event process via two indexes that describe the shape and size of the process. The proposed model overlaps with or includes many existing recurrent event models and retain the interpretability of regression coefficients via monotonic constraints. We develop a two-step estimating approach to estimate the indexes: We first eliminate the size parameters by conditioning on the event number and estimate the shape parameters via maximum rank correlation estimation; we then plug in the estimated shape parameters and estimate the size parameter via monotone rank estimation. Large sample properties of the estimators are established. Simulation studies and a data example are presented to illustrate the proposed methodology.

E1182: **Classification of competing risk outcomes using transition biomarkers**
Presenter: **Feng-Chang Lin**, University of North Carolina at Chapel Hill, United States
Co-authors: Quefeng Li, Jessica Lin
In the Plasmodium vivax malaria infection, relapse from previous infections and reinfection from a new mosquito bite can be considered as competing risks. Classification of the recurrent infection to either relapse or new infection becomes critical when the researcher tries to detect genetic signatures of relapse that is key to evaluating anti-vivax interventions. While one can use baseline information to build up a nave classifier for the recurrent infection, little has been suggested to use transition biomarkers that appear in the recurrent infection for classification. We will introduce a newly developed classifier that uses the transition biomarker information to enhance the accuracy of classification. The approach was examined in extensive simulation experiments when the underlying outcome is known, with superior sensitivity and specificity. A real data from 78 Cambodian Plasmodium vivax malaria patients was analyzed to demonstrate the practical use of the proposed method.

E1394: **Recurrent events analysis with data collected at informative clinical visits in electronic health records**
Presenter: **Chiung-Yu Huang**, University of California, San Francisco, United States
Co-authors: Yifei Sun
Although increasingly used as a data resource for assembling cohorts, electronic health records (EHR) pose a number of analytic challenges because they are primarily collected for clinical encounters rather than for research purpose. In particular, a patients' health status influences when and what data are recorded, generating sampling bias in the collected data. We consider recurrent event analysis using EHR data. Conventional methods for event risk analysis usually require the values of covariates to be observed throughout the follow-up period. In EHR databases, time-dependent covariates are intermittently measured during clinical visits, and the timing of these visits is informative in the sense that it depends on the disease course. Simple methods, such as the last-observation-carried-forward approach, can lead to biased estimation. On the other hand, complex joint models require additional assumptions on the covariate process and cannot be easily extended to handle multiple longitudinal predictors. By incorporating sampling weights derived from estimating the observation time process, we develop a novel estimation procedure for the semiparametric proportional rate model of recurrent events. The proposed methods do not require model specifications for the covariate processes and can easily handle multiple time-dependent covariates. The estimators for the regression parameters are asymptotically unbiased and normally distributed with a root-n convergence rate.

E1430: **Semiparametric estimation of the accelerated mean model with panel count data under informative examination times**
Presenter: **Sy Han Chiou**, University of Texas at Dallas, United States
Co-authors: Gongjun Xu, Chiung-Yu Huang, Jun Yan
Panel count data arise when the number of recurrent events experienced by each study subject is observed intermittently at discrete examination times. The validity of existing methods usually requires the examination time process being independent of the underlying recurrent event process; however, this independence assumption fails to hold in many applications. We consider a semiparametric accelerated mean model for the underlying recurrent event process and allow the two processes to be correlated through shared frailty. The model allows the regression parameters to have a simple marginal interpretation of modifying the time scale of the cumulative mean function of the event process. A novel estimation procedure for the regression parameters and the baseline rate function is proposed. In contrast to existing methods, the proposed method is robust in the sense that it requires neither the strong Poisson-type assumption for the underlying recurrent event process nor a parametric assumption on the distribution of the unobserved frailty. The asymptotic consistency of the estimator is established, and the variance of the estimator is estimated by a model-based smoothed bootstrap procedure. Numerical studies demonstrated that the proposed point estimator and variance estimator performs well with practical sample sizes. The methods are applied to data from a skin cancer chemoprevention trial.

EO416 Room H1 STATISTICAL MODELS FOR ENVIRONMENTAL PROCESSES AND HUMAN ACTIVITIES Chair: Clara Grazian

E0574: **Quantifying personal exposure to air pollution from smartphone-based location data**
Presenter: **Lucia Paci**, Universita Cattolica SC, Italy
Co-authors: Francesco Finazzi

Personal exposure assessment is a challenging task that requires both measurements of the state of the environment as well as individual's movements and their activity patterns. While ambient exposure is well studied, learning people movements represents an open issue. We show how location data collected by smartphone applications are exploited to quantify the individual exposure of a large group of people to air pollution. A Bayesian approach that blends air quality monitoring data with individual location data is proposed to assess the personal exposure over time, under uncertainty on both pollutant level and individual location. A comparison with personal exposure obtained assuming fixed locations for the individuals is also provided. Location data collected by the Earthquake Network research project are employed to quantify the dynamic exposure to fine particulate matter of around 2500 people living in Santiago (Chile) over a 4-month period. For around 30% of individuals, the personal exposure based on people movements emerges significantly different over the static exposure. The approach is flexible and can be adopted to quantify the personal exposure based on any location-aware smartphone application.

E0589: **A new design test based on a bootstrap method for response-adaptive clinical trials**
Presenter: **Marco Novelli**, Novelli, Italy
Co-authors: Alessandro Baldi Antognini, Maroussa Zagoraiou

The problem of testing hypothesis in sequential clinical trials for treatment comparisons managed via response-adaptive (RA) randomization is addressed. We propose a new bootstrap test which is more efficient and robust with respect to the classical tests proposed in the literature. In particular, through a suitably choice of the target, we introduce a new test statistic based on the treatment allocation proportion and its bootstrap estimate of the variance. We derive the theoretical properties of the suggested procedure in terms of power and ethical gain; moreover, its performance is illustrated through simulations, also compared with those of other tests suggested in the literature, showing a significant improvement from the viewpoint of inferential precision and ethical concerns as well.

E0798: **Constructing priors for varying coefficient models**
Presenter: **Maria Franco Villoria**, University of Torino, Italy
Co-authors: Massimo Ventrucci, Haavard Rue

Varying coefficient models arise naturally as a flexible extension of a simpler model where the effect of the covariate is constant. We present varying coefficient models in a unified way using the recently proposed framework of penalized complexity (PC) priors to build priors that allow proper shrinkage to the simpler model, avoiding overfitting. We illustrate their application in a case study on air pollution and hospital admissions in Turin (Italy).

E0977: **Multivariate Bayesian change-point model with concurrent breaking points**
Presenter: **Gianluca Mastrantonio**, Politecnico of Turin, Italy
Co-authors: Giovanna Jona Lasinio, Alessio Pollice, Giulia Capotorti, Lorenzo Teodonio, Giulio Genova, Carlo Blasi

Extreme temperatures and precipitations have been recorded from 1951 to 2010 in 360 stations across Italy. Motivated by this real data, we present a new multivariate change-point model. The time series are modelled through a spatio-temporal/seasonal Gaussian process, a mean dependent on elevation and with independent trivariate residuals that follow change-point models, one for each station. The change point models take into account possible temporal drifts and parameters and breaking-points can be both shared across stations. Our model is specified using the Dirichlet process in a Bayesian framework. The clusterizations are then compared with the Italian Ecorigion, that are ecologically homogeneous areas of similar potential as regards the climate, physiography, hydrography, vegetation and wildlife, used as geographic framework for interpreting ecological processes, disturbance regimes, and vegetation patterns and dynamics.

EO104 Room I1 FLEXIBLE PARAMETRIC DISTRIBUTIONS: THEORY AND APPLICATIONS Chair: Christophe Ley

E0247: **A new Fourier series based construction for circulas**
Presenter: **Arthur Pewsey**, University of Extremadura, Spain

A new Fourier series based construction is proposed for the analogues of copulas for circular distributions recently coined 'circulas'. Different patterns of real Fourier coefficients are used to generate eight different classes of models, one of which is well-known within the literature. Approaches to model identification, fitting and goodness-of-fit testing are illustrated in the analysis of bird migration data.

E0249: **On a general structure for hazard-based regression models**
Presenter: **Francisco Javier Rubio**, King's College London, United Kingdom
Co-authors: Laurent Remontet, Nicholas Jewell, Aurelien Belot

The proportional hazards model represents the most commonly assumed hazard structure when analysing time to event data using regression models. The context of excess hazard models, which is of great interest in cancer epidemiology, is also dominated by the proportional hazards assumption. We will give a brief introduction to excess hazard regression models, and we will present a general hazard structure which contains, as particular cases, proportional hazards, accelerated hazards, and accelerated failure time structures, as well as combinations of these. We combine these different hazard structures with a flexible parametric distribution (exponentiated Weibull) for modelling the baseline hazard. This distribution allows us to cover the basic hazard shapes of interest in practice: constant, bathtub, increasing, decreasing, and unimodal. An example with real data will be used to illustrate the usefulness of this model. We also illustrate the importance of studying flexible parametric distributions, with interpretable parameters and good inferential properties that control the shape of the hazard.

E0272: **Modulated-symmetry-type skew distributions for directional data**
Presenter: **Christophe Ley**, Ghent University, Belgium

The modulation or perturbation of symmetry is one of the most popular methods to produce skew distributions on the real line and in R^k. The main driving force in this research area was a seminal paper that introduced the skew-normal distribution. The idea of symmetry-modulation is simple: take a symmetric density and modulate it by multiplication with a skewing function. The resulting density is of a simple form and exhibits many nice properties. We will show how, over the past years, the idea of symmetry-modulation has been successfully extended to the world of directional data, be it on the circle, sphere or cylinder.

E0698: **Small area estimation of inequality measures under alternative distribution models**
Presenter: **Silvia Pacei**, University of Bologna, Italy
Co-authors: Maria Rosaria Ferrante

Small area statistics on economic inequality are becoming important for better planning public regional policies. We focus on the estimation of entropy inequality measures in Italian provinces by using data taken from the EU-SILC sample survey for Italy. As EU-SILC survey is planned

to provide reliable estimates for areas that are larger than those we are interested in, the number of units sampled at provincial level is generally too small to obtain reliable estimates, and the use of small area estimation models is advisable. We consider small area models specified at area level that include the direct survey weighted estimators. In these models direct estimators are assumed to be unbiasedness and normally distributed. Due to the range of values that these estimators can assume and to the possible asymmetry of their distribution, the normality assumption could be inadequate in small samples. Therefore, more flexible distributions are compared and explored as alternative to the normal one. Moreover, inequality direct estimators are known to be biased for small sample sizes. A correction that can produce approximately unbiased direct estimators, taking into account the complexity of the survey design, is also proposed.

EO044 Room L1 ON SOME RECENT RESULTS IN SUPERVISED AND UNSUPERVISED CLASSIFICATION I Chair: Geoffrey McLachlan

E1108: **Fast Gaussian mixture model estimation using online EM algorithms**
Presenter: **Hien Nguyen**, La Trobe University, Australia
The use of the Gaussian mixture model for model-based clustering and classification is ubiquitous in the modern analysis of multivariate data. Unfortunately, the estimation of Gaussian mixture models is often computationally burdensome, especially when the dimension of the data and the number of observations are large. Modern trends in optimisation theory have leaned towards the usage of stochastic algorithms for high-dimensional and big data optimisation problems. A number of interesting online algorithms for stochastic estimation of Gaussian mixture models have been made available. We implement some of these algorithms in R and compare the performance of such algorithms against some algorithms that are currently available and implemented in R.

E1149: **Extending robust fuzzy clustering to skew data**
Presenter: **Francesca Greselin**, University of Milano Bicocca, Italy
Co-authors: Luis Angel Garcia-Escudero, Agustin Mayo-Iscar
Clustering is an important technique in exploratory data analysis, with applications in image processing, object classification, target recognition, data mining etc. The aim is to partition data according to natural classes present in it, assigning data points that are more similar to the same cluster. We solved this ill-posed problem by adopting a fuzzy clustering method, based on mixtures of skew Gaussian, endowed by the joint usage of trimming and constrained estimation of scatter matrices. A set of membership values are used to fuzzy partition the data and to contribute to the robust estimates of the mixture parameters. The purpose is to adopt the basic skew Gaussian component for the mixture and apply impartial trimming to the data, to model the skew core of the clusters and to adapt to any type of tail behaviour. The choice of the skew Gaussian components is motivated by the fact that, with the increased availability of multivariate datasets, often underlying asymmetric structures appear. In these cases, the extremely useful paradigm for clustering given by the mixtures of Gaussian distributions appeared somehow unrealistic. Moreover, impartial trimming provides robust ML estimation, even in presence of outliers in the data. Finally, synthetic and real data are analyzed, to show how intermediate membership values are estimated for observations lying at cluster overlap, while cluster cores are composed by observations that are assigned to a cluster in a crisp way.

E0989: **Matrix sketching and supervised classification**
Presenter: **Roberta Falcone**, University of Bologna, Italy
Co-authors: Laura Anderlucci, Angela Montanari
Matrix sketching is a recently developed data compression technique. An input matrix A is efficiently approximated with a smaller matrix B, so that B preserves most of the properties of A up to some guaranteed approximation ratio. In so doing numerical operations on big data sets become faster. Sketching algorithms generally use random projections to compress the original dataset and this stochastic generation process makes them amenable to statistical analysis. The statistical properties of sketched regression algorithms have been widely studied previously. We study the performances of sketching algorithms in the supervised classification context, both in terms of misclassification rate and of boundary approximation, as the degree of sketching increases. We also address, through sketching, the issue of unbalanced classes, which hampers most of the common classification methods.

E0252: **Bayesian mixtures of multiple scale distributions**
Presenter: **Florence Forbes**, INRIA, France
Co-authors: Alexis Arnaud, Benjamin Lemasson, Emmanuel Barbier, Russel Steele
Multiple scale distributions are multivariate distributions that exhibit a variety of shapes not necessarily elliptical while remaining analytical and tractable. We consider mixtures of such distributions for their ability to handle non standard, typically non-Gaussian clustering tasks. We propose a Bayesian formulation of the mixtures and a tractable inference procedure based on variational approximation. The interest of such a Bayesian formulation is illustrated on an important mixture model selection task, which is the issue of selecting automatically the number of components. We derive promising procedures that can be carried out in a single-run, in contrast to the more costly comparison of information criteria.

EO048 Room M1 FUNCTIONAL DATA ANALYSIS AND MORE Chair: Jane-Ling Wang

E0476: **Bootstrapping max statistics in high dimensions: Near-parametric rates and application to functional data analysis**
Presenter: **Miles Lopes**, UC Davis, United States
Co-authors: Zhenhua Lin, Hans-Georg Mueller
In recent years, bootstrap methods have drawn attention for their ability to approximate the laws of "max statistics" in high-dimensional problems. A leading example of such a statistic is the coordinate-wise maximum of a sample average of n random vectors in R^p. Existing results for this statistic show that bootstrap consistency can be achieved when $n << p$, and rates of approximation (in Kolmogorov distance) have been obtained with only logarithmic dependence in p. Nevertheless, one of the challenging aspects of this setting is that established rates tend to scale like $n^{-1/6}$ as a function of n. The main purpose is to demonstrate that improvement in rate is possible when extra model structure is available. Specifically, we show that if the coordinate-wise variances of the observations exhibit decay, then a nearly $n^{-1/2}$ rate can be achieved, independent of p. Furthermore, a surprising aspect of this dimension-free rate is that it holds even when the decay is very weak. As a numerical illustration, we show how these ideas can be used in the context of functional data analysis to construct simultaneous confidence intervals for the Fourier coefficients of a mean function.

E0551: **Prediction using averaging estimated functional linear regression models**
Presenter: **Jeng-Min Chiou**, Academia Sinica, Taiwan
Co-authors: Xinyu Zhang, Yanyuan Ma
A novel model averaging approach is proposed to predict the functional response variable. We develop a cross-validation model averaging estimator based on functional linear regression models in which the response and the covariate are both treated as random functions. The weights chosen by the method are asymptotically optimal in the sense that the squared error loss of the predicted function is as small as that of the infeasible best possible averaged function. Monte Carlo studies and a data application indicate that in most cases the approach performs better than model selection.

E0670: **On the optimal reconstruction of partially observed functional data**
Presenter: **Alois Kneip**, University of Bonn, Germany
Co-authors: Dominik Liebl

A new reconstruction operator is proposed that aims to recover the missing parts of a function given the observed parts. This new operator belongs to a new, very large class of functional operators which includes the classical regression operators as a special case. We show the optimality of our reconstruction operator and demonstrate that the usually considered regression operators generally cannot be optimal reconstruction operators. Our estimation theory allows for autocorrelated functional data and considers the practically relevant situation in which each of the n functions is observed at *m* discretization points. We derive rates of consistency for our nonparametric estimation procedures using double asymptotics. For data situations, as in our real data application where *m* is considerably smaller than *n*, we show that our functional principal components based estimator can provide better rates of convergence than any conventional nonparametric smoothing method.

E1177: **Analyzing functional data over complex multi-dimensional domains**
Presenter: **Laura Sangalli**, Politecnico di Milano, Italy
Co-authors: Eleonora Arnone, Luca Negri

A novel class of models is presented for the analysis of functional data defined over complex multidimensional domains, including curved bi-dimensional domains and complex three-dimensional domains. This class of models includes smoothing methods, regression methods and principal component analysis methods. These are implemented using numerical techniques such as finite elements and they are based on the idea of differential regularizations. We will illustrate the methods via an application to the study of neuroimaging data. In this applicative domain, the proposed methods offer important advantages with respect to the best state-of-the-art techniques, allowing to correctly take into account to complex anatomy of the brain.

EO512 Room N1 COMPLEX DEPENDENCE IN EXTREMES Chair: Jenny Wadsworth

E0405: **A generalized additive framework for estimating covariate effects on extremal dependence based on threshold exceedances**
Presenter: **Thomas Opitz**, BioSP, INRA, France
Co-authors: Valerie Chavez-Demoulin, Linda Mhalla

The probability and structure of co-occurrences of extreme values in multivariate data may critically depend on auxiliary information provided by covariates. We develop a flexible generalized additive modeling framework based on high threshold exceedances for estimating covariate-dependent joint tail characteristics for regimes of asymptotic dependence and asymptotic independence. The framework is based on suitably defined marginal pretransformations and projections of the random vector along the directions of the unit simplex, which lead to convenient univariate representations of multivariate exceedances based on the exponential distribution. We illustrate this modeling framework on a large dataset of nitrogen dioxide measurements recorded in France between 1999 and 2012, where we use the generalized additive framework for modeling marginal distributions and tail dependence in monthly maxima. Results imply asymptotic independence of data observed at different stations. We find that the estimated coefficients of tail dependence decrease as a function of spatial distance. Differences further arise in the patterns for different years and for different types of stations (traffic vs. background).

E0462: **Hypothesis testing for tail dependence parameters on the boundary of the parameter space**
Presenter: **Anna Kiriliouk**, University of Namur, Belgium

Modelling multivariate tail dependence is one of the key challenges in extreme-value theory. The max-linear model is a parametric tail dependence model which is dense in the class of multivariate extreme-value models. Being non-differentiable, it cannot be estimated using likelihood-based methods, so that minimum distance estimation forms a valuable alternative. Currently, estimation is limited to the set-up where the number of factors and/or the structure of the model is defined a priori, because answering these questions necessitates estimation and testing at the boundary of the parameter space. The main goal is to propose hypothesis tests for tail dependence parameters that, under the null hypothesis, are on the boundary of the alternative hypothesis. We give the asymptotic distribution of the weighted least squares estimator when the true parameter is on the boundary of the parameter space, and we propose two test statistics whose asymptotic distribution is easily computable. An extensive simulation study evaluates the performance of the test statistics, which are then applied to the stock market prices of two NYSE companies.

E0591: **Extremal dependence properties of vine copulas**
Presenter: **Emma Simpson**, Lancaster University, United Kingdom
Co-authors: Jenny Wadsworth, Jonathan Tawn, Ingrid Hobaek Haff

Vine copulas form a class of multivariate dependence model, and are composed of a series of bivariate copulas with certain underlying dependence structures. These models are flexible, and the use of pair copulas in their construction means that they extend well to moderate or high dimensions. We investigate the tail dependence properties of such models. In particular, we examine some of the extremal dependence structures that can be achieved in vine copula modelling, and demonstrate how to calculate the coefficients of tail dependence, denoted by η, for this class of models. We present examples in the trivariate case, with pair copulas being either logistic (asymptotically dependent) or inverted logistic (asymptotically independent), and use geometric approaches to find the corresponding bivariate and trivariate η values. We also explain how the same approaches can be extended for studying vine copulas in higher dimensions.

E1303: **An asymptotically justified framework for modelling extreme ocean states with Markov processes and tail graphs**
Presenter: **Ioannis Papastathopoulos**, University of Edinburgh, United Kingdom

Recent developments on asymptotic characterizations of extremes of Markov processes are described. We flesh out flexible extreme value models that are used to infer spatio-temporal characteristics of extreme ocean states. The advantage of the proposed statistical models is that they facilitate spatio-temporal graphical modelling of variables such as significant wave height, wind speed and current speed, but also combinations of the different variables within one model, with quite general extremal dependence structure between nodes. This is particularly useful because the most important characteristics of environmental extremes are not contemporaneous. For example, the extremum of significant wave height within a storm may not coincide with extrema of wind speed or wave peak frequency and the proposed models offer a way of understanding this temporal incoherence. The methodology is also appropriate to describe the evolution of individual storm events including the evolution of inter dependent wave, wind and current fields in space and time. We outline a likelihood based procedure, model selection criteria and shrinkage methods for estimating graphical structures and illustrate the proposed methods with an application to extremes of North Sea waves.

EO350 Room P1 SHRINKAGE METHODS FOR ANALYZING COMPLEX DATA Chair: Ines Wilms

E0523: **High-dimensional variable selection when features are sparse**
Presenter: **Jacob Bien**, University of Southern California, United States
It is common in modern prediction problems for many predictor variables to be counts of rarely occurring events. This leads to design matrices in which a large number of columns are highly sparse. The challenge posed by such "rare features" has received little attention despite its prevalence in diverse areas, ranging from biology (e.g., rare species) to natural language processing (e.g., rare words). We show, both theoretically and empirically, that not explicitly accounting for the rareness of features can greatly reduce the effectiveness of an analysis. We next propose a framework for aggregating rare features into denser features in a flexible manner that creates better predictors of the response. An application to online hotel reviews demonstrates the gain in accuracy achievable by proper treatment of rare words.

E0537: **Nearest comoment estimation with unobserved factors**
Presenter: **Dries Cornilly**, Vrije Universiteit Brussel, Belgium
Co-authors: Kris Boudt, Tim Verdonck
A minimum distance estimator is proposed for the higher order comoments of a multi-dimensional distribution exhibiting a lower dimensional latent factor structure. We derive the influence function of the proposed estimator and prove its consistency and asymptotic normality. The simulation study confirms the large gains in accuracy compared to the traditional sample comoments. The empirical usefulness of the novel framework is shown in applications of portfolio allocation under non-Gaussian objective functions and factor extraction in a dataset containing mental ability scores.

E0839: **Sparse identification and estimation of high-dimensional vector autoregressive moving averages**
Presenter: **Sumanta Basu**, Cornell University, United States
Co-authors: Ines Wilms, David Matteson, Jacob Bien
The Vector AutoRegressive Moving Average (VARMA) model is fundamental to the study of multivariate time series. However, estimation becomes challenging in even relatively low-dimensional VARMA models. With growing interest in the simultaneous modeling of large numbers of marginal time series, many authors have abandoned the VARMA model in favor of the Vector AutoRegressive (VAR) model, which is seen as a simpler alternative, both in theory and practice, in this high-dimensional context. However, even very simple VARMA models can be very complicated to represent using only VAR modelling. We develop a new approach to VARMA identification and propose a two-phase method for estimation. Our identification and estimation strategies are linked in their use of sparsity-inducing convex regularizers, which favor VARMA models that have only a small number of nonzero parameters. We establish sufficient conditions for consistency of sparse infinite-order VAR estimates in high dimensions, a key ingredient for our two-phase sparse VARMA estimation strategy. The proposed framework has good estimation and forecast accuracy under numerous simulation settings. We illustrate the forecast performance of the sparse VARMA models for several application domains, including macro-economic forecasting, demand forecasting, and volatility forecasting. The proposed sparse VARMA estimator gives parsimonious forecast models that lead to important gains in relative forecast accuracy.

E0918: **Sparse semiparametric canonical correlation analysis for data of mixed types**
Presenter: **Irina Gaynanova**, Texas A and M University, United States
Co-authors: Grace Yoon, Raymond Carroll
Canonical correlation analysis investigates linear relationships between two sets of variables, but often works poorly on modern data sets due to high-dimensionality and mixed data types (continuous/binary/zero-inflated). We propose a new approach for sparse canonical correlation analysis of mixed data types that does not require explicit parametric assumptions. The main contribution is the use of truncated latent Gaussian copula to model the data with excess zeroes, which allows us to derive a rank-based estimator of latent correlation matrix without the estimation of marginal transformation functions. The resulting semiparametric sparse canonical correlation analysis method works well in high-dimensional settings as demonstrated via numerical studies, and application to the analysis of association between gene expression and micro RNA data of breast cancer patients.

EO204 Room Q1 OUTLIERS AND STRUCTURAL BREAKS Chair: Alexander Duerre

E0310: **Detecting multiple generalised change-points by isolating single ones**
Presenter: **Andreas Anastasiou**, The London School of Economics and Political Science, United Kingdom
Co-authors: Piotr Fryzlewicz
A new approach is introduced, called Isolate-Detect (ID), for the consistent estimation of the number and location of multiple generalised change-points in noisy data sequences. Examples of signal changes that ID can deal with, are changes in the mean of a piecewise-constant signal and changes in the trend, accompanied by discontinuities or not, in the piecewise-linear model. The method is based on an isolation technique, which prevents the consideration of intervals that contain more than one change-point. This isolation allows for detection in the presence of frequent changes of possibly small magnitudes. Thresholding and model selection through an information criterion are the two stopping rules are described. A hybrid of both criteria leads to a general method with very good practical performance and minimal parameter choice. Examples from both simulated and real-life data will be given; ID is at least as accurate as the state-of-the-art methods.

E0410: **Estimation of the spatial weighting matrix for spatiotemporal data with structural breaks**
Presenter: **Philipp Otto**, European University Viadrina, Germany
Co-authors: Rick Steinert
A two-step penalized regression approach is proposed to estimate the entire spatial dependence structure of a spatiotemporal process under the presence of structural breaks in the mean. Simultaneously, we address an important problem in spatial econometrics. The classical approach is to replace the unknown spatial dependence structure by a linear combination of a scalar and a predefined, non-stochastic weighting matrix describing the dependence. In contrast to this classical approach, we estimate all entries of this weighting matrix by a penalized regression approach. In addition, we suppose that there might be an unknown number of structural breaks in the data. These breaks can occur at different time points for each location and they can be of different magnitude. For estimation of the model parameters, we propose a two-step estimation approach. In the first step, we determine a set of candidate change points assuming independent univariate time series. Consequently, the spatial dependence structure is ignored in this first step. This set is then passed to the full model to estimate the changes and the spatial dependence simultaneously by an adaptive lasso approach.

E1068: **Robust change point tests using bounded transformations**
Presenter: **Alexander Duerre**, TU Dortmund, Germany
Co-authors: Roland Fried
Classical moment based change point tests like the cusum test are very powerful under Gaussian time series with no more than one change point but behave poorly under heavy tailed distributions and corrupted data. A new class of robust change point tests based on cusum statistics of

robustly transformed observations is proposed. This framework is very flexible, depending on the used transformation one can detect amongst others changes in the mean, scale or dependence of a possibly multivariate time series. The calculation of p-values can be simplified by using asymptotics which yields a computational complexity of $Tlog(T)$ where T is the number of observations. We apply our general approach to detect changes in the covariance structure of a multivariate time series. Simulations indicate high power under Gaussianity as well as heavy tails.

E1337: **Comparison of different estimators of the long run variance for the cusum test**
Presenter: **Julia Duda**, TU Dortmund, Germany
Co-authors: Alexander Duerre, Roland Fried

The cusum test is one of the most popular tools in change-point detection. Under short range dependence and fairly mild technical conditions cusum type tests depend only on one nuisance parameter, often called long run variance, but apart from that, they are distribution free. There are many proposals to estimate the long run variance non-parametrically which are all challenged by simultaneously being unbiased under the null hypothesis and giving reasonable results under the alternative. If one does not account for a possible change-point, estimators of the long run variance get inflated in case of a level shift yielding a considerable loss in power. On the other hand, accounting for a change-point often leads to a conspicuous bias in small samples resulting in substantially anti-conservative tests. We review and compare different proposals to estimate the long run variance, mainly concentrating on the two most popular approaches: kernel and bootstrap estimators. An extensive simulation study also reveals suitable tuning parameters of the presented procedures.

EO658 Room O2 ECOSTA JOURNAL PART B: STATISTICS I Chair: Byeong Park

E1122: **Jensen-Shannon divergence as a goodness-of-fit measure for maximum likelihood estimation and curve fitting**
Presenter: **Mark Levene**, Birkbeck University of London, United Kingdom

The coefficient of determination, known as R^2, is commonly used as a goodness-of-fit criterion for fitting linear models. R^2 is somewhat controversial when fitting nonlinear models, although it may be generalised on a case-by-case basis to deal with specific models such as the logistic model. Assume we are fitting a parametric distribution to a data set using the maximum likelihood estimation method. A general approach to measure the goodness-of-fit of the fitted parameters, which we advocate herein, is to use a nonparametric measure for model comparison between the raw data and the fitted model. In particular, for this purpose we put forward the *Jensen-Shannon divergence* (JSD) as a metric, which is bounded and has an intuitive information-theoretic interpretation. We demonstrate, via a straightforward procedure making use of the JSD, that it can be used as part of maximum likelihood estimation or curve fitting as a measure of goodness-of-fit, including the construction of a confidence interval for the fitted parametric distribution. We also propose that the JSD can be used more generally in nonparametric hypothesis testing for model selection.

E1617: **Least squares and ML estimation approaches of the sufficient reduction for matrix valued predictors**
Presenter: **Efstathia Bura**, Vienna University of Technology, Austria
Co-authors: Ruth Pfeiffer

In some regression and classification settings, as in longitudinal data analysis, predictors are matrix valued. In earlier work, we proposed first-moment-based sufficient dimension reduction (SDR) methods, such as Longitudinal Sliced InverseRegression (LSIR), for combining several longitudinally measured markers into a composite marker score for prediction or regression modeling, under a mild distributional assumption and by exploiting their matrix structure. We project the dimension reduction subspace onto the tensor product of the column and row vector space of the conditional predictor mean. We propose least squares and maximum likelihood based approaches to estimate optimal combinations of matrix-valued predictors that comprise sufficient reductions in regression and classification problems and obtain a score with improved predictive accuracy as well as computational efficiency. We establish the connection with 2D-LDA (2-dimensionalLinear Discriminant Analysis), a machine learning method for the statistical analysis of images and face recognition, for which we offer estimation algorithms with optimal statistical properties.

E0662: **Prediction and robustness: Calibration of inequality indices in small areas**
Presenter: **Elvezio Ronchetti**, University of Geneva, Switzerland
Co-authors: Setareh Ranjbar, Stefan Sperlich

Firstly, a general discussion of the robustness issues in a prediction framework is provided and their implications in different areas, including classification, insurance, and estimation in finite populations is analyzed. Secondly, we illustrate more specifically these issues in the prediction of nonlinear indices (such as inequality or poverty measures) for small areas and in the presence of outliers. We propose two approaches to calibrate for the bias of nonlinear functionals, such as the Gini index and when the so-called representative outliers come from a skewed heavy tail distribution. These methods can also be used to impute missing income values, a common occurrence e.g. in labour force surveys.

E1424: **Empirical likelihood for general conditional estimating equations**
Presenter: **Valentin Patilea**, CREST-Ensai, France
Co-authors: Matthieu Marbac

The empirical likelihood (EL) is a prominent statistical inference approach that has extensively studied over the last two decades. Its fast and still ongoing development is due to some important features guaranteed by the fact that it combines the flexibility of the nonparametric methods with the effectiveness of the likelihood approach. Meanwhile, many, if not most, statistical models could be written under the form of conditional moment equations. We propose a new approach for EL in general models defined by conditional moments. The method is based on an equivalent characterization of the initial conditional moment restrictions using a set of unconditional moments that is not increasing with the sample size when the dimension of the parameter of interest is fixed. We allow for the presence of a nuisance parameter of infinite dimension and characterize a class of conditional moment equations models for which the likelihood ratio remains asymptotic pivotal chi-square distributed. The class includes many common semiparametric regression models. Some simulation experiments illustrate the effectiveness of our approach.

EO424 Room Q2 J-ISBA SESSION: ADVANCES IN BAYESIAN NONPARAMETRICS Chair: Isadora Antoniano-Villalobos

E0788: **Nonnegative Bayesian nonparametric factor models with completely random measures**
Presenter: **Fadhel Ayed**, Oxford, United Kingdom
Co-authors: Francois Caron

Bayesian nonparametric Poisson factor models for community detection are studied. We assume that each user is affiliated to an infinite number of communities, which are modeled using a Completely Random Measure (CRM). The model is flexible and allows the number of active communities to grow unboundedly with the number of users. We show how the properties of the CRM relate to the growth of the number of active communities. We also derive a Markov chain Monte Carlo algorithm for posterior inference for this class of models.

E0916: **Posterior contraction rates for Bayesian functional linear regression**
Presenter: **Giuseppe Di Benedetto**, University of Oxford, United Kingdom
Co-authors: Judith Rousseau

Functional linear regression (FLR) has been thoroughly studied in the frequentist literature. We consider a Bayesian approach to FLR with scalar

response and random functional covariate. Using a sieve prior for the slope parameter, we investigate the asymptotic properties of its posterior distribution. The model can be studied as an ill-posed inverse problem and we provide contraction rates for the direct and inverse problems, namely the posterior contraction rates with respect to the prediction risk and the L_2 norm.

E0965: **PABC: Probably approximate Bayesian computation**
Presenter: **James Ridgway**, CFM, France
Approximate Bayesian computation (ABC) is a widely used inference method in Bayesian statistics to bypass the point-wise computation of the likelihood. We develop theoretical bounds for the distance between the statistics used in ABC. We show that some versions of ABC are inherently robust to misspecification. The bounds are given in the form of oracle inequalities for a finite sample size. The dependence on the dimension of the parameter space and the number of statistics is made explicit. We apply our theoretical results to given prior distributions and data generating processes, including a non-parametric regression model. We will also show how to use a sequential Monte Carlo sampler (SMC) to sample from the pseudo-posterior, improving upon the state-of-the-art samplers.

E1041: **Bayesian nonparametric mixed effects models in microbiome data analysis**
Presenter: **Boyu Ren**, Dana Farber Cancer Institute, United States
Co-authors: Sergio Bacallado, Stefano Favaro, Curtis Huttenhower, Lorenzo Trippa
Detecting associations between microbial composition and sample characteristics is one of the most important tasks in microbiome studies. Most of the existing methods apply univariate models to single microbial species separately, with adjustments for multiple hypothesis testing. We propose a Bayesian nonparametric analysis for a generalized mixed effects linear model tailored to this application. The marginal prior on each microbial composition is the Dirichlet process, and dependence across compositions is induced through a linear combination of individual covariates, such as disease biomarkers or the subject's age, and latent factors. The latent factors capture residual variability and their dimensionality is learned from the data in a fully Bayesian procedure. We propose an efficient algorithm to sample from the posterior and visualizations of model parameters which reveal associations between covariates and microbial composition. The proposed model is validated in simulation studies and then applied to analyze a microbiome dataset for infants with Type I diabetes.

CI011 Room A0 RESAMPLING AND TIME SERIES Chair: Dimitris Politis

C0177: **Functional data analysis for continuous functions**
Presenter: **Holger Dette**, Ruhr-Universitaet Bochum, Germany
The aim is to develop data analysis methodology for functional time series in the space of all continuous functions. Our methodology is motivated by the fact that objects with rather different shapes may still have a small L_2-distance and are therefore identified as similar when using an L_2-metric. However, in applications it is often desirable to use metrics reflecting the visualization of the curves in the statistical analysis. The methodological contributions are focused on developing two-sample and change-point tests as well as confidence bands, as these procedures appear to be conducive to the proposed setting. Particular interest is put on relevant differences; that is, on not trying to test for exact equality, but rather for pre-specified deviations under the null hypothesis. The procedures are justified through large-sample theory. To ensure practicability, non-standard bootstrap procedures are developed and investigated addressing particular features that arise in the problem of testing relevant hypotheses. The finite sample properties are explored through a simulation study.

C0179: **Sieve bootstrap for functional time series**
Presenter: **Efstathios Paparoditis**, University of Cyprus, Cyprus
A new bootstrap procedure for functional time series is proposed which exploits a basic representation of the vector time series of Fourier coefficients appearing in the Karhunen-Loeve expansion of the functional process. A double, sieve-type bootstrap method to generate functional pseudo-time series is developed, which uses a finite set of functional principal components to capture the essential driving parts of the infinite dimensional process and a finite order parametric process to mimic the temporal dependence structure of the corresponding vector time series of Fourier coefficients. By allowing the number of functional principal components as well as the order of the model used to increase to infinity (at an appropriate rate) as the sample size increases, consistency of the sieve bootstrap procedure for a wide range of statistics is established. Some interesting applications in the context of variance estimation, fully functional testing and the construction of prediction intervals are discussed. An automatic data-driven method to select the bootstrap parameters is also proposed. In this context, a new variance-ratio criterion is used which explicitly takes into account the dependence of the functional time series. Some numerical examples illustrate the finite sample performance of the new bootstrap methodology proposed.

C0178: **Empirical likelihood under short- and long-range dependence**
Presenter: **Soumendra Lahiri**, North Carolina State University, United States
Different versions of block empirical likelihood methods for time series data withshort- and long-range dependence are considered. As it is well known, the standard overlapping block empirical likelihood requires a choice of the block size and also a scale adjustment to achieve a distribution free limit law. We consider nonstandard variants of the blocking rule that overcome these limitations in the short-range dependent case. We also investigate if similar properties continue to hold for these methods when the underlying process is long-range dependent.

CO548 Room A2 FINANCIAL MODELLING AND FORECASTING Chair: Ekaterini Panopoulou

C0262: **The role of jump activity and signed jumps in forecasting realized volatility**
Presenter: **Rodrigo Hizmeri**, Lancaster University, United Kingdom
Co-authors: Marwan Izzeldin, Anthony Murphy, Mike Tsionas
The gains from various jumps classifications are examined: signed, finite and infinite jumps in volatility forecasting. Using a Heterogeneous Autoregressive (HAR) framework, we illustrate gains from considering various jumps specifications. We consider different sampling frequencies and microstructure noise and document the impact on the model forecasting performance. We find that on average, jumps improve volatility forecasts. Negative jumps are more important for short horizons whilst positive jumps achieve greater gains at longer horizons. Controlling for market microstructure noise at higher frequencies, results in substantial out-of-sample improvements for both short and longer horizons. Findings from the use of a model averaging approach based on the selected models by the model confidence set, outperforms the benchmark model under the various scenarios considered.

C0407: **Analysing implied volatilities between stock and dividend markets**
Presenter: **Enoch Nii Boi Quaye**, University of Kent, United Kingdom
Co-authors: Radu Tunaru
A computational approach is proposed to compare the information in the implied volatility of stock index options to the information in the implied volatility of index dividend options. The approach uses the implied volatility surface (IVs) as a stochastic state variable accounting for the evolution of the underlying asset price process. The proposed method explicitly allows for variability in time-to-maturity, and outlines a computational

process for the aggregation of volatility measures under the Black-Scholes, the Black model and model-free approaches. The study illustrates how the implied volatility term-structure of stock index option contracts with time-to-maturity exceeding “9-months" move enough to be justified by subsequent fluctuations in dividends although contracts with time-to-maturities around “1-month", “1-3 months" and “3-9 months" move too much to be justified by subsequent changes in dividends. The IVs term-structure shows that the implied volatility of stock index options consistently exceeds that of index dividend options, thereby confirming previous criticism based on novel financial data and instruments. However, we show that the magnitude of excess implied volatility declines with long-dated time-to-maturities, suggesting that the discrepancy between the two implied volatilities is sensitive to investment horizon.

C0435: **Return prediction models and asset allocation**
Presenter: **Ekaterini Panopoulou**, University of Kent, United Kingdom
Co-authors: Iason Kynigakis
The aim is to examine the out-of-sample return predictability for stock, bond and commodity indices using a large set of economic indicators. The methods used to construct the return forecasts are from the shrinkage, latent factor and forecast combination literature. The economic and statistical value of the forecasts is examined during recessions and expansions and for periods with negative and positive returns. The benefits of return predictability and of including commodities in a portfolio are further investigated through an out-of-sample multivariate asset allocation exercise based on the mean-variance optimization framework. The effects of return predictability are assessed around business cycles, different levels of risk aversion, investment constraints, transaction costs, alternative risk measures and different ways of estimating the covariance matrix.

C0456: **Tail optimal combinations**
Presenter: **Christos Argyropoulos**, Lancaster University, United Kingdom
Co-authors: Ekaterini Panopoulou
The utility of combining densities in improving the forecasting accuracy of risk measures is investigated. Specifically, we propose the Tail Quantile Score (TQS) rule which focuses directly on the tails of the distribution by taking into account the severity of the losses alongside the probability of events, across the tail of the distribution. Our simulation exercise suggests that TQS isolates efficiently the tail related performance of the methods when compared to competing tail related scoring rules. Furthermore, we develop time-varying weighting schemes in order to evaluate the benefits of combining the densities on the tail regions and the respective benefits on the accuracy of the Value at Risk and Expected Shortfall risk forecasts. Our results suggest that the optimal weights outperform the naive combination schemes with significant improvements in the forecasting accuracy of the respective risk measures.

CO554 Room B2 HIGH-FREQUENCY FINANCIAL ECONOMETRICS Chair: Bezirgen Veliyev

C1073: **Modelling limit order book data by state-dependent Hawkes processes**
Presenter: **Mikko Pakkanen**, Imperial College London, United Kingdom
Co-authors: Maxime Morariu-Patrichi
During the past ten years, self-exciting Hawkes processes have become a popular model for high-frequency financial data, as they are able to capture the endogeneity and feedback effects in order flow data at very short time scales. In such a model based on Hawkes processes, the arrival rate of new orders depends on the past order flow, but it cannot directly depend on any state variables of the limit order book, such as the current bid/ask price or queue imbalance. To address this limitation, we introduce a novel state-dependent extension of a Hawkes process. The new framework couples the Hawkes process to a state process that influences the arrival rate of new orders whilst the arriving orders may, reciprocally, prompt the state process to move to a new state. We develop maximum likelihood estimation methodology for the new class of processes and apply it to NASDAQ limit order book data. The empirical results indicate that excitation effects in order flow depend strongly on queue imbalance.

C0337: **A regime-switching stochastic volatility model for forecasting electricity prices**
Presenter: **Peter Exterkate**, University of Sydney, Australia
Co-authors: Oskar Knapik
Three crucial challenges outstanding in the area of electricity price forecasting are addressed. Specifically, we show the importance of considering fundamental price drivers in modelling, develop new techniques for probabilistic (i.e. interval or density) forecasting of electricity prices, and introduce a universal technique for model comparison. We propose a new regime-switching stochastic volatility model with three regimes, which may be interpreted as negative jump or “drop”, normal price or “base”, and positive jump or “spike", respectively. The transition matrix between these regimes is allowed to depend on explanatory variables. Bayesian inference is employed in order to obtain predictive densities. The main focus is on short-term density forecasting in the Nord Pool intraday market. We show that the proposed model outperforms several benchmark models at this task. In particular, the incorporation of stochastic volatility, regime switching, information from the day-ahead market, and exogenous information from weather reports into the model are all shown to improve its predictive performance, without falling prey to curse of dimensionality problems.

C0391: **Parametric and semi-parametric renewal based high-frequency volatility estimator**
Presenter: **Yifan Li**, The University of Manchester, United Kingdom
High-frequency volatility is proposed to be estimated parametrically based on a renewal process in business time. We show that based on this estimator, an instantaneous volatility estimator can be constructed without involving infill asymptotics. We study the property of the integrated variance process under the assumption that the calendar time point process constructed based on the observed price process follows some parametric autoregressive processes, such as variants of the autoregressive conditional duration model or autoregressive conditional intensity model. We derive asymptotic results for our parametric volatility estimator, proving its consistency and providing a formula to estimate its asymptotic variance. Moreover, we show that a semi-parametric volatility estimator can be constructed, which is more robust to model misspecifications than a full parametric structure at the cost of some estimation efficiency. We provide simulation and empirical evidence for the validity of the estimator.

C0928: **Exploiting news analytics for volatility forecasting**
Presenter: **Simon Bodilsen**, Aarhus University, CREATES, Denmark
Using a large database of macroeconomic and firm-specific news, we study whether news sentiment can be used to enhance prediction of stock market volatility. We construct two types of news indices at the daily frequency, by properly aggregating the sentiment scores of past macroeconomic- and firm-specific news, respectively. Using reduced-form time series models for realized measures of volatility, we find evidence that the index of domestic macroeconomic news is very useful in order to predict future levels of volatility for both individual stocks and the S&P 500 Index. In particular, we find large gains in the predictions of volatilities at long horizons by including the macroeconomic index in the time series regressions. On the other hand, the predictive power of firm-specific news are found to modest in the general framework.

CO078 Room D2 FINANCIAL NETWORKS **Chair: Massimiliano Caporin**

C0877: **Financial networks via conditional autoregressive expected shortfall**
Presenter: **Giovanni Bonaccolto**, University of Enna Kore, Italy
Co-authors: Massimiliano Caporin

Several approaches have been developed to build financial networks, emphasizing the distress state of the involved companies to assess their systemic relevance. Building on this strand of the literature, we propose a new method that connects the expected shortfalls of N financial institutions. In particular, the network structure is retrieved from N expectile regression models, each of them is calibrated to estimate the expected shortfall of a company at the level θ conditional to the expected shortfalls of the $N-1$ remaining institutions. Our method is designed to deal with large values of N, as we impose an l_1-norm penalty on each expectile regression model, an effective tool that allows us to select the relevant connected institutions. In our model, we capture the temporal persistence of expectiles by including latent autoregressive components. As a last exercise, we interpret the obtained network as a financial portfolio and compare it with the optimal portfolio with the minimum expected shortfall. We implement our method on the STOXX 600 financials constituents, highlighting interesting empirical findings, especially during distress periods.

C1051: **Modeling realized volatility in big data panel**
Presenter: **Michele Costola**, SAFE, Goethe University Frankfurt, Germany
Co-authors: Massimiliano Caporin, Mauro Bernardi

A Bayesian approach to the problem of variable selection and shrinkage in high dimensional sparse vector-HAR model is proposed. The regularisation method is an extension of a previous lasso. The model allows us to include the realized volatility of a large number of assets by taking into account in each equation of the system the HAR and co-jumps components for all the considered assets. The empirical analysis is performed on high-frequency data from the US market. The model estimates, performed on a rolling basis (and with a future time-varying parameter specification) will be used to retrieve the information needed for the identification of financial networks and to evaluate, ex-post, the estimated network structures.

C0721: **Credit rating migration risk and interconnectedness in a corporate lending network**
Presenter: **Masayasu Kanno**, Nihon University, Japan

The aim is to assess the credit rating migration risk and interconnectedness among bank-to-listed firms and insurer-to-listed firms in Japan's corporate lending market during the fiscal years 2008-2015. First, a portfolio credit risk analysis is conducted by using outstanding lending data with borrowers and lenders names. The results show an expected shortfall with tail dependence of t-copula captures the heavy-tailed risk of Japanese institutions. Subsequently, the network structure of lending contracts is analyzed using network centrality measures. From the perspective of network, institutions play a central role in terms of degree centrality. Further, a stress test is undertaken by using a historical economic scenario pertaining to a credit rating migration matrix shortly after the Lehman Brothers' bankruptcy. The test finds a much sparser network structure because of a large number of firm defaults. The analysis offers banks and insurers important implications regarding the credit risk management of corporate lending.

C1462: **Maximum-entropy models in economics and finance**
Presenter: **Tiziano Squartini**, IMT School for Advanced Studies Lucca, Italy

Entropy-maximization represents the unifying concept underlying the definition of a number of methods which are now part of the discipline known as "network theory". Despite the perfect generality of this approach, a particularly fruitful application of it has been observed in disciplines like economics and finance. The methodological aspects of the aforementioned approach are illustrated, with particular emphasis on the definition of null models. The latter can be employed in a number of applications, ranging from pattern detection to network reconstruction: examples will be provided of both, by taking as case studies real-world systems, as the World Trade Web and the Dutch Interbank Network. The aforementioned framework also allows one to properly model fluctuations: the latter can be interpreted as errors affecting the estimation of the quantities of interest and strongly depend on the kind of constraints defining the maximization procedure. In order to illustrate how different reconstruction algorithms perform, a number of proposed approaches will be compared on the aforementioned real-world systems.

CO601 Room E2 MODELLING EXPECTATIONS: DIFFERENT ANALYTICAL PERSPECTIVES **Chair: Maritta Paloviita**

C0409: **Effects of monetary policy decisions on professional forecasters' expectations and expectations uncertainty**
Presenter: **Maritta Paloviita**, Bank of Finland, Finland
Co-authors: Sami Oinonen, Matti Viren

The aim is to examine how professional forecasters' expectations and expectation uncertainty have reacted to the ECB interest rate decisions and non-conventional monetary policy measures during the period 1999-2017. The analysis makes use of a conventional dif-in-dif type set up. It indicates that expectations have been sensitive to policy actions, but is not clear that these forecasters' reactions follow some basics of economic theory. This is true both for the response of expectations to policy actions and the relationship between the level of the policy rate and inflation and output point forecasts. Moreover, short- and long term reaction are sometimes puzzlingly different. The most interesting results are obtained with the subjective forecast uncertainty measures which seem to be surprisingly sensitive to policy measures. Uncertainty measures, including long-term inflation uncertainty are problematic from the point of view of anchoring inflation expectations to the inflation target.

C0866: **Macroeconomic literacy and expectations**
Presenter: **Luba Petersen**, Simon Fraser University, Canada
Co-authors: Michael Mirdamadi

The effects of macroeconomic literacy on expectation formation is explored in an experimental economy where participants' aggregated expectations endogenously influence macroeconomic variables. We systematically vary the information participants receive about the economy's data-generating process (no information, qualitative information, and quantitative information) and the predictability of the economy. Our experimental results suggest there are many advantages to providing precise quantitative training about the macroeconomy. Compared to an environment where forecasters have no initial information about the structure of the economy, quantitative information about the underlying data-generating process consistently reduces inflation forecast errors, reduce disagreements about inflation, and encourages a larger reaction to past forecast errors. Inflation variability is on average lower with quantitative information. Qualitative information, by contrast, is inconsistently effective at influencing forecasting behavior.

C1265: **Disagreement in consumer inflation expectations**
Presenter: **Tomasz Lyziak**, National Bank of Poland, Poland
Co-authors: Xuguang Sheng

It is posited that consumers form expectations about inflation by combining two sources of information: their beliefs from shopping experience and news about inflation they learn from experts. Disagreement among consumers in our model comes from four sources: (i) consumers' divergent prior beliefs, (ii) heterogeneity in their propensities to learn from experts, (iii) experts' different views about future inflation, and (iv) difference in

mean expectations between consumers and experts. By carefully matching the datasets from the Michigan survey of consumers with the survey of professional forecasters, we find that inflation expectations between households and experts differ substantially and persistently from each other, and households pay close attention to salient price changes, while experts respond more to monetary policy and macro indicators. Our empirical estimates imply economically significant degrees of information rigidity and these estimates vary substantially across households. This significant heterogeneity poses a great challenge for the canonical sticky-information model that assumes a single rate of information acquisition and for noisy-information model in which all agents place the same weight on new information received.

C1582: **Consumers' perception of inflation in inflationary and deflationary environment**
Presenter: **Ewa Stanislawska**, Narodowy Bank Polski, Poland

Survey data on quantitative inflation perceptions are employed to investigate the formation of consumers' opinions about current price developments. Firstly, we compare Polish consumers' estimates of price changes to the consumer price index (CPI) and find out that consumers react more quickly to inflation increases than decreases, and that they ignore small moves in inflation. Moreover, previously stable relation between inflation perception and the CPI inflation was seriously disrupted during deflationary period, leading to smaller perception bias. Secondly, we relax the assumption that consumers perceive price changes in the CPI terms and show that prices of food, housing, water, gas, electricity and clothing - contrary to transport prices - have greater impact on inflation perception than on CPI inflation. Selective attention of consumers to price changes and asymmetric reaction to increases and falls in prices do not explain inflation perception during deflation.

CO486 Room G2 MIXTURE MODELS, IDENTIFICATION, AND FINANCIAL MODELING Chair: Markus Haas

C0592: **Likelihood-based dynamic asset pricing**
Presenter: **Dennis Umlandt**, Kiel University, Germany
Co-authors: Stefan Reitz

A new parametric approach is proposed to estimate linear factor models with time-varying risk premia. In contrast to recent contributions in the literature, our framework abstains from introducing instrument variables to describe the time variation of risk prices. This is particularly useful in situations where instrument variables are unavailable or of poor quality and misspecification should be circumvented. Risk prices are derived from a generalized autoregressive score (GAS) model where parameters dynamics are driven by the scaled score of the observation density. Estimation and inference is conducted by likelihood maximization. We assess the potential improvement for predicting risk prices in a simulation study. Moreover, several applications to classical factor pricing models like the CAPM and the Fama-French 3-factor model are presented.

C0615: **Changes on realized correlations, betas and volatility spillover in the agricultural commodity market**
Presenter: **Matteo Bonato**, IPAG Business School, Switzerland

The aim is to provide new insights on the changes in the dynamics of price correlations and spillover effects in the commodity market. Using US-traded futures price data at a 1-minute frequency over the 2002- 2017 period, we consider the interaction within soft and grain commodities and between these commodities and oil. We rely on a recently introduced volatility model - the realized Beta GARCH model. Our results reveal that soft commodities were segmented prior to 2008 and became correlated thereafter. The nature of the increase in correlation is only temporary. The correlations within grains - already significant and positive - increased only marginally, indicating that this group has been less affected by recent events. The correlation between oil and agricultural commodities, which reached its peak in 2008, has also reverted to pre-crisis level. Spillover effects between oil and commodities have become more prominent prior to the commodity price crash. However, this increase in volatility transmission tends to precede the increase in correlations. Finally, the impact of these findings on the performance of hedging strategies is discussed. Our results are important for investors exposed to the commodity market as they show that while the diversification benefits of investing in this market have decreased, volatility transmission risk and hedging costs have increased.

C0486: **Model risk of expected shortfall**
Presenter: **Emese Lazar**, University of Reading, United Kingdom
Co-authors: Ning Zhang

The model risk of Expected Shortfall (ES) is studied, extending previous results on model risk of Value-at-Risk (VaR). We propose a correction formula for ES based on passing three backtests. Our results show that for the DJIA index, the smallest corrections are required for the ES estimates built using GARCH models. Furthermore, the 2.5% ES requires smaller corrections for model risk than the 1% VaR, which advocates the replacement of VaR with ES as recommended by the Basel Committee. Also, if the model risk of VaR is taken into account, then the correction made to ES estimates reduces by 50% on average.

C0828: **Conditional skewness and kurtosis of aggregated normal mixture and Markov-switching GARCH returns**
Presenter: **Markus Haas**, University of Kiel, Germany

The first four conditional moments of aggregated returns generated by normal mixture and Markov-switching GARCH processes are derived. The results illustrate the considerable flexibility and great variety of the conditional density profiles that can emerge from this class of conditional volatility models. Moreover, the usefulness of the results for approximating multi-step-ahead conditional densities and Value-at-Risk measures is demonstrated.

CO192 Room I2 MULTIVARIATE VOLATILITY AND RISK Chair: Jean-Michel Zakoian

C0412: **Volatility estimation when observations are missing**
Presenter: **Genaro Sucarrat**, BI Norwegian Business School, Norway
Co-authors: Natalia Bahamonde, Hamdi Raissi

In empirical practice observations are often missing. This invalidates standard estimation methods of Generalised Autoregressive Conditional Heteroscedasticity (GARCH) models because of a repeated invertibility problem induced at each missing location. To sidestep this problem we propose a log-ARCH model – i.e. no GARCH-terms – with stochastic conditioning covariates (e.g. volatility proxies) that can be estimated with least squares methods. Apart from omitted GARCH terms, however, the model is very general and flexible: It is asymmetric, multivariate, allows for (unknown) Dynamic Conditional Correlations (DCCs), and non-negativity constraints are not needed on the parameters nor on the covariates. We derive a least squares equation-by-equation estimator of the model, and prove its Consistency and Asymptotic Normality (CAN) when the missing data process is stationary, unknown and not necessarily independent of the log-ARCH process itself. Our results are illustrated in a simulation study, and in an empirical application.

C0688: **A multivariate dynamic mixture model for discrete price changes at high frequency**
Presenter: **Leopoldo Catania**, Aarhus BBS, Denmark
Co-authors: Paolo Santucci de Magistris, Roberto Di Mari

High frequency price changes of financial assets are usually assumed to follow a distribution with continuous support and time-varying parameters. However, the tick structure of the financial markets entails that price changes observed at very high frequency are discrete. We start from this

empirical evidence to develop a new model able to describe the dynamic properties of a multivariate time-series of high frequency price changes, including the high probability of observing no variations (zeroes). We assume the existence of two independent latent processes determining the dynamic properties of the price changes and the probability of the occurrence of zeroes. Given the probabilistic structure embedded in our modelling framework we analyze the different sources of this large amount of zeroes as for example: absence of news, same magnitude of positive and negative news, and periods of market illiquidity. Furthermore, we propose a multivariate model to investigate the dynamics of the zeroes across several assets.

C0756: **Asymptotics of Cholesky GARCH models and time-varying conditional betas**
Presenter: **Christian Francq**, CREST and University Lille III, France
Co-authors: Sebastien Laurent, Serge Darolles
A new model with time-varying slope coefficients is proposed. The model, called CHAR, is a Cholesky-GARCH model, based on a previous Cholesky decomposition of the conditional variance matrix introduced in the context of longitudinal data. We derive stationarity and invertibility conditions and prove consistency and asymptotic normality of the full and equation-by-equation QML estimators of this model. We then show that this class of models is useful to estimate conditional betas and compare it a previous approach, the dynamic conditional beta model. Finally, we use real data in a portfolio and risk management exercise. We find that the CHAR model outperforms a model with constant betas as well as the dynamic conditional beta model.

C1117: **Virtual historical simulation for estimating the conditional VaR of large portfolios**
Presenter: **Jean-Michel Zakoian**, CREST, France
Co-authors: Christian Francq
In order to estimate the conditional risk of a portfolio's return, two strategies can be advocated. A multivariate strategy requires estimating a dynamic model for the vector of risk factors, which is often challenging, when at all possible, for large portfolios. A univariate approach based on a dynamic model for the portfolio's return seems more attractive. However, when the combination of the individual returns is time varying, the portfolio's return series is typically non stationary which may invalidate statistical inference. An alternative approach consists in reconstituting a virtual portfolio, whose returns are built using the current composition of the portfolio and for which a stationary dynamic model can be estimated. The asymptotic properties of this method, that we call Virtual Historical Simulation (VHS), are established. Numerical illustrations on simulated and real data are provided.

CO090 Room M2 TOPICS IN MACROECONOMETRICS Chair: Alessia Paccagnini

C0208: **Identification and inference in a vector autoregressive model with common frailty**
Presenter: **Federico Carlini**, USI, Lugano, Switzerland
Co-authors: Patrick Gagliardini
The aim is to study a vector autoregressive model of order one augmented by unobservable factors with a dynamic described by a vector autoregression of order one. A detailed discussion of different identification strategies is provided. We estimate the model parameters by means of a 3-step procedure such that each step is in closed-form (up to eigenvalues-eigenvectors decomposition of matrices of small dimension). We study the asymptotic and finite-sample properties of the estimators.

C0536: **Debt overhang and monetary policy transmission: An international perspective**
Presenter: **Eleonora Granziera**, Bank of Finland, Finland
The aim is to investigate how the level of household indebtedness affects the monetary transmission mechanism in advanced economies, using state-dependent local projection methods. In particular, we explore whether the impact of monetary policy shocks on output, bank credit and other macroeconomic and financial variables are less pronounced during periods of high household debt, reminiscent of a debt overhang. We exploit the time-series variation within countries, as well as the cross-sectional variation across countries such as the prevalence of fixed- versus adjustable-rate mortgages, to investigate this issue. We then build a small-scale model, where households face collateral and debt-service constraints and are subject to income shocks, to rationalize these facts. The model points to the weakening of the home equity loan channel and increased debt aversion, especially during recessions, as a possible reason for the decline in monetary policy effectiveness when initial debt levels are high.

C1096: **Detecting breaks in the group-structure of high-dimensional data**
Presenter: **Stefano Soccorsi**, Department of Economics, Lancaster University Management School, United Kingdom
Co-authors: Haeran Cho
The focus is on large panels of time series with common factors pervasive to all cross sectional units and group factors pervasive only within an unknown cluster of time series according to a group structure which is subject to multiple change points. We study the problem of estimating group memberships; extending previous results, we do so while allowing them to change over time in a piecewise-stationary framework. Consistent change-point detection, factor estimation and clustering are established, while in an empirical application on stock market data we show the usefulness of allowing for time variation in second order property of the data and their cluster memberships.

C1692: **Uncertainty and financial stability: A VAR analysis**
Presenter: **Chiara Scotti**, Board of Governors of the Federal Reserve System, United States
Co-authors: Dario Caldara, Molin Zhong
The aim is to study the relative importance of three shocks for macroeconomic and financial conditions: real, financial and uncertainty shocks. We find that shocks to financial uncertainty play a supplementary role in addition to real and financial shocks, and lead to a deterioration of the macroeconomic and financial outlook. Asset valuations in equities, corporate markets, and real estate decline. Businesses and households increase their savings and start a long-lasting process of deleveraging. The supply of credit to the economy retrenches, with underwriting standards tightening especially for commercial real estate and commercial and industrial loans. The financial sector experiences a deleveraging in banks, and a buildup of leverage in other nonfinancial institutions.

CO480 Room N2 RECENT ISSUES ON THE IDENTIFICATION OF SVAR MODELS Chair: Emanuele Bacchiocchi

C1150: **The structure condition in SVAR identification**
Presenter: **Riccardo Lucchetti**, Universita Politecnica delle Marche, Italy
Co-authors: Emanuele Bacchiocchi
A criterion for checking local identification in Structural VAR based on the structure of the constraints has been previously proposed. While the structure condition is in fact necessary and sufficient, we prove that the checking procedure originally proposed contained a flaw and put forward an amendment.

C1184: **Robust shrinkage for set-identified SVARs**
Presenter: **Alessio Volpicella**, Queen Mary University of London, United Kingdom
Set-identified SVARs, which relax exclusion restrictions and rely on weaker assumptions such as sign restrictions, are increasingly common. However, a known drawback is that the inference is rarely informative. It is shown that robust restrictions on the Forecast Error Variance (FEV) decomposition may dramatically shrink the inference. Specifically, these restrictions are consistent with the implications of a variety of different DSGE models, with both real and nominal frictions, and with sufficiently wide ranges for their parameters. First, in a bivariate and trivariate setting, restrictions on the FEV decomposition are proven to be more informative than traditional sign restrictions. Second, sufficient conditions are provided to guarantee that the identified set is non-empty and convex. Finally, two applications are provided: using models of monetary policy and technology shocks, restrictions on the FEV decomposition tend to be highly informative, greatly shrink and even change the inference of models originally identified via traditional sign restrictions. Remarkably, shrinkage in inference is robust to the recent concerns over the unintended consequences of rotation matrix prior.

C1213: **The dark side of the SVAR: A trip into the local identification world**
Presenter: **Emanuele Bacchiocchi**, University of Milan, Italy
Co-authors: Toru Kitagawa
The focus is on structural vector autoregressions where the identification issue can be addressed only locally. In this particular case, there are different isolated points in the parametric space satisfying the imposed restrictions. Unfortunately, all these points are observationally equivalent and it is impossible to chose among them simply by considering the likelihood function. The estimation of the parameters, thus, is subject to the algorithm and the related starting values used for maximizing the likelihood function. We address this problem by considering all the locally identified parameters and use Bayesian techniques to make inference on the estimated impulse responses.

C1248: **Identifying shocks via time-varying volatility**
Presenter: **Daniel Lewis**, Federal Reserve Bank of New York, United States
Under specific parametric assumptions, an n-variable structural vector auto-regression (SVAR) can be identified (up to $n!$ shock orderings) via heteroskedasticity of the structural shocks. We show that misspecification of the heteroskedasticity process can bias results derived from these identification schemes. We propose a new identification method that identifies the SVAR up to $n!$ shock orderings by using only moment equations implied by an arbitrary stochastic process for the variance. Unlike previous work, this result requires only weak technical conditions. In particular, it requires neither parametric assumptions nor the specification of variance regimes. We propose intuitive criteria to select among the orderings and show that this selection does not impact inference asymptotically. As an empirical illustration, we consider oil prices and their macroeconomic effects. This exercise strengthens his results by failing to reject his lower-triangular assumption and replicating his macroeconomic conclusions.

CO627 Room P2 BAYESIAN HIERARCHICAL MODELLING Chair: Helga Wagner

C0210: **Implicit copulas from Bayesian regularized regression smoothers**
Presenter: **Nadja Klein**, Humboldt University of Berlin, Germany
Co-authors: Michael Smith
The aim is to extract the implicit copula of a response vector from a Bayesian regularized regression smoother with Gaussian disturbances. The copula can be used to compare smoothers that employ different shrinkage priors and function bases. We illustrate with three popular choices of shrinkage priors - a pairwise prior, the horseshoe prior and a g-prior augmented with a point mass as employed for Bayesian variable selection - and both univariate and multivariate function bases. The implicit copulas are high-dimensional, have flexible dependence structures that are far from that of a Gaussian copula, and are unavailable in closed form. However, we show how they can be evaluated by first constructing a Gaussian copula conditional on the regularization parameters, and then integrating over these. Combined with non-parametric margins the regularized smoothers can be used to model the distribution of non-Gaussian univariate responses conditional on the covariates. Efficient Markov chain Monte Carlo schemes for evaluating the copula are given for this case. Using both simulated and real data, we show how such copula smoothing models can improve the quality of resulting function estimates and predictive distributions.

C0232: **Effect selection in distributional regression**
Presenter: **Manuel Carlan**, University of Goettingen, Germany
Co-authors: Nadja Klein, Thomas Kneib, Stefan Lang, Helga Wagner
A spike-and-slab prior specification is proposed that allows us to carry the concept of Bayesian variable selection over to general effect selection within the class of structured additive distributional regression models comprising various effect types such as non-linear effects, varying coefficients, spatial effects, random effects as well as potentially hierarchical regression structures. The spike-and-slab prior is assigned to the prior standard deviation of blocks of regression coefficients which allows us to work with a scalar quantity instead of dealing with possibly high-dimensional effect vectors. Furthermore, we specify the model in a redundant parameterisation with parameter expansion that yields improved shrinkage and sampling performance compared to the classical normal-inverse-gamma prior. We investigate the propriety of the posterior distribution, show that the proposed prior structure yields desirable shrinkage properties, propose an interpretable way of eliciting prior parameters and provide an efficient Markov Monte Carlo sampling scheme. Using both simulated data and three real data sets, we show that our approach is applicable for data with a potentially large number of covariates, multilevel predictor structures accounting for hierarchically nested data and a wide range of non-standard response distributions such as bivariate normal or zero-inflated Poisson with regression effects on all distributional parameters.

C0442: **Bayesian effect fusion for categorical predictors in logistic regression**
Presenter: **Magdalena Leitner**, Johannes Kepler University Linz, Austria
Co-authors: Helga Wagner
For regression type models sparsity is an important goal as usually a large number of covariates is available in applications. Many methods have been developed for variable selection and for the selection of level effects of categorical covariates. However, for a categorical covariate, which is captured by a group of level effects, a sparse representation of its effect can also be achieved by fusing predictor levels that have essentially the same effect on the response. In a Bayesian framework, this can be accomplished by specifying appropriate prior distributions. In order to encourage fusion of level effects in logistic regression models, we extend methods developed for Bayesian effect fusion in linear regression. In the first method a joint multivariate Normal prior is specified on all level effects associated with one covariate. Equivalently, spike and slab priors can be specified on the effect differences of a covariate, considering the linear restrictions between them. The second method relies on a sparse finite mixture prior of spiky components constructed to encourage clustering of level effects that are almost identical. For both prior distributions, MCMC sampling is feasible using data augmentation with latent Polya-Gamma variables. We compare the performance of both approaches for simulated data and illustrate the methods by analyzing a real data set.

C0597: **Variable selection in Bayesian latent class analysis using shrinkage priors**
Presenter: **Gertraud Malsiner-Walli**, WU Vienna University of Business and Economics, Austria
Co-authors: Bettina Gruen

Latent class analysis uses mixture models to model multivariate categorical data where dependencies between variables are observed due to the presence of latent groups. Each latent group corresponds to a component in the mixture model and the variable distributions are independent within components. Applications of the latent class model are widespread within fields where categorical data are frequently collected such as medicine or the social sciences. Crucial issues in performing latent class analysis are to select a suitable number of filled components and the variables which are most informative for distinguishing between the components. Within a Bayesian context we propose an approach which addresses these two issues simultaneously. We specify suitable shrinkage priors for the component weights as well as the component-specific success probabilities. They induce sparsity with respect to the number of filled components as well as heterogeneity of the success probabilities across components. Standard estimation methods for Bayesian mixture models can be employed since the only difference to a standard Bayesian latent class analysis lies in the specification of suitable hierarchical priors with appropriate hyperparameter values. The application of this approach is investigated using simulation studies as well as real data.

CC645 Room F2 CONTRIBUTIONS IN TIME SERIES I Chair: Cristina Amado

C1516: **The identification problem for linear rational expectations models**
Presenter: **Majid Al Sadoon**, Universitat Pompeu Fabra, Spain
Co-authors: Piotr Zwiernik

The problem considered is that of the identification of stationary linear rational expectations models from the second moments of observable data. Observationally equivalence is characterized and necessary and sufficient conditions are provided for: (i) identification under affine restrictions, (ii) generic identification under affine restrictions of analytically parametrized models, and (iii) local identification under general non-linear parametrization or non-linear constraints. The results strongly resemble the classical theory for VARMAX models although significant points of departure are also documented.

C1235: **The T ratio test for a bilinear unit root under general conditions with applications**
Presenter: **Julio Angel Afonso-Rodriguez**, University of the Balearic Islands, Spain

Nonstationarity is a great stylized fact of many macroeconomic and financial time series, where the particular type of nonstationary behaviour determines its theoretical properties and conditions the inferential procedures to be used in their empirical analysis. We study a general class of globally nonstationary processes, called a stochastic unit root (STUR), which generalizes the fixed unit root case generating periods of stationary, nonstationary and explosive behaviour, and that contains several different particular cases as, e.g., the so-called bilinear unit root process (BLUR). For this process we propose a general weak limiting distribution in the case of possibly serially correlated error terms, and study the properties of some unit root tests under this representation, both against a stationary and a STUR alternative. Among these test statistics, we found that the T-ratio test for a BLUR alternative based on an augmented auxiliary regression automatically corrects for serially correlated regular error terms and for any choice of the number of lags (thus controlling the empirical size), and only shows a slight power loss for high values of the number of lags. We illustrate all these theoretical findings with an extensive simulation exercise and with several empirical applications.

C1350: **Robust estimation of a dynamic spatiotemporal model with structural change for count data**
Presenter: **Charlene Mae Celoso**, University of the Philippines Diliman, Philippines

A dynamic spatiotemporal model with count response is estimated with a hybrid of forward search algorithm and bootstrap embedded into the backfitting algorithm. The method is evaluated for its robustness in some count data generating process. Simulation studies indicated that the method performs better in terms of median absolute deviation (MAD) when there are more time points than observations units in space and when the covariates contributes more than the spatial externalities. Also, the bias and the standard error of the parameter estimates indicated its suitability for count data across a wide variety of conditions.

C1426: **Inference on irregularly spaced time series**
Presenter: **Fulvia Marotta**, Queen Mary University of London, United Kingdom

Time series data are usually recorded at regularly spaced time intervals. In spite of this, for a variety of reasons, in many fields time series can be recorded as irregularly spaced observations. Irregularly spaced time series refers to the case when the sampled time series presents irregularly spaced observations. Point estimation and large sample statistical inference are developed for time series with irregularly spaced data. Specifying the setting we make a distinction between calendar time and intrinsic, operational time. When these two times do not coincide, we have unevenly spaced elements in the sample. We first focus on the question of estimating the sample mean when data is not regularly spaced. We provide an expression for the sample mean estimator and we establish asymptotic properties and central limit theorem. Subsequently, we construct a consistent estimator for the variance of the sample mean estimator. Finite sample properties of the estimator are investigated in a Monte Carlo study which confirms the good performance of such an estimator.

Friday 14.12.2018 16:50 - 18:30 Parallel Session E – CFE-CMStatistics

EI452 Room Sala Convegni ADVANCES IN FUNCTIONAL DATA ANALYSIS Chair: Jeng-Min Chiou

E0164: **Additive regression with Hilbertian responses**
Presenter: **Byeong Park**, Seoul National University, Korea, South
A foundation of methodology and theory for the estimation of structured nonparametric regression models with Hilbertian responses is developed. The method and theory are focused on the additive model, while the main ideas may be adapted to other structured models. For this, the notion of Bochner integration is introduced for Banach-space-valued maps as a generalization of Lebesgue integration. Several statistical properties of Bochner integrals, relevant for our method and theory, and also of importance in their own right, are presented for the first time. Our theory is complete. The existence of our estimators and the convergence of a practical algorithm that evaluates the estimators are established. These results are nonasymptotic as well as asymptotic. Furthermore, it is proved that the estimators achieve the univariate rates in pointwise, L2 and uniform convergence, and that the estimators of the component maps converge jointly in distribution to Gaussian random elements. Our numerical examples include the cases of functional, density-valued and simplex valued responses, which demonstrates the validity of our approach.

E1140: **Modeling longitudinal compositional data as trajectories on Riemannian manifolds**
Presenter: **Hans-Georg Mueller**, University of California Davis, United States
Co-authors: Xiongtao Dai
Longitudinal compositional data exhibit dependencies over time as well as among the p components of the compositional vectors, as these are constrained to be non-negative and to sum to 1. Such data are encountered in various applications, including the longitudinal modeling of behaviors for samples of individuals and in many other contexts, for example in metabolomics. This type of data can be represented as trajectories on the positive quadrant of a $(p-1)$-dimensional sphere. This motivates the study of functional data with trajectories on smooth Riemannian manifolds, with spheres as a special case. Of particular interest is the associated Riemannian functional principal component analysis. The proposed methods are supported by theory and will be illustrated with samples of trajectories consisting of compositional as well as other types of data.

E1445: **Subject-specific functional prediction from electronic health records using an enriched Dirichlet process mixture model**
Presenter: **Jason Roy**, Rutgers University, United States
In many modern applications, there is interest in predicting subject-specific functions of a variable over time. For example, we might want to know patient-specific trends in a biomarker over time. Modeling is needed if there is measurement error in the variable, or if gaps between data collection times is too wide. We propose a novel semiparametric model for the joint distribution of a continuous longitudinal outcome and the baseline covariates using an enriched Dirichlet process (EDP) prior. This joint model decomposes into subject-specific linear mixed models for the outcome given the covariates and simple marginals for the covariates. The nonparametric EDP prior is placed on the regression and spline coefficients, the error variance, and the parameters governing the predictor space. We predict the outcome at unobserved time points for subjects with data at other time points as well as for completely new subjects with covariates only. We find improved prediction over mixed models with Dirichlet process (DP) priors when there are a large number of covariates. The method is demonstrated with electronic health records consisting of initiators of second generation antipsychotic medications, which are known to increase the risk of diabetes. We use our model to predict laboratory values indicative of diabetes for each individual and assess incidence of suspected diabetes from the predicted dataset.

EO528 Room A0 INTERACTIONS BETWEEN COMPUTATION AND INFERENCE IN HIGH-DIMENSIONAL DATA Chair: Xiaodong Li

E0289: **On the likelihood ratio test in high-dimensional logistic regressions**
Presenter: **Yuxin Chen**, Princeton University, United States
Co-authors: Pragya Sur, Emmanuel Candes
Logistic regression is used thousands of times a day to fit data and assess the statistical significance of explanatory variables. When used for statistical inference, logistic models produce p-values for the regression coefficients by using a large-sample approximation to the distribution of the likelihood-ratio test (LRT). However, this asymptotic approximation is grossly incorrect when the number p of explanatory variables is comparable to the sample size n; in fact, this approximation produces p-values that are far too small (under the null hypothesis). We show that in the high-dimensional regime, the LRT converges to a rescaled chi-square, where the rescaling factor can be determined by solving a nonlinear system of two equations with two unknowns. We also complement our mathematical study by showing that the new limiting distribution is accurate for finite sample sizes. The results also extend to some other regression models such as the probit regression model.

E0357: **Statistical inference in large Ising graphical models via quadratic programming**
Presenter: **Zhao Ren**, University of Pittsburgh, United States
Co-authors: Cun-Hui Zhang, Harrison Zhou, Sai Li
The high dimensional graphical model, a powerful tool for studying conditional dependency relationship of random variables, has attracted great attention in recent years. A statistical inference of each edge for large Ising graphical models is investigated. Significant progress has been achieved recently in computing confidence intervals and p-values for each edge. The key role in these new inferential methods is played by a linear projection method to de-bias an initial regularized estimator. Major drawback of this approach in Ising models is that an extra sparsity assumption on the linear projection coefficient besides the sparsity of the graph itself is required, which cannot be checked in practice. In addition, efficiency is often compromised by the usage of sample splitting in these methods. We propose a novel estimator of each edge via quadratic programming and show that our estimator is asymptotically normal without the above mentioned extra sparsity condition. Our proof applies a novel low dimensional maximum likelihood method for the de-bias procedure and a data swap technique to avoid loss of efficiency. We further show that whenever the extra sparsity condition is satisfied, our estimator is adaptively efficient and achieves the Fisher information. Otherwise, we still provide a restricted Fisher information as a lower bound.

E0356: **Edge sampling using network local information**
Presenter: **Can Minh Le**, University of California, Davis, United States
Edge sampling is an important topic in network analysis. It provides a natural way to reduce network size while retaining desired features of the original network. Sampling methods that only use local information are common in practice as they do not require access to the entire network and can be parallelized easily. Despite promising empirical performance, most of these methods are derived from heuristic considerations and therefore still lack theoretical justification. To address this issue, we study a simple edge sampling scheme that uses network local information. We show that when local connectivity is sufficiently strong, the sampled network satisfies a strong spectral property. We quantify the strength of local connectivity by a global parameter and relate it to more common network statistics such as clustering coefficient and Ricci curvature. Based on this result, we also derive a condition under which a hypergraph can be sampled and reduced to a weighted network.

E0684: **Early stopping for gradient type algorithms**
Presenter: **Yuting Wei**, Stanford University, United States
The behavior of boosting or gradient-type algorithms in non-parametric estimation will be discussed. While non-parametric models offer great flexibility, they can lead to overfitting and thus poor generalization performance. For this reason, procedures for fitting these models must involve some form of regularization. Although early-stopping of iterative algorithms is a widely-used form of regularization in statistics and optimization, it is less well-understood than its analogue based on penalized regularization. We will establish a direct connection between these two through a general bound of the excess risk for penalized M-estimators. Based on this new insight, we are able to give an explicit and optimal stopping criteria for boosting algorithms run in reproducing kernel Hilbert spaces which is standard in non-parametric estimation, and then generalize it to broader classes of functions.

EO118 Room A1 STATISTICAL METHODS IN NEUROSCIENCE Chair: Jeff Goldsmith

E0181: **Nonlinear normalization methods for harmonization in neuroimaging data**
Presenter: **Russell Shinohara**, University of Pennsylvania Perelman School of Medicine, United States
Magnetic resonance imaging (MRI) allows for the in vivo study of neurological and psychiatric disorders. Conventional MRI, which is widely used in both research and clinical settings, is acquired in arbitrary units and differences across scanners vary nonlinearly in volumetric and intensity space. We present and contrast statistical approaches for reducing inter-scan and inter-scanner variation for improved generalizability and comparability of MRI-based measures of volume, brain tissue integrity, and function in the presence of focal and distributed pathology.

E0423: **Addressing partial volume effects using intra-subject locally adjusted cerebral blood flow images**
Presenter: **Kristin Linn**, University of Pennsylvania, United States
Co-authors: Russell Shinohara, Alessandra Valcarcel, Simon Vandekar
Local cortical coupling is a subject-specific measure of the spatially varying relationship between cortical thickness and sulcal depth. Although it is a promising first step towards understanding local covariance patterns between two image-derived measurements, a more general coupling framework that can accommodate multiple volumetric imaging modalities is warranted. We first introduce Inter-Modal Coupling (IMCo), an analogue of local coupling in volumetric space that can be used to produce subject-level, spatially varying feature maps derived from two volumetric imaging modalities. We then leverage IMCo to address partial volume effects when studying localized relationships between gray matter density and cerebral blood flow (CBF) among participants in the Philadelphia Neurodevelopmental Cohort. We show that when CBF images are adjusted for partial volume effects at the subject level using our method, we have more power to detect non-linear interactions between age and sex in voxelwise analyses. We call the proposed IMCo-adjusted CBF images Intra-Subject Locally Adjusted Cerebral Blood Flow (ISLA-CBF) images.

E0840: **Joint and individual non-Gaussian component analysis**
Presenter: **Benjamin Risk**, Emory University, United States
Co-authors: Irina Gaynanova
As advances in technology allow the acquisition of complementary information, it is increasingly common for scientific studies to collect multiple data sets. Large-scale neuroimaging studies often include multiple imaging modalities (e.g., functional MRI, diffusion MRI, and/or structural MRI) and behavioral data, with the aim to understand the relationships between data sets. Common approaches to data integration utilize transformations that maximize covariance or correlation, but measures of information using higher order moments may reveal additional structure. We introduce Joint and Individual Non-Gaussian component analysis (JIN) for data integration. We focus on information shared in subject score subspaces estimated using non-quadratic nonlinearities, and we also examine information unique to each data set. We apply our method to data from the Human Connectome Project.

E1208: **Latent time joint mixed effect models**
Presenter: **Michael Donohue**, University of Southern California, United States
Co-authors: Dan Li, Samuel Iddi, Wesley Thompson
Characterization of long-term disease dynamics, from disease-free to end-stage, is integral to understanding the course of neurodegenerative diseases such as Parkinson's and Alzheimer's, and ultimately, how best to intervene. Natural history studies typically recruit multiple cohorts at different stages of disease and follow them longitudinally for a relatively short period of time. We propose a latent time joint mixed effects model to characterize long-term disease dynamics using this short-term data. Markov chain Monte Carlo methods are proposed for estimation, model selection, and inference. We apply the model to data from the Alzheimers disease neuroimaging initiative, and discuss applications to Alzheimer's prognosis and clinical trial simulations.

EO687 Room B1 GRAPHICAL MARKOV MODELS III Chair: Nanny Wermuth

E1439: **Direct and indirect effect for a class of discrete regression graph models**
Presenter: **Monia Lupparelli**, University of Bologna, Italy
In linear regression modelling the distortion of effects after marginalizing over variables of the conditioning set has been widely studied in several contexts. For Gaussian variables, the relationship between marginal and partial regression coefficients is well-established. Possible generalizations beyond the linear Gaussian case have been developed, nevertheless the case of discrete variables is still challenging, in particular in medical and social science settings. A multivariate regression framework is proposed for binary data with regression coefficients given by the logarithm of relative risks and a multivariate relative risk formula is derived to define the relationship between marginal and conditional relative risks. The method is illustrated through the analysis of the morphine data in order to assess the effect of preoperative oral morphine administration on the postoperative pain relief.

E1466: **A Bayesian approach to coloured graphical Gaussian models**
Presenter: **Helene Massam**, York University, Canada
Co-authors: Qiong Li, Xin Xin Gao
The focus is on graphical Gaussian models $N_p(0,\Sigma)$ Markov with respect to an undirected graph G, with additional symmetry constraints on the entries of the precision matrix $K = \Sigma^{-1}$. We give an overview of recent results for estimation and model selection in this class of models: the Diaconis-Ylvisaker conjugate prior, called the coloured G-Wishart, a Bayesian estimate of Σ and K, its asymptotic behaviour when p is fixed and the number n of sample points tends to infinity or, when both p and n tend to infinity, and also an efficient double reversible jump Markov chain Monte Carlo algorithm for estimating Bayes factors in model selection.

E1622: **On maximum likelihood estimation for mean zero versus general Ising graphical Markov models**
Presenter: **Giovanni Maria Marchetti**, University of Florence, Italy
Co-authors: Nanny Wermuth
The properties of the class of mean zero Ising models in fitting graphical Markov models to binary data are summarized. Moreover, we address

parameter estimation by maximum likelihood with a comparison to the larger class of general Ising models. We discuss how to simplify estimation using mean zero Ising models when the marginal distributions of data have skewed margins.

E1680: **Bayesian diagnostics for chain event graphs**
Presenter: **Rachel Wilkerson**, University of Warwick, United Kingdom
Co-authors: Jim Smith
The class of chain event graphs has now been established as a practical Bayesian graphical tool for modeling a variety of processes. However, although a number of techniques for estimating this and performing model selection on this class have now been developed no bespoke methods of diagnostically checking representatives within this family have been yet developed. We rectify this situation and provide a number of new Bayesian diagnostics that parallel those available for the more restrictive class of Bayesian network models. These are designed to check the continued validity of the selected model as data about a population continues to be collected.

EO631 Room C1 STATISTICS IN COSMOLOGY Chair: Armin Schwartzman

E1277: **Minkowski functionals: Constraining cosmology with non Gaussianity**
Presenter: **Anne Ducout**, University of Tokyo, Japan
Non Gaussianity (NG) i.e any deviation of a random field from the purely Gaussian distribution is a key observable in cosmology. It can be either primordial in origin, describing the statistics of the primordial phases of the universe and observed through the distribution of the Cosmic Microwave Background (CMB) anisotropies. It can also be the result of non linear gravitational processes and can then appear in the structures of matter at large scales, being observed in lensed galaxies surveys. To measure these potential NG features in cosmological datasets, numerous methods have been developed, optimised and used, and the focus will be on one in particular, the Minkowski Functionals (MFs). MFs describe the morphology of random fields – topology and geometry – and are sensitive to any NG, at any order. Being generic and easy to implement, MFs have been applied to the Planck CMB data and could help to constrain the inflation paradigm and gave the best constraint on NG sourced by cosmic strings arising in supersymmetric and grand unified theories. Used on future weak lensing galaxy surveys (cosmic shear) they will help to break degeneracies between parameters to give stronger constraints on dark energy.

E1446: **Multiple testing of local maxima for detection of peaks on the (celestial) sphere**
Presenter: **Armin Schwartzman**, University of California, San Diego, United States
Co-authors: Dan Cheng, Valentina Cammarota, Yabebal Fantaye, Domenico Marinucci
A topological multiple testing scheme for detecting peaks on the sphere under isotropic Gaussian noise is presented. The tests are performed at local maxima of the observed field filtered by the spherical needlet transform. Motivated by point-source detection in cosmic Microwave Background radiation (CMB) data, we focus on cases where a single realization of a smooth isotropic Gaussian random field on the sphere is observed, and a number of well-localized signals are superimposed on such background field. The proposed algorithms, combined with the Benjamini-Hochberg procedure for thresholding p-values, provide asymptotic control of the False Discovery Rate (FDR) and power consistency as the signal strength and the frequency of the needlet transform get large.

E1463: **Unexpected topology of the cosmic microwave background**
Presenter: **Pratyush Pranav**, Ecole Normale Superieure de Lyon, France
The aim is to study the topology generated by the temperature fluctuations of the Cosmic Microwave Background (CMB) radiation, as quantified by the number of components and holes in the growing excursion sets. We compare CMB maps observed by the Planck satellite with a thousand simulated maps generated according to the LCDM paradigm with Gaussian distributed fluctuations. The comparison is multi-scale, being performed on a sequence of degraded maps with mean pixel separation ranging from 0.05 to 7.33 degrees. The parametric χ^2-test shows differences between observations and simulations, yielding p-values at per-cent to less than per-mil levels roughly between 2 and 7 degrees, with the difference in the number of components and holes peaking at more than 3σ sporadically at these scales. There are reports of mildly unusual behaviour of the Euler characteristic at 3.66 degrees, which is phenomenologically related to the strongly anomalous behaviour of components and holes. It is also the scale at which the observed maps exhibit low variance compared to the simulations, and approximately the range of scales at which the power spectrum exhibits a dip with respect to the theoretical model. Non-parametric tests show even stronger differences at almost all scales. The results motivate a closer look at the standard cosmological paradigm, including primordial non-Gaussianity.

E1402: **Application of the second order Gaussian kinematic formula to CMB data analysis**
Presenter: **Yabebal Fantaye**, African Institute for Mathematical Sciences, South Africa
Co-authors: Domenico Marinucci
Recent advances in mathematical literature have established an analytical result for the second order Gaussian kinematic formula. Using this result it is shown that the variance of Minkowski Functionals (MFs) becomes smaller when the higher order terms are subtracted out. We will show that under ideal setup, on full sky observation, second order subtracted MFs are very sensitive to small deviations from Gaussianity. All our results are validated by numerical experiments that show a good agreement between theoretical predictions and Monte Carlo simulations.

EO028 Room D1 RECENT ADVANCES IN BAYESIAN APPROACHES FOR CAUSAL INFERENCE Chair: Michael Daniels

E0473: **Bayesian regression tree models for causal inference: Regularization, confounding and heterogeneity**
Presenter: **Richard Hahn**, Arizona State University, United States
A semi-parametric Bayesian regression model is described for estimating heterogeneous treatment effects from observational data. Standard nonlinear regression models, which may work quite well for prediction, can yield badly biased estimates of treatment effects when fit to data with strong confounding. Our Bayesian causal forests model avoids this problem by directly incorporating an estimate of the propensity function in the specification of the response model, implicitly inducing a covariate-dependent prior on the regression function. This new parametrization also allows treatment heterogeneity to be regularized separately from the prognostic effect of control variables, making it possible to informatively shrink to homogeneity, in contrast to existing Bayesian non- and semi-parametric approaches.

E0581: **Bayesian non-parametric G-computation in the presence of non-ignorable dropout and death**
Presenter: **Maria Josefsson**, Centre for Demographic and Ageing Research, Sweden
Co-authors: Michael Daniels
Causal inference with observational longitudinal data and time-varying exposures is often complicated by time-dependent confounding and attrition. G-computation is one method used for estimating a causal effect when time-varying confounding is present. The parametric modeling approach most often used in practice relies on strong modeling assumptions for valid inference, and moreover depends on an assumption of missing at random, which is generally invalid when the missingness is non-ignorable or due to death. We develop a flexible Bayesian non-parametric G-computation approach for assessing the causal effect on the subpopulation that would survive irrespective of exposure, in a setting with non-ignorable dropout. The approach is to specify models for the observed data using Bayesian additive regression trees, and then use assumptions with embedded

sensitivity parameters to identify and estimate the causal effect. The proposed approach is motivated by a longitudinal cohort study on cognition, health, and ageing. We apply our approach to study the effect of becoming a widow on memory.

E0635: **Assessing causal effects in the presence of treatment switching through principal stratification**
Presenter: **Fabrizia Mealli**, University of Florence, Italy

Clinical trials, focusing on survival outcomes for patients suffering from AIDS-related illnesses and particularly painful cancers, often allows patients in the control arm to switch to the treatment arm if their physical conditions are worse than certain tolerance levels. The ITT analysis provides valid estimates of the effect of assignment, it does not give information about the effect of the actual receipt of the treatment. Other existing methods propose to reconstruct the outcome a unit would have had if s/he had not switched. But these methods usually rely on strong assumptions. We propose to redefine the problem of treatment switching using principal stratification and introduce new causal estimands, principal causal effects for patients belonging to subpopulations defined by the switching behavior under control. For inference, we use a Bayesian approach to properly take into account that (i) switching happens in continuous time generating a continuum of principal strata; (ii) switching time is not defined for units who never switch in a particular experiment; and (iii) survival time, the outcome of primary interest, and switching time are subject to censoring. We illustrate our framework using simulated data based on the Concorde study, a randomized trial aimed to assess causal effects on time-to-disease progression or death of immediate versus deferred treatment with zidovudine for HIV patients.

E1036: **Common support diagnostic for heterogeneous treatment effect**
Presenter: **Jennifer Hill**, New York University, United States

Robust estimation of average causal effects requires strong assumptions that are often untestable. The common support assumption is testable in theory however in with high dimensional confounders this can become computationally challenging. When the goal is estimation of subgroup effects or individual-level treatment effects identifying sufficient common support is even more difficult. Bayesian nonparametric strategies for identifying common causal support for heterogeneous treatment effects will be presented. These have the advantage of more robust estimation of targeted treatment effects (for subgroups and individuals) and superior uncertainty quantification as compared to common frequentist alternative. Moreover, these approaches have the advantage of a natural and effective strategy for identifying observations and subgroups for whom common causal support is likely to be satisfied.

EO526 Room E1 ANALYSIS OF LARGE AND COMPLEX DATA Chair: Johannes Lederer

E1147: **Robust machine learning via median-of-means**
Presenter: **Guillaume Lecue**, CNRS and ENSAE, France
Co-authors: Matthieu Lerasle

Median-of-means (MOM) based procedures have been recently introduced in learning theory. These estimators outperform classical least-squares estimators when data are heavy-tailed and/or are corrupted. None of these procedures can be implemented, which is the major issue of current MOM procedures. We introduce minmax MOM estimators and show that they achieve the same sub-Gaussian deviation bounds as the alternatives, both in small and high-dimensional least-squares regression. In particular, these estimators are efficient under moments assumptions on data that may have been corrupted by a few outliers. Besides these theoretical guarantees, the definition of minmax MOM estimators suggests simple and systematic modifications of standard algorithms used to approximate least-squares estimators and their regularized versions. As a proof of concept, we perform an extensive simulation study of these algorithms for robust versions of the lasso.

E0609: **Manifold learning using kernel density estimation and local principal component analysis**
Presenter: **Hariharan Narayanan**, TIFR, India
Co-authors: Kitty Mohammed

The problem of recovering a d dimensional manifold M when provided with noiseless samples from M is considered. There are many algorithms (e.g., Isomap) that are used in practice to fit manifolds and thus reduce the dimensionality of a given data set. Ideally, the estimate $M put$ of M should be an actual manifold of a certain smoothness; furthermore, $M put$ should be arbitrarily close to M in Hausdorff distance given a large enough sample. Generally speaking, existing manifold learning algorithms do not meet these criteria. An algorithm has been previously developed whose output is provably a manifold. The key idea is to define an approximate squared-distance function (asdf) to M. Then, $M put$ is given by the set of points where the gradient of the asdf is orthogonal to the subspace spanned by the largest $n-d$ eigenvectors of the Hessian of the asdf. As long as the asdf meets certain regularity conditions, $M put$ is a manifold that is arbitrarily close in Hausdorff distance to M. Two asdfs are defined which can be calculated from the data. It is shown that they meet the required regularity conditions. The first asdf is based on kernel density estimation, and the second is based on estimation of tangent spaces using local principal component analysis.

E1530: **Fridge: Focused fine-tuning of ridge regression for personalized predictions**
Presenter: **Kristoffer Hellton**, University of Oslo, Norway

Penalized regression methods, depending on one or more tuning parameters, require fine-tuning to achieve optimal prediction performance. For ridge regression, there exist numerous approaches with cross-validation as the standard procedure, but common for all is that one single parameter is chosen for all future predictions. To better adapt to heterogeneity in high-dimensional data, we propose a focused ridge regression, the fridge procedure, with a unique tuning parameter for each covariate vector for which we wish to produce a prediction. The covariate vector specific tuning parameter is defined as the minimizer of the theoretical mean square prediction error, which is explicitly given in case of ridge regression. We propose to estimate the resulting tuning parameter through a plugin approach, and for high-dimensional data, ridge regression with cross-validation is used as the plugin estimate. The procedure is extended to logistic ridge regression by utilizing parametric bootstrap. Simulations show that fridge gives lower average prediction error than standard ridge regression in heterogeneous data, and we illustrate the method in an application of personalized medicine, predicting individual disease risk based on gene expression data.

E1175: **Fresh ideas for tuning parameter calibration**
Presenter: **Johannes Lederer**, University of Washington, United States

Large and high-dimensional data has become a major source of knowledge in economics, biology, astronomy, and many other fields. However, lasso and other standard methods for such data depend on tuning parameters that are difficult to calibrate. We introduce novel approaches to this calibration and demonstrate their features in theory, computations, and applications.

EO490 Room F1 RECENT ADVANCES IN COMPUTATION FOR STATISTICAL MACHINE LEARNING Chair: Irina Gaynanova

E0221: **Optimization in high dimensional additive models**
Presenter: **Noah Simon**, UW Biostatistics, United States
A general framework for sparse additive regression is discussed. We allow the class of each additive component to be quite general (characterized by semi-norm smoothness this includes monotonicity of derivatives, variation/Sobolev smoothness). We show that by minimizing a simple convex problem, we can estimate these functions at the minimax rate (over functions in that smooth additive class). In addition we show that the penalized regression problem can be efficiently solved using a proximal gradient descent algorithm: Each prox-step decouples into p-univariate penalized regression problems; each of these univariate penalized problems can in turn be written as a simple update of the solution to a univariate non-parametric regression problem (for which we often have efficient algorithms). In addition, we characterize the statistical performance of the output of our algorithm after a finite number of steps.

E0522: **An explicit mean-covariance parameterization for multivariate response linear regression**
Presenter: **Aaron Molstad**, Fred Hutchinson Cancer Research Center, United States
Co-authors: Adam Rothman, Charles Doss, Guangwei Weng
A new method is developed to fit the multivariate response linear regression model that exploits a parametric link between the regression coefficient matrix and the error covariance matrix. Specifically, we assume that the correlations between entries in the multivariate error random vector are proportional to the cosines of the angles between their corresponding regression coefficient matrix columns, so as the angle between two regression coefficient matrix columns decreases, the correlation between the corresponding errors increases. This assumption can also be motivated through an error-in-variables formulation. Motivated by this parameterization, we propose a class of estimators that minimizes a non-convex loss plus penalty. The optimization problem is solved with an accelerated proximal gradient descent algorithm. We show our method can outperform competitors in both real and simulated data examples.

E0632: **High-dimensional Gaussian graphical model for network-linked data**
Presenter: **Tianxi Li**, University of Virginia, United States
Co-authors: Cheng Qian, Liza Levina, Ji Zhu
Graphical models are commonly used to represent conditional independence relationships between variables, and estimating them from high-dimensional data has been an active research area. However, almost all existing methods rely on the assumption that the observations share the same mean, and that they are independent. At the same time, datasets with observations connected by a network are becoming increasingly common, and tend to violate both these assumptions. We develop a Gaussian graphical model for settings where the observations are connected by a network and have potentially different mean vectors, varying smoothly over the network. We propose an efficient estimation method for this model and demonstrate its effectiveness in both simulated and real data, obtaining meaningful interpretable results on a statistician's coauthorship network. We also prove that our method estimates both the inverse covariance matrix and the corresponding graph structure correctly under the assumption of network cohesion, which refers to the empirically observed phenomenon of network neighbors sharing similar traits.

E0888: **A convex optimization formulation for multivariate regression**
Presenter: **Yunzhang Zhu**, Ohio State University, United States
Multivariate regression (or multi-task learning) concerns the task of predicting the value of multiple responses from a set of covariates. We will present a convex optimization formulation for high-dimensional multivariate linear regression under general error covariance structure. The main difficulty for simultaneous estimation of the regression coefficients and the error covariance lies in the fact that the negative log-likelihood function is not jointly convex. To overcome this difficulty, a new parameterization is proposed, under which the negative log-likelihood function is convex. It will be demonstrated that the new parameterization is particularly useful for covariate-adjusted graphical modeling. The proposed method compare favorably to existing high dimensional multivariate linear regression methodologies that are based either on minimizing non-convex criteria or certain two-step procedures. Finally, we present some theoretical properties and applications to gene network analysis.

EO156 Room G1 RECENT DEVELOPMENTS IN STATISTICAL MODELS FOR SURVIVAL DATA Chair: Marialuisa Restaino

E0639: **Using the Dagum distribution in survival regression models**
Presenter: **Mariangela Zenga**, Universita degli Studi di Milano-Bicocca -DISMEQ, Italy
Co-authors: Filippo Domma, Juan Eloy Ruiz-Castro
The Dagum distribution is a special case of the Generalized Beta distribution with three parameters. In the last years, it has been introduced in the field of the survival analysis and the reliability. It fact, it has been demonstrated that the hazard rate of distribution is very flexible: according to the parameters values, the hazard rate has a decreasing, or an upside-down bathtub, or bathtub and then upside-down bathtub failure rate. We will consider the Dagum distribution in survival regression models: assuming right censored data, we will consider the maximum likelihood inference for analysis and a graphical method to test the goodness of fit for residuals.

E0807: **Statistical boosting for time-to-event data: An overview on recent developments**
Presenter: **Andreas Mayr**, University of Bonn, Germany
Co-authors: Matthias Schmid, Elisabeth Waldmann
Statistical boosting algorithms combine a powerful machine learning approach with classical statistical modelling, offering various practical advantages like automated variable selection and implicit regularization of effect estimates. In the context of time-to-event data, the general boosting concept was also extended beyond the classical Cox framework. For example, we developed a boosting approach to directly estimate prediction models that are optimal with respect to the concordance index (Harrell's C). Another recent approach focuses on joint models for longitudinal and time-to-event data. The aim is to provide a short introduction to the concept of boosting and its application for the analysis of survival data, highlighting both advantages and limitations for practical data analysis.

E1009: **Survival models for highly clustered censored data: Accurate inference based on integrated likelihoods**
Presenter: **Giuliana Cortese**, University of Padua, Italy
Co-authors: Nicola Sartori
Clustering structures are frequently encountered in censored time-to-event-data. Often the main interest is not in the cluster-related parameters, which are then treated as nuisance. When inference is on a parameter of interest in presence of many nuisance parameters, standard likelihood methods perform very poorly and may lead to severe bias. This problem is particularly evident in survival models where the number of clusters is high compared to the within-cluster size. We consider clustered failure time data under independent censoring and propose inference based on integrated likelihoods. This approach provides very accurate inference in presence of many nuisance parameters or small sample sizes. The regression models of interest can be parametric or semiparametric survival models. We show some applications of the proposed method in different types of regression approaches. A data example about late-stage HIV-infected patients is used to compare the new approach with the existing alternatives, such as frailty models and stratified Cox's models. Simulation studies show that appropriately defined integrated likelihoods provide

very accurate inferential results in all circumstances, such as for highly clustered data or heavy censoring, even in extreme settings where standard likelihood procedures lead to strongly misleading results. We show that the proposed method performs as well as the frailty model and it is superior when the frailty distribution is misspecified.

E1123: **Random forest under random censoring applied to the prediction of the duration of an insurance contract**
Presenter: **Olivier Lopez**, Universita Pierre et Marie Curie Paris 6, France

In the insurance broker market, commissions received by brokers are closely related to so-called "customer value": the longer a policyholder keeps their contract, the more profit there is for the company and therefore the broker. Hence, predicting the time at which a potential policyholder will surrender their contract is essential in order to optimize a commercial process and define a prospect scoring. We propose a weighted random forest model to address this problem. Our model is designed to compensate for the impact of random censoring. We investigate different types of assumptions on the censoring, studying both the cases where it is independent or not from the covariates. We compare our approach with other standard methods which apply in our setting, using simulated and real data analysis. We show that our approach is very competitive in terms of quadratic error in addressing the given problem.

EO170 Room H1 FLEXIBLE MODELS AND METHODS FOR CATEGORICAL DATA Chair: Rosaria Simone

E0206: **Analyzing large matrices of ordinal data**
Presenter: **Margot Selosse**, Universite de Lyon, France
Co-authors: Julien Jacques, Christophe Biernacki

A co-clustering strategy for analyzing large matrix of ordinal data is presented. For this, a model-based co-clustering algorithm for ordinal data is proposed. This algorithm relies on the latent block model embedding a probability distribution specific to ordinal data (the so-called BOS or Binary Ordinal Search distribution). Model inference relies on a stochastic EM algorithm coupled with a Gibbs sampler, and the ICL-BIC criterion is used for selecting the number of co-clusters (or blocks). The main advantage of this ordinal dedicated co-clustering model is its parsimony, the interpretability of the co-cluster parameters (mode, precision) and the possibility to take into account missing data. The usefulness of the method is illustrated by analyzing a psychological survey on women affected by a breast tumor.

E0464: **Modelling perceived choice variety by a mixture model for rating data**
Presenter: **Marica Manisera**, University of Brescia, Italy
Co-authors: Paola Zuccolotto, Eugenio Brentari

In consumer research, marketing, public policy and other fields, individuals choice depends on the number of possible alternatives. In addition, according to the literature, the choice satisfaction is influenced not only by the number of options but also by the perceived variety. The aim is to apply a novel statistical approach to model perceived variety, in order to better understand the perceptions of individuals about the variety of the possible choice options. We resort to the class of CUB (Combination of Uniform and Binomial random variables) models, in particular to the Nonlinear extension of CUB, in order to (i) provide a measure for perceived variety, (ii) add a measure of uncertainty, (iii) give insights on the state of mind of respondents toward the response scale. The application of the Nonlinear CUB to real data shows interesting results.

E0506: **Hierarchical models for rater agreement and the evergreen kappa statistic**
Presenter: **Mauro Gasparini**, Politecnico di Torino, Italy

The kappa statistic as a measure of rater agreement will turn 60 soon and yet some of its properties are still attracting a lot of interest among statisticians. In particular, the appropriate modeling of rater agreement at the population level and the use of kappa for inferential purposes have been studied by several authors with different approaches. Some of the asymptotic distributions of kappa, an approach based on hierarchical models and their use in clinical and psychological research will be presented.

E0814: **A family of models for multivariate rating scale data accounting for response styles**
Presenter: **Sabrina Giordano**, University of Calabria, Italy

Rating scales have been widely used to investigate attitudes and opinions in sociological and psychological contexts. However, observed ratings may not represent the true opinion. In fact, it is generally observed that unaware respondents may use only a few of the given options irrespective of their real opinion. Someone has a strong tendency to mark the endpoints, others take shelter in the middle category, someone else responds with agreement(disagreement) regardless of item content, optimists (pessimists) may prefer the positive (negative) side of the scale. These behaviours are typically considered as response styles. A recently proposed family of models for multivariate rating scale data distinguishes aware from unaware responses. It directly and explicitly models the response styles by specifying the distribution of the unaware responses. We present an alternative approach defining the unaware distribution as a transformation of the distribution of the aware responses. To this aim, we use a known adjacent categories model accounting for response styles. Moreover, we aim also at investigating the influence of covariates on response styles. Competing models are compared by the Vuong test for misspecified non-nested models as the distributions that model response styles are only an approximation of the true mechanism generating unaware responses.

EO164 Room I1 ADVANCES IN INFERENCE AND DISTRIBUTION THEORY Chair: Inmaculada Barranco-Chamorro

E0595: **Confidence regions for Pareto parameters from single and independent samples**
Presenter: **Udo Kamps**, RWTH Aachen University, Germany

Based on a single and on two independent samples, joint confidence regions for parameters of Pareto distributions are proposed with minimum volume properties and without assigning the confidence level to dimensions. In the one-sample case, comparisons are made to former simultaneous confidence sets for Pareto parameters by means of simulation and a real data set. The two-sample case is studied in various set-ups and comprises, e.g., a simultaneous confidence region for the scale parameters and a common shape parameter. Such confidence regions can also be developed for type-II right censored data.

E0742: **Multi-output conditional inference trees applied to the electricity market: Variable importance analysis**
Presenter: **Ismael Ahrazem Dfuf**, Technical University of Madrid, Spain
Co-authors: Jose Mira, Camino Gonzalez

Random forests algorithm has been applied extensively due to its high prediction accuracy, interpretability, ability to deal with high dimensional data and to assess the relevance of highly correlated variables in complex non-linear models. We propose an alternative framework to assess the variable importance in multivariate response scenarios based on the permutation importance method using the conditional inference trees algorithm. To build the solution, a ϕ-divergence measure from information theory is used. The main goal of divergence measures is to provide a distance between probability distributions, in our case, the observations and predicted values. The solution was tested in simulated examples and also in a real case, where we assessed and ranked the most relevant predictors for price and demand of electricity jointly. The results show that the new method outperforms in most cases the outcomes achieved by the recently proposed variable importance technique, Intervention Prediction Measure.

E1070: **Methodological advances in slash distributions**
Presenter: **Inmaculada Barranco-Chamorro**, Fundacion de investigacion de la Universidad de Sevilla (FIUS), Spain

In real-world data, it is quite common to find symmetrical and unimodal histograms with heavy tails that do not fit well to a normal distribution. Slash models are a good option to deal with this kind of situations, in which departures of Gaussianity are a serious problem for the data analyst. This is one the main reasons why slash distributions have received a great deal of attention during the last decades. Symmetrical and unimodal slash models are discussed. These models are obtained as the quotient of two independent continuous random variables. Slash distributions are models with heavy tails, and three parameters (location, scale and kurtosis). The emphasis is on the estimation of kurtosis parameter. It is shown that traditional inference methods of estimation fail in slash models. Some improvements are proposed. A practical application to real data, of interest in economics, is included.

E1499: **Analytical results for distributions of condition numbers from dual noncentral Wishart type matrices**
Presenter: **JT Ferreira**, University of Pretoria, South Africa
Co-authors: Andriette Bekker, Mohammad Arashi

The analytical characterisation of MIMO (multiple input multiple output) statistics is investigated; in particular that of the condition number. The channel propagation matrix is assumed to be complex matrix variate elliptically distributed: this assumption allows the analysis of the condition number in broad generality. The complex matrix variate elliptical class includes the complex matrix variate normal- and t distributions as special cases. Specifically, the probability density function and cumulative distribution function of the condition number of a dual noncentral Wishart type matrix with arbitrary degrees of freedom is derived and studied for the complex noncentral matrix variate t case, and comparatively investigated versus the well-studied normal model. This dual setting is of interest stemming from a practical consideration; viz. dual-branch systems which are equipped with two transmit- and receive antennas. A numerical- and comparative study illustrates and motivates the analytical expressions.

EO268 Room L1 ON SOME RECENT RESULTS IN SUPERVISED AND UNSUPERVISED CLASSIFICATION II Chair: Geoffrey McLachlan

E1143: **Model-based tools for the analysis of flow and mass cytometric data**
Presenter: **Sharon Lee**, University of Queensland, Australia

Cytometry plays an important role in clinical diagnosis and monitoring of lymphoma and leukaemia. However, analysis of modern cytometric data is challenging due to their high dimensionality, large number of observations, as well as complex distributional features such as multimodality, asymmetry, and other non-normal characteristics. Firstly, a mixture model-based tool is presented to automatically segment and perform dimension reduction of high-dimensional cytometry data. Secondly, the tasks of unsupervised clustering and supervised classification of multiple heterogeneous cytometric samples are presented. We adopt a linear mixed model approach to handle inter-sample variations, and flexible component densities to cater for non-normal cluster shapes. The usefulness and effectiveness of these model-based tools are demonstrated using a number of real data from flow and mass cytometry experiments.

E0408: **Overlapping mixture models for network data (manet) with covariates adjustment**
Presenter: **Saverio Ranciati**, Universita di Bologna, Italy
Co-authors: Giuliano Galimberti, Veronica Vinciotti, Ernst Wit

Network data often come in the form of actor-event information, where two types of nodes comprise the very fabric of the network. Examples of such networks are: people voting in an election, users liking/disliking media content, or, more generally, individuals - actors - attending events. Interest lies in discovering communities among these actors, based on their patterns of attendance to the considered events. To achieve this goal, we propose an extension of a previous model: covariates are injected into the model, leveraging on parsimony for the parameters and giving insights about the influence of such characteristics on the attendances. We assess the performance of our approach in a simulated environment.

E1278: **Model-based clustering in very high dimensions via adaptive projections**
Presenter: **Bernd Taschler**, German Center for Neurodegenerative Diseases, Germany
Co-authors: Frank Dondelinger, Sach Mukherjee

Model-based clustering is considered in high-dimensional settings where the dimension p is large relative to sample size n and where either or both of means and covariance structures may differ between the latent groups. We propose an approach called *Model-based Clustering via Adaptive Projections* or *MCAP*. Instead of estimating mixtures in the original space, we work in a low-dimensional space obtained by linear projection. The projection dimension plays an important role and governs a type of bias-variance trade-off. MCAP sets the projection dimension automatically in a data-adaptive manner. The mixture modelling itself is done using a full covariance formulation and this, combined with the adaptive projection, allows detection of both mean and covariance signals in very high dimensional problems. We show real-data examples in which covariance signals are reliably detected in problems with $p \sim 10^4$ or more, and examples where MCAP maintains performance even when the mean signal is entirely removed, leaving differential covariance structure in the high-dimensional space as the only signal. Across a number of regimes, MCAP performs as well or better than a range of existing methods, including a recently-proposed ℓ_1-penalized approach, and performance remains broadly stable with increasing dimension, at low computational cost.

E1494: **Classification using distance nearest neighbours with adjusted pseudolikelihood**
Presenter: **Lida Fallah**, University College Dublin, Ireland
Co-authors: Nial Friel

The distance nearest neighbour (DNN) model, offers a probabilistic classification algorithm, modelling the joint distribution of the training and test data as a Markov random field. The essence of the DNN model is to account for the distance, in some sense, between feature vectors so that two feature vectors which are closer to each other are more likely to share the same class label. However, in its original formulation, it is computationally expensive due to an intractability of the likelihood. The pseudolikelihood offers a tractable alternative, however it is well understood that this can result in biased parameter estimation. To address this we consider a transformed pseudolikelihood approximation so that its mode and curvature (at the mode) coincide with that of the intractable likelihood. An additional advantage of our approach is it offers the possibility to carry out feature selection using Bayesian model selection.

EO558 Room M1 STATISTICAL METHODOLOGIES WITH COMPLEX INFORMATION Chair: Maria Brigida Ferraro

E0643: **Using interval-wise testing to investigate high-resolution "omics" data at multiple locations and scales**
Presenter: **Marzia Cremona**, The Pennsylvania State University, United States
Co-authors: Francesca Chiaromonte, Kateryna Makova

Interval-Wise Testing (IWT) is a non-parametric inferential procedure for functional data that exploits the ordered nature of measurements to adjust *P*-values and control the interval-wise error rate. We present an extended version of IWT implemented in the R/Bioconductor package IWTomics designed for complex, high-resolution "Omics" data. IWTomics allows one to compare several"Omics" features over groups of genomic regions at multiple locations and scales, and outputs those at which each feature shows significant effects. We describe three collaborative projects in which IWTomics has been successfully employed to "Omits" studies. In the first, we analyze the profiles of a large collection of genomic landscape features in human and mouse, to identify features that influence integration and fixation of endogenous retroviruses (ERVs). In a similar framework, the second project investigates the dynamic landscape of long interspersed elements-1 (L_1) transposition in the human genome. Finally, the third project investigates how non-canonical 3D-conformations in the genome affect the local speed of DNA polymerization, and how this is associated to sequencing errors and to mutations in living cells.

E0673: **Alzheimer's disease risk prediction with multidimensional biomarkers**
Presenter: **Zheyu Wang**, Johns Hopkins University, United States

The initiation of the Alzheimer's disease (AD) pathogenic process is typically unobserved and has been thought to precede the first symptoms by 20 years or more. This imposes a major challenge in investigating risk factors for early AD detection, because using clinical diagnosis as the reference point can be inaccurate, especially in the early course of the disease. Until technology advance allows for brain examination with autopsy level clarity in vivo, an appropriate statistical method that directly addresses the unobservable nature of preclinical AD progression is necessary for any rigorous AD biomarker evaluation and for efficient analyzing AD study data where only clinical data are available and neuropathology data are not yet available. Moreover, AD progression is recognized as a multidimensional cascade, biomarkers reflect different aspects along AD pathogensis pathway should be examined jointly and in the context of normal aging process.

E0697: **Reconstruction of functional fragments**
Presenter: **Marco Stefanucci**, University of Rome - Sapienza, Italy
Co-authors: David Kraus

Often in the framework of functional data analysis the curves might be observed only on a subset of the whole domain, invalidating most of the existing statistical procedures. We propose a method able to handle this kind of data, in particular in the context of curve reconstruction, where the aim is to predict the unobserved part of each curve. More specifically, we study the role of regularization in the estimation of the reconstruction operator. The performances of different type of regularization are presented and the main differences with the standard estimator are discussed. The relevance of the proposed methodology is further pointed out through an application to a real dataset.

E1017: **M-quantile regression for multivariate longitudinal data**
Presenter: **Maria Francesca Marino**, University of Florence, Italy
Co-authors: Marco Alfo, Maria Giovanna Ranalli, Nicola Salvati, Nikos Tzavidis

Recently, there has been an increasing interest in the analysis of longitudinal data via quantile and M-quantile regression with the aim of studying the effect of observed covariates at different levels of the response distribution. When compared to mean regression, such approaches offer a more complete picture of the response of interest. We propose a multivariate finite mixture of M-quantile regression models to deal with multivariate longitudinal responses. Discrete, individual-specific, random parameters are used to account for both dependence within the same response recorded at different time occasions and association between multiple responses observed on the same unit at a given time. The distribution of the random parameters is left unspecified and is directly estimated from the observed data. Furthermore, to account for potential endogeneity of observed covariates, we propose the definition of an auxiliary regression model. Within and between effects of time-varying covariates on the M-quantiles of the response distribution are separately modeled to avoid bias. An extended EM algorithm is proposed to derive parameter estimates under a maximum likelihood approach. The model is applied to the analysis of data from the millennium cohort study on children internalizing and externalizing disorders.

EO234 Room N1 STATISTICS OF ENVIRONMENTAL EXTREMES Chair: Raphael Huser

E0278: **A hierarchical max-infinitely divisible process for spatial precipitation modeling**
Presenter: **Gregory Bopp**, Pennsylvania State University, United States
Co-authors: Raphael Huser, Benjamin Shaby

The hazards of atmospheric phenomena such as extreme precipitation are often largely determined by their spatial extent. Different choices of extremal dependence classes can lead to vastly different conclusions about the risk of such hazards. We will present a class of models for spatial extremes that allows for a smooth transition between extremal dependence types. The conditional representation of the proposed model allows for full-likelihood based inference via Markov chain Monte Carlo that scales to large datasets. The model extends a familiar max-stable class to a broader family of max-infinitely divisible processes that allows for more flexible spatial dependence types. Due to a construction in terms of flexible random basis functions that are estimated from the data, straightforward inspection of the predominant spatial patterns of extremes is also possible. The proposed model is applied to extreme precipitation to examine flood risk over hydrologically defined watersheds in eastern North America.

E0682: **Inference for extremal-*t* and skew-*t* max-stable models in high dimensions**
Presenter: **Boris Beranger**, University of New South Wales, Australia
Co-authors: Scott Sisson, Alec Stephenson

Environmental phenomena are spatial processes by nature as a single extreme event (heat waves, floods, storms, etc.) often has repercussions at multiple locations. For risk management purposes it is important to have a good understanding of the dependence structure that links such events in order to make predictions on future phenomena, that can have a major impact on real life. Moreover, available data at different sites can exhibit asymmetric distributions proving the necessity for max-stable processes that can handle skewness. The extremal-*t* and skew-*t* processes possess such flexible dependence structure between extremes and inference for these max-stable models can be performed via composite likelihood based methods. However, the computational demands remains a burden in the scenario where the processes are observed at a large number of spatial locations. Assuming the time of occurrence of maxima known, the efficiency of moderately large order composite likelihood estimates is compared to those of the full likelihood approach when high dimensional information is available. Finally, an illustration using maximum temperatures in the region of Melbourne, Australia, is provided.

E1013: **Are extreme rainfalls in northeastern USA becoming more frequent, or bigger, or both?**
Presenter: **Holger Rootzen**, Chalmers, Sweden
Co-authors: Helga Olafsdottir, David Bolin
Records may become more extreme because the underlying distribution changes, or because one makes more tries, or both. We use NOAA data from 17 stations to study large rainstorms in northeastern USA. The stations were selected because they show an increasing trend in annual maxima. These trends did not appear in some nearby stations. Data on individual rainstorms provide the most direct path to understanding the development of rainstorms. However, annual maxima data are more widely available, and often of higher quality. We hence use both kinds of data, together with the close relation between the PoT method with generalized Pareto distributed excesses and the annual maxima method with the generalized extreme value distribution. This relation is closely related to Langbeins formula, which is widely used in hydrology to connect partial duration series with annual maxima. A (very) preliminary answer is that rainstorms are becoming more frequent, but that the distribution of the total amount of rain in individual rainstorm is not changed.

E1033: **Aggregation of extreme rainfall**
Presenter: **Anthony Davison**, EPFL, Switzerland
There is often a mismatch between observation of a process at individual sites and the need for inferences on aggregated output. Areal reduction factors, for example, transform estimates of extreme rainfall at a point to an estimate over a spatial domain, and are commonly used by hydrologists for flood risk estimation. The dependence structure of extreme rainfall plays a key role in this, since areal extremes may behave quite differently depending on whether the extreme rainfall is asymptotically dependent or independent. We shall discuss this problem and some of its ramifications, using simulated and real data.

EO576 Room O1 REGULARIZATION AND PARAMETER ESTIMATION IN ORDINARY DIFFERENTIAL EQUATIONS Chair: Nicolas Brunel

E0467: **Application of one-step method to parameter estimation in ODE models**
Presenter: **Itai Dattner**, University of Haifa, Israel
Co-authors: Shota Gugushvili
An application of Le Cam one-step method to parameter estimation in ordinary differential equation models is presented. This computationally simple technique can serve as an alternative to numerical evaluation of the popular non-linear least squares estimator, which typically requires the use of an iterative algorithm and repetitive numerical integration of the ordinary differential equation system. The one-step method starts from a preliminary $\sqrt{n}$-consistent estimator of the parameter of interest and next turns it into an asymptotic (in the sample size n) equivalent of the least squares estimator through a numerically straightforward procedure. We demonstrate performance of the one-step estimator via extensive simulations and real data examples. The method enables the researcher to obtain both point and interval estimates. The preliminary $\sqrt{n}$-consistent estimator that we use depends on non-parametric smoothing, and we provide a data-driven methodology for choosing its tuning parameter and support it by theory. An easy implementation scheme of the one-step method for practical use is pointed out.

E0569: **Regularization of parameter estimation in ordinary differential equations via discrete optimal control theory**
Presenter: **Quentin Clairon**, Newcastle University, United Kingdom
A parameter estimation method in Ordinary Differential Equation (ODE) models from partial, noisy observations is presented. Due to complex relationships between parameters and states, standard techniques such as nonlinear least squares can lead to presence of poorly identifiable parameters. Moreover, ODEs are generally approximations of the true process and influence of this misspecification on inference is often neglected. Control theories have been used to regularize the problem of parameter estimation in this context. In these methods, a perturbation is added to the ODE to facilitate data fitting and to represent model misspecifications. The estimation is done by solving a trade-off between data and model fidelity which leads to solve an optimal control problem. However, these approaches based on continuous control theory are computationally intensive and rely on a nonparametric state estimator known to be biased in sparse sample case. We construct a criterion based on discrete control theory. A computational efficient method which also bypasses the presmoothing step of signal estimation is developed. First, we expose how the estimation problem is turned into a control one and the numerical method used to solve it. Then, we derive the consistency with root-n convergence rate of our estimator in the well-specified case. Simulation in models with poorly identifiable parameters and misspecifications presence show our method gives accurate estimates.

E0744: **Learning large scale ordinary differential equation systems**
Presenter: **Niels Richard Hansen**, University of Copenhagen, Denmark
Co-authors: Frederik Vissing Mikkelsen
Learning large scale nonlinear ordinary differential equation (ODE) systems from data is known to be computationally and statistically challenging. We present a framework together with the adaptive integral matching (AIM) algorithm for learning polynomial or rational ODE systems with a sparse network structure. The framework allows for time course data sampled from multiple environments representing e.g. different interventions or perturbations of the system. The algorithm AIM combines an initial penalised integral matching step with an adapted least squares step based on solving the ODE numerically. The R package episode implements AIM together with several other algorithms and is available from CRAN. It is shown that AIM achieves state-of-the-art network recovery for the in silico phosphoprotein abundance data from the eighth DREAM challenge with an AUROC of 0.74, and it is demonstrated via a range of numerical examples that AIM has good statistical properties while being computationally feasible even for large systems.

E0815: **Optimisation and selection strategies for parameter estimation in ODE models with generalised smoothing**
Presenter: **Beatrice Laroche**, INRA, France
Co-authors: Nicolas Brunel, Daniel Goujot, Simon Labarthe
Let $(Y_{ik})_{i=1\ldots n;k=0\ldots K}$ be observations of the coordinates of a n-dimensional vector data collected at times $(t_{ik}) \in [0,T]$, modelled by a parametric ODE $\frac{dX}{dt} = f(X,\theta) + u$ and an observation equation $Y_{ik} = X_i(t_{ik}) + \eta_{ik}$, where X is the n-dimensional vector state of the model, $u = \frac{dX}{dt} - f(X,\theta)$ is the model error, $\eta_{ik} = Y_{ik} - X_i(t_{ik})$ are measurement errors and θ is the parameter. Methods from functional data analysis consist in replacing X by an approximation $\widehat{X}$ on a functional base (e.g. B-splines). Inspired by one of them, the Generalized Smoothing, the estimation problem can be formulated as the joint estimation of the parameters and the coefficients in the approximation basis (gathered for all coordinates of $\widehat{X}$ in the vector C) as $\min_{\theta,C}\left(\sum_{i,k}|Y_{ik} - \widehat{X}_i(t_{ik})|^2 + \frac{\lambda}{T}\int_0^T \|\frac{d\widehat{X}}{dt} - f(\widehat{X},\theta)\|^2 dt\right)$ where $\lambda \geq 0$ allows the trade-off between the measurement and model errors. The formulation, properties and numerical resolution of the estimation problem for a fixed λ will be discussed, as well as strategies for choosing λ. The influence of optimisation and hyperparameter selection will be tackled, both on synthetic and real data, on models of interactions networks in ecology or chemistry.

EO476 Room P1 STATISTICAL LEARNING AND ANALYSIS WITH COMPLEX FEATURED DATA Chair: Grace Yi

E0183: **Statistical inference with non-probability survey samples**
Presenter: **Changbao Wu**, University of Waterloo, Canada
A general framework for statistical inferences with non-probability survey samples is established. We develop a rigorous procedure for estimating the propensity scores for units in the non-probability sample, and construct robust estimators for finite population means. Variance estimation is discussed under the proposed framework. Results from simulation studies and real data analysis show the robustness and the efficiency of our proposed estimators as compared to existing methods.

E0634: **Perturbation-based model tests with application to the Clayton model**
Presenter: **Wenqing He**, University of Western Ontario, Canada
The perturbation resampling method can be employed to estimate the covariance matrix of an estimator when the estimator is obtained through minimizing a U-process. This perturbation resampling is proposed to establish general tests for the detection of model misspecification or for model checking. The proposed tests enjoy the simplicity and a theoretical justification. We apply the proposed method to modify previous tests for the assessment of Clayton models in multivariate survival analysis, where the asymptotic variance is intractable. The proposed tests present a promising performance in the simulation studies and have simpler procedures than the nonparametric bootstrap which can also be applied to approximate the covariance matrix. A colon cancer study further illustrates the proposed methods.

E1034: **Uncertainty quantification of treatment regime in precision medicine by confidence distributions**
Presenter: **Min-ge Xie**, Rutgers University, United States
Personalized decision rule in precision medicine is a 'discrete parameter, for which theoretical development of statistical inference is lacking. A new way to quantify the estimation uncertainty in a personalized decision based on confidence distribution (CD) is proposed. Suppose, in a regression setup, the optimal decision for treatment versus control for an individual z is determined by a linear decision rule $D = I(m_1(z) > m_0(z))$, where $m_1(z)$ and $m_0(z)$ are the expectations of potential outcomes of treatment and control, respectively. The estimated D has uncertainty. We propose to find a CD for $v = m_1(z)m_0(z)$ and compute a 'confidence measure of the decision $D = 1 = v > 0$. This measure, with value in [0,1], provides a frequency-based assessment about the decision. For example, if the measure for $D = 1$ is 63%, then, out of 100 patients the same as patient z, 63 will benefit using treatment and 37 will be better off in control group. This confidence measure is shown to match well with the classical assessments of sensitivity and specificity, but without the need to know the true $D = 1$ or $D = 0$. Utility of the development is demonstrated in an adaptive clinical trial.

E1063: **Design efficient composite likelihoods**
Presenter: **Cristiano Varin**, Ca Foscari University of Venice, Italy
Composite likelihood is an inference function constructed by compounding component likelihoods based on low dimensional marginal or conditional distributions. Since the components are multiplied as if they were independent, the composite likelihood inherits the properties of likelihood inference from a misspecified model. The virtue of composite likelihood inference is combining the advantages of likelihood with computational feasibility. Given the wide applicability, composite likelihoods are attracting interest as scalable surrogate for intractable likelihoods. Despite the promise, application of composite likelihood is still limited by theoretical and computational issues that have received only partial or initial responses. Open theoretical questions concern characterization of general model conditions assuring validity of composite likelihood inference, optimal selection of component likelihoods and precise evaluation of estimation uncertainty. Computational issues concern how to design composite likelihoods to balance statistical efficiency and computational efficiency. After a critical review of composite likelihood theory, we shall focus on the potential merits of composite likelihood inference in modeling complex forms of dependence in discrete and categorical data.

EO428 Room Q1 ADVANCES IN COMPUTING FOR ROBUSTNESS Chair: Emmanuele Sordini

E0981: **Real-time outlier detection based on DetMCD**
Presenter: **Iwein Vranckx**, KU Leuven, Belgium
Co-authors: Bart de Ketelaere, Mia Hubert, Peter Rousseeuw
Modern, state-of-the-art sorting machines can use robust covariance-based classification methods in order to separate the regular samples from outliers (raisins versus glass). Compared to well-known machine learning methods like deep neural networks and SVMs, robust statistical classifiers offer comparable classification efficiencies, in addition to being highly resistant against trainings-set contamination at the same time. However, industrial machines generate several gigabytes of spectroscopic measurements in milliseconds, frequently pushing the boundaries of currently available computational power. Due to the time criticalness of industrial classification tasks in day-to-day operations, combined with the vast amount of spectroscopic data, high performance is essential. The presented research therefore focuses specifically on the computational improvement of the DetMCD algorithm: a highly robust and deterministic estimator of location and scatter. To illustrate the performance of our accelerated DetMCD estimator, the algorithm is applied to industrial spectroscopic datasets of various food related products. We demonstrate the corresponding machine efficiency improvements and highlight the improved classification training times.

E1066: **Computational tools and methods for statistical inference based on using the characteristic functions**
Presenter: **Viktor Witkovsky**, Slovak Academy of Sciences, Slovakia
The exact distributions of many estimators and test statistics can be specified by their characteristic functions. Frequently, such distributions can be expressed as linear combinations or products of independent random variables with known parametric and/or non-parametric (empirical) distributions, specified by their characteristic functions. We consider computational methods and algorithms for specific problems (as e.g. the exact tests in multivariate analysis or non-parametric methods) of statistical inference based on combining characteristic functions and their numerical inversion to evaluate the associated CDF/PDF and quantiles. New methods and algorithms are being continuously developed and implemented into the MATLAB resp. R version of CharFunTool - the Characteristic Functions Toolbox. We shall present current status and basic functionality of the toolbox and illustrate its usage for selected problems of statistical inference.

E1291: **Dynamic graphics for robust multivariate analysis in R**
Presenter: **Emmanuele Sordini**, Joint Research Centre of the European Commission (JRC), Italy
Co-authors: Valentin Todorov, Aldo Corbellini
The monitoring of robust estimates computed over a range of key parameter values is a technique advocated in a number of recent articles. Through this approach the diagnostic tools of choice can be tuned in such a way that highly robust estimators which are as efficient as possible are obtained. The forward search for multivariate analysis is an algorithm for avoiding outliers by recursively constructing subsets of good observations and the underlying idea can be extended to many other techniques like S- and MM-estimates. To illustrate the forward search analysis, we start with a simple example and then analyze a real life data set. The analysis is conducted with the R package "fsdaR", which makes the analytical and graphical tools of the MATLAB FSDA library available to R users. The estimations are presented in monitoring plots of all n squared Mahalanobis

distances which can be combined with brushing to relate Mahalanobis distances to data points exhibited in scatterplot matrices. In this way, a straight relationship between statistical results and individual observations is established.

E1522: **Diagnostics for scale-shape models based on robust statistics**
Presenter: **Peter Ruckdeschel**, University of Oldenburg, Germany
Co-authors: Nataliya Horbenko, Matthias Kohl
Infrastructure and robustness-based diagnostics for scale-shape models available in R package RobExtremes (on CRAN since 08/2018) are presented. These scale-shape models cover amongst others, Pareto, generalized Pareto, generalized extreme value, gamma, and Weibull distributions. Lacking equivariance in the shape, these models call for refined computational techniques to achieve acceptable timings. RobExtremes provides speeded up optimally-robust estimators for these models together with high-breakdown starting estimators. Of course, MLEs and Minimum-Distance-Estimators (MDEs) are also available through R package distrMod. Diagnostics from R package ismev are available for our model fits, as well as the robustness-based diagnostic plots from R package RobAStBase such as outylingness plots, influence curve plots, information plots. In addition, we provide non-parametric confidence bands for qqplots and return level plots. We demonstrate these diagnostics with data from hydrology, hospital length of stay, and finance.

EO504 Room P2 BAYESIAN ANALYSIS WITH LARGE DATA Chair: Radu Craiu

E0279: **Bayesian spatiotemporal modeling using hierarchical spatial priors with applications to fMRI**
Presenter: **Galin Jones**, University of Minnesota, United States
A spatiotemporal Bayesian variable selection model is proposed for detecting activation in functional magnetic resonance imaging (fMRI) settings. Following recent research in this area, we use binary indicator variables for classifying active voxels. We assume that the spatial dependence in the images can be accommodated by applying an areal model to parcels of voxels. The use of parcellation and a spatial hierarchical prior (instead of the popular Ising prior) results in a posterior distribution amenable to exploration with an efficient Markov chain Monte Carlo algorithm. We study the properties of our approach by applying it to simulated data and fMRI data sets.

E0385: **Bayes calculations from quantile implied likelihood**
Presenter: **George Karabatsos**, University of Illinois-Chicago, United States
Co-authors: Fabrizio Leisen
A Bayesian model can have a likelihood function that is analytically or computationally intractable, perhaps due to large data sample size or high parameter dimensionality. For such a model, a likelihood function is introduced which approximates the exact likelihood through its quantile function, and is defined by an asymptotic chi-square distribution based on confidence distribution theory. This Quantile Implied Likelihood (QIL) gives rise to an approximate posterior distribution, which can be estimated either by maximizing the penalized log-likelihood, or by any standard Monte Carlo algorithm. The QIL approach to Bayesian Computation is illustrated through the Bayesian analysis of simulated and real data sets having sample sizes that reach the millions, involving models for univariate or multivariate iid or non-iid data. They include the Student's t, g-and-h, and g-and-k distributions; the Bayesian logit regression model; Exponential random graph model, a doubly-intractable model for networks; the multivariate skew normal model for robust inference of large inverse-covariance matrices; the Wallenius distribution model for preference data; and a novel high-dimensional Bayesian nonparametric model for distributions under unknown stochastic precedence order-constraints.

E0441: **Automated scalable Bayesian inference via data summarization**
Presenter: **Tamara Broderick**, MIT, United States
The use of Bayesian methods in large-scale data settings is attractive because of the rich hierarchical relationships, uncertainty quantification, and prior specification these methods provide. Many standard Bayesian inference algorithms are often computationally expensive, however, so their direct application to large datasets can be difficult or infeasible. Other standard algorithms sacrifice accuracy in the pursuit of scalability. We take a new approach. Namely, we leverage the insight that data often exhibit approximate redundancies to instead obtain a weighted subset of the data (called a "coreset") that is much smaller than the original dataset. We can then use this small coreset in existing Bayesian inference algorithms without modification. We provide theoretical guarantees on the size and approximation quality of the coreset. In particular, we show that our method provides geometric decay in posterior approximation error as a function of coreset size. We validate on both synthetic and real datasets, demonstrating that our method reduces posterior approximation error by orders of magnitude relative to uniform random subsampling.

E0963: **Bayesian regularization and computation for graphical models**
Presenter: **Feng Liang**, University of Illinois at Urbana-Champaign, United States
Co-authors: Lingrui Gan, Naveen Naidu Narisetty
A Bayesian framework is considered for estimating a high-dimensional sparse precision matrix, in which adaptive shrinkage and sparsity are induced by a mixture of Laplace priors. Besides discussing our formulation from the Bayesian standpoint, we investigate the MAP (maximum a posteriori) estimator from a penalized likelihood perspective that gives rise to a new non-convex penalty approximating the L_0 penalty. Optimal error rates for estimation consistency in terms of various matrix norms along with selection consistency for sparse structure recovery are shown for the unique MAP estimator under mild conditions. For fast and efficient computation, an EM algorithm is proposed to compute the MAP estimator of the precision matrix and (approximate) posterior probabilities on the edges of the underlying sparse structure. Through extensive simulation studies and a real application to a call center data, we have demonstrated the fine performance of our method compared with existing alternatives.

EO280 Room Q2 ADVANCES IN BAYESIAN MODELLING Chair: Raffaele Argiento

E0333: **Varying-sparsity regression models with application to cancer proteogenomics**
Presenter: **Francesco Stingo**, University of Florence, Italy
Co-authors: Veerabhadran Baladandayuthapani
Identifying patient-specific prognostic biomarkers is of critical importance in developing personalized treatment for clinically and molecularly heterogeneous diseases such as cancer. We propose a novel regression framework, Bayesian hierarchical varying-sparsity regression (BEHAVIOR) models to select clinically relevant disease markers by integrating proteogenomic (proteomic+genomic) and clinical data. Our methods allow flexible modeling of protein-gene relationships as well as induces sparsity in both protein-gene and protein-survival relationships, to select genomically driven prognostic protein markers at the patient-level. Simulation studies demonstrate the superior performance of BEHAVIOR against competing method in terms of both protein marker selection and survival prediction. We apply BEHAVIOR to The Cancer Genome Atlas (TCGA) proteogenomic pan-cancer data and find several interesting prognostic proteins and pathways that are shared across multiple cancers and some that exclusively pertain to specific cancers.

E0824: **Generalised graphical models for the analysis of phospho-flow cytometry data from drug combination experiments**
Presenter: **Andrea Cremaschi**, Universitetet i Oslo, Norway
Co-authors: Manuela Zucknick, Kjetil Tasken, Sigrid Skanland

The study of drug interaction via concentration-response experiments has recently received increasing attention. One technique used to produce such data is phospho-flow cytometry, measuring the expressions of a set of proteins of interest, pre-selected according to the study to be undertaken (e.g. proteins targeted by a specific drug or involved in the evolution of a type of cancer). Hence, the phosphorylation level of such proteins can be seen as a real-valued vector of the same dimension as the number of selected proteins. A typical application is the study of drugs that act on cancer cells by stimulating certain signalling pathways. For proteins that are in such pathways, we can expect that their level of activity as expressed by their phosphorylation level will depend on the concentration of the tested drugs. Signalling pathways imply interaction between the proteins of interest and can be depicted as graphs, but knowledge about these structures is based on experiments and typically incomplete. We propose a Bayesian generalised graphical model to study the relation among such proteins, when two compounds are tested simultaneously over a set of pairs of concentrations. In particular, the vectors of phosphorylation levels are modelled via Gaussian graphical models where the graph depends on the concentrations of the two drugs. The proposed framework is applied to a Chronic Lymphocytic Leukemia (CLL) dataset where two drugs used in CLL management are combined.

E0994: **Exact inference for Cox-Ingersoll-Ross driven hidden Markov models**
Presenter: **Matteo Ruggiero**, University of Torino, Italy
Co-authors: Guillaume Kon Kam King, Omiros Papaspiliopoulos

Inference is considered on the hidden signal and on the respective parameters for a hidden Markov model driven by a Cox-Ingersoll-Ross diffusion, with discretely collected observations from a marginally conjugate emission density. We show that a set of sufficient conditions related to dual processes that allow us to derive in closed forms all quantities of interest are in this case met, and provide recursive forms for the filtering and the smoothing distributions, as well as for the marginal likelihood of the observations. All these expressions are computable in that their evaluation entails only a finite computational effort. We investigate the implications of our results, which easily accommodate certain pruning techniques for speeding up inference, which are in turn tested against competitive alternative methods.

E1167: **Streaming statistical models via merge & reduce**
Presenter: **Katja Ickstadt**, TU Dortmund University, Germany

Merge & Reduce is a general algorithmic scheme in the theory of data structures. Its main purpose is to transform static data structures into dynamic data structures with as little overhead as possible. This can be used to turn classic off-line algorithms for summarizing and analyzing data into streaming algorithms. We transfer these ideas to the setting of statistical data analysis in streaming environments. Our approach is conceptually different from previous settings where Merge & Reduce has been employed. Instead of summarizing the data, we combine the Merge & Reduce framework directly with statistical models. This enables performing computationally demanding data analysis tasks on massive data sets. The computations are divided into small tractable batches independent of the total number of observations n and the results are combined in a structuredway at the cost of a bounded $O(logn)$ factor in their memory requirements. It is only necessary (though non-trivial) to choose an appropriate statistical model and implement merge and reduce operations for the specific type of model. We illustrate our Merge & Reduce schemes on simulated and real world data employing Bayesian linear regression models, Gaussian mixture models, and generalized linear models.

CO254 Room A2 FREQUENCY DYNAMICS OF ECONOMIC AND FINANCIAL VARIABLES Chair: Jozef Barunik

C0531: **The yield curve and the stock market: Mind the long run**
Presenter: **Fabio Verona**, Bank of Finland, Finland

Central banks' monetary policy actions affect a broad spectrum of interest rates, which in turn have an impact on stock markets and, ultimately, on stockholders' wealth. It is thus important for central banks to better understand the effects of interest rates movements on the stock market. A variable of interest to policymakers and financial markets participants alike is the slope of the yield curve, also known as the term spread of interest rates. We extract cycles from the term spread and study their role for predicting the equity premium using linear models. When properly computed, the trend of the term spread is a strong and robust out-of-sample equity premium predictor, both from a statistical and an economic point of view. Properly means that it is crucial the way the higher-frequency fluctuations are eliminated, as certain filtering methods (namely wavelet filters) enable the extraction of a low-frequency component with equity premium forecasting performance clearly superior to that other filters (like band-pass filters). For policymakers and market participants interested in gauging equity market developments, the proper trend of the term spread can thus be a promising variable to look at.

C0701: **Horizon-specific risks, higher moments, and asset prices**
Presenter: **Josef Kurka**, UTIA AV CR, v.v.i., Czech Republic
Co-authors: Jozef Barunik

Asset pricing traditionally works with information aggregated over horizons, however investors' preferences are horizon-specific. Decomposing returns, and risk factors to components representing individual horizons may hence provide valuable insights into pricing mechanisms of investors. With increasing size of factor-investing literature, the number of factors approximating risk, and possibly explaining the cross section of returns is growing rapidly. However, most of the factors perform poorly in subsequent out-of-sample testing. Therefore, attention should be turned to theory-based factors approximating the risks such as moments of the return distribution that are found to be priced empirically. We derive an asset pricing model that contains second, third and fourth centralized moments of returns on aggregate wealth decomposed to short-run and long-run components. Empirical results show that horizon-specific risk from higher moments is priced, and uncover different effects of the moment-based risk factors in short-run, and long-run.

C0703: **The variance-frequency decomposition as an instrument for the identification of SVAR models**
Presenter: **Yuliya Lovcha**, Universitat Rovira i Virgili, Spain
Co-authors: Alejandro Perez Laborda

A framework is proposed to study the identification of structural VAR models. The framework focuses on the contribution of the identified shock to the variance of the variables in the business cycle frequency range. The discussion is organized around the identification of technology shocks, since it has attracted considerable attention. First, we conduct a Monte-Carlo study to analyze within this framework the properties of a set of identification schemes for the technology shock. After that, we propose a new identification method, based on the variance-frequency decomposition, which delivers a reliable estimate of the response of hours. The empirical application of this scheme is illustrated in two datasets. Finally, we show that, aside from its use as a pure identification mechanism, the proposed method may be employed to evaluate the consistency between parameterized models and the data.

C1057: **Time-frequency response analysis of monetary policy transmission**
Presenter: **Lubos Hanus**, Charles University, Czech Republic
Co-authors: Lukas Vacha
A new approach is considered to look at the effects of economic shocks to dynamics of economic systems. We analyse the widely known phenomenon of price puzzle in a time-varying environment using the frequency decomposition. We use the frequency response function to measure the power of a shock transferred to different economic cycles. Considering both the time-variation of the system and frequency analysis, we can quantify the local dynamics of shocks at given time and over frequencies, and reveal broader policy implications the system can provide. While studying the monetary policy transmission of the U.S., the empirical evidence shows that low-frequency cycles are prevalent, yet, their amplitudes vary significantly in time.

CO296 Room B2 ADVANCES IN EMPIRICAL FINANCE AND ECONOMETRICS Chair: Jose Olmo

C0694: **Characteristic function-based approach for pricing long-run market uncertainty**
Presenter: **Abderrahim Taamouti**, Durham University Business School, United Kingdom
Co-authors: Julian Williams, Handing Sun, Yang Zhang
High-order moments derivatives such as variance and skewness swaps are increasingly popular risk management tools for foreign exchange exposures. We propose a simple characteristic function based method for determining the risk-neutral value of a spanning contract contingent on the future outcome of an asset price. Contrary to the existing approaches that require using large cross sectional data on option prices, the price of such contract can be computed by a single quadrature evaluation of the characteristic function and can be used to calculate the moment swaps beyond variance, such as skewness and kurtosis. Based on our approach, fractional moment swaps can be estimated directly from spot and yield curve data via maximum likelihood. Finally, based on our method, we show that the GARCH based predicted moment swaps perform well against both option implied variance and skewness swaps and traded variance swaps.

C0658: **Multivariate quantile impulse response functions**
Presenter: **Gabriel Montes-Rojas**, Universidad de Buenos Aires, Argentina
A multivariate vector autoregression quantile (VARQ) model is developed which is the solution to a collection of directional quantile models for a fixed orthonormal basis, in which each component represents a directional quantile that corresponds to a particular endogenous variable. This corresponds to a reduced form multivariate quantile autoregressive model is developed to study heterogeneity in the effects of macroeconomic shocks. The VARQ estimator allows us to forecast the future performance of the multivariate time-series conditional on the available information, which depends on multivariate quantile indexes. From the forecasting procedure we define an impulse-response function (IRF) model that computes the effect of a given perturbation in a (some) variable(s). This thus generalizes the mean-based IRF analysis to the multivariate quantile framework. This analysis explores potential dynamic heterogeneity not covered by the mean-based IRF analysis using mean-based VAR. In particular we can study the realization of particular sequences of events, as defined by particular values of the multivariate quantile indexes, defined as quantile paths. The model is applied to study monetary shocks in a three-variable macroeconomic model (output gap, inflation, Fed Funds rate) for the U.S. for the period 1980 to 2010.

C0660: **Measuring asset market linkages: Nonlinear dependence and tail risk**
Presenter: **Juan Carlos Escanciano**, Universidad Carlos III de Madrid, Spain
Co-authors: Javier Hualde
Traditional measures of dependence in time series are based on correlations or periodograms. These are adequate in many circumstances but, in others, especially when trying to assess market linkages (e.g., financial contagion), might be inappropriate. In particular, tail dependence measures based on correlations of single tail events have limited information on tail risk. We propose a new nonparametric measures of dependence and show how they characterize conditional dependence and persistence. We propose simple estimates for these measures and establish their limiting properties. We employ the proposed methods to analyze the persistence properties of some of the major international stock market indices during and after the 2007-2009 financial crisis. The results uncover a leading role of US and London in international diversification. Tail dependence, as quantified with the new measures, is more informative than the popular marginal expected shortfall for the US. We find a ubiquitous nonlinear persistence in conditional variance across all markets that is not explained by popular parametric models. Market crashes also show substantial persistence in likelihood and their systemic effects.

C1043: **A re-examination of the size effect: The influence of winning stocks in size portfolios**
Presenter: **Jose Olmo**, University of Southampton, United Kingdom
Co-authors: Richard McGee
The influence on the size effect of the top performing stocks on a cross-section of risky assets separated by industry is empirically investigated. To do this, we propose a conditional logit model for ranking different investment portfolios based on size and assess the robustness of the ranking to the inclusion/exclusion of the best performing stocks in the cross-section. We apply this methodology for analysing the performance of different size portfolios constructed for 20 U.S. industries over the period January 1970 to December 2015. Our results show that the size effect is spurious for most industries once we remove the winning stocks of the size portfolios. The application of this analysis to asset pricing shows that standard asset pricing models fail to correctly specify the risk premium on risky assets when the industry winners are excluded from the construction of the size factor portfolio.

CO334 Room C2 SYSTEMIC RISK Chair: Massimiliano Caporin

C1354: **Systemic-systematic risk in financial system: A dynamic ranking based on expectiles**
Presenter: **Laura Garcia-Jorcano**, Universidad de Castilla-La Mancha, Spain
Co-authors: Lidia Sanchis
An international dynamic ranking comparison for systematic and systemic risk in the financial system is provided based on a coherent downside risk measure, the expected shortfall (ES) computed from expectiles. This approach has the advantage of avoiding distributional assumptions and providing a more efficient estimation procedure than using quantiles. From these ES, estimates are obtained for static and dynamic, systematic and systemic rankings for different banks, insurances and financial services institutions by using principal component. In general, it is expected that this new approach is competitive, because it is more sensitive to the magnitude of extremes losses than conventional quantiles. The main evidence is that banks are more systemic and systematic in periods of crisis, but in quiet periods (pre and post) insurances are more systemic and systematic. For countries, institutions from Asia are more systematic than systemic and from Europe and North America more systemic than systematic. In the crisis, North American institutions are less systematic and systemic, while Asian are more systematic and systemic. These results have implications for supervisors regarding the regulation of financial firms and for investors regarding diversifiable and non-diversifiable risks in their portfolios. Moreover, it is provided additional evidence on the necessity of macro and micro-prudential regulation not only in the banking sector but also in the insurance sector.

C1613: **Systemic stress testing under central and non-central clearing**
Presenter: **Elena Kalotychou**, Cyprus University of Technology, Cyprus
Co-authors: Barbara Casu, Petros Katsoulis
Following the global financial crisis, central counterparties (CCPs) have become key participants in the OTC derivatives markets clearing an increasing proportion of contracts in accordance with international regulations aimed at mitigating systemic risk. Furthermore, since 2016 non-centrally cleared contracts have also been subject to stringent clearing regulations in order to facilitate their effective risk management and incentivize market participants to migrate to central clearing. We empirically assess the effects of the introduction of non-central clearing on counterparty, liquidity and systemic risks in the presence of central clearing by developing a dynamic macroprudential stress testing network model of the largest market participants in the OTC derivatives market. We show that non-central clearing significantly reduces counterparty and systemic risks under all market conditions, at the expense of higher liquidity risk. In addition, the expansion of central clearing further reduces systemic risk but the CCP's propensity for contagion increases as a result, becoming a significant source of stress for the market participants when extreme market conditions manifest.

C0438: **A meta-analysis of systemic risk measures for gauging financial stability**
Presenter: **Jean-Charles Garibal**, Laboratoire Economie Orleans, France
Co-authors: Massimiliano Caporin, Michele Costola, Bertrand Maillet
After the last major financial crisis of 2008, several systemic risk measures have been proposed in the financial literature as attempts for quantifying the magnitude of the financial system distress. We suggest the construction of an overall meta-index for the measurement of systemic risk based on a sparse principal component analysis of main systemic risk measures, which ultimately aims to provide an index with a more stable dynamic and which is explicitly linked to severe economic recessions.

C0439: **Traffic lights system for systemic Stress: TALIS3**
Presenter: **Massimiliano Caporin**, University of Padova, Italy
Co-authors: Juan-Angel Jimenez-Martin, Laura Garcia-Jorcano
A TrAffic LIght System of Systemic Stress (TALIS-cube) is proposed which provides a comprehensive color-based classification that groups companies according to both the level of stress reaction of the system when the company is in distress and the level of stress of the company. Our proposal builds on the Conditional Value-at-Risk measure extended by introducing Filtered Historical Simulation, preferred to the choice of a specific parametric density for the innovations. We classify companies by resorting two loss functions, one for the system and one for each company, evaluated over time and in the cross-section. From the color-based classification of companies we also recover an aggregated index. TALIS-cube can be used to enhance the performance and robustness of current systemic risk measures. We provide an empirical analysis on the US market and several robustness checks evaluating different underlying models driving variance and correlation dynamic and different tuning parameters on the loss functions and company rankings.

CO582 Room D2 ROUGH VOLATILITY Chair: Mathieu Rosenbaum

C0695: **Ergodic properties of certain financial models**
Presenter: **Miklos Rasonyi**, Renyi Institute, Budapest, Hungary
Co-authors: Balazs Gerencser
Markov chains are considered in random environments and results on their ergodic behaviour that go well beyond current literature are presented. We establish stochastic stability and the law of large numbers for functionals of these processes. Such properties are useful when analysing e.g. recursive algorithms. The results are pertinent to certain models of mathematical finance, notably to the fractional stochastic volatility model.

C0713: **Affine forward variance models**
Presenter: **Martin Keller-Ressel**, TU Dresden, Germany
Co-authors: Jim Gatheral
The class of affine forward variance (AFV) models is introduced which includes the Heston model and the rough Heston model. We show that AFV models can be characterized by the affine form of their cumulant generating function (CGF), which is obtained as solution of a convolution Riccati equation. We further introduce the class of affine forward order flow intensity (AFI) models, which are structurally similar to AFV models, but driven by jump processes. We show that the AFI model's CGF satisfies a generalized convolution Riccati equation and that a high-frequency limit of AFI models converges to the AFV model.

C0796: **Volatility derivatives in rough forward variance models**
Presenter: **Stefano De Marco**, Ecole Polytechnique, France
Forward variance models are models for the joint dynamics of an asset price and the implied volatility of its variance swaps. The works in this field can be traced back to the early 90's. Since 2008, these models have been successfully applied to the derivatives market on the VIX index. More recently, the new stream of research on rough volatility modeling has pushed forward new instances within this family notably, the rough Bergomi model. We will consider a class of models that embeds the examples above, and present some of its major properties, with a focus on the model-generated term structure of volatilities of volatilities. In particular, we will present a non-log normal extension of the rough Bergomi model that is able to accommodate smiles of VIX options in this (rough) forward variance setting, and present its appealing properties in the calibration to the VIX market.

C1168: **Learning rough volatility**
Presenter: **Blanka Horvath**, Kings College London, United Kingdom
Calibration time is the bottleneck for models with rough volatility. We present ways for substantial speed-ups, along every step of the calibration process: In a first step we describe a powerful numerical scheme (based on functional central limit theorems) for pricing a large family of rough volatility models. In a second step we discuss various machine learning methods that significantly reduce calibration time for these models. By simultaneously calibrating several (classical and rough) models to market data as a byproduct of our calibration results, we re-confirm that volatility is rough, calibration performance being best for very small Hurst parameters in a multitude of market scenarios.

CO382 Room F2 NEW DEVELOPMENTS IN NONLINEAR SPATIAL AND TEMPORAL MODELLING Chair: Maria Kyriacou

C1231: **Model averaging estimation for conditional heteroscedasticity model family**
Presenter: **Qingfeng Liu**, Otaru University of Commerce, Japan
Co-authors: Qingsong Yao, Guoqing Zhao

The model averaging estimation for the conditional heteroscedasticity model family is considered. We first consider the case of zero conditional mean. Given a set of candidate models of different functional forms, we propose a model averaging estimator for the conditional variance and construct the corresponding weight choosing criterion. It is shown that the weight minimizing the criterion asymptotically minimizes the true KL divergence as well as the discrete and continuous Itakura-Saito distance (a measurement of the distance between two positive vectors or measurable functions). As an extension, we also extend the method to the framework where both the conditional mean and variance require estimation. On this condition, we conduct the averaging procedure for the conditional mean and variance in a separate manner and derive the double averaging estimator (DAE). Monte Carlo experiments show that the model averaging estimation leads to higher estimation accuracy, which is in favor of our newly-proposed method. For empirical application, we apply the model averaging method to the estimation of the conditional volatility of Shanghai composite index.

C1241: **Nonparametric regression with network data**
Presenter: **Swati Chandna**, Birkbeck, University of London, United Kingdom
Co-authors: Pierre-Andre Maugis

Nonparametric methods are introduced which address the setting where a sample of small networks, along with additional information, is observed. For example, in a connectome study, for each individual in the sample both a structural brain network is observed, along with covariates such as age, gender, etc. We work under the framework of exchangeability commonly used to model network data where the node labels carry no information. Under this formulation, estimation of the limit object termed 'graphon', has attracted significant attention in the nonparametric literature on networks. Building upon the standard graphon model, we provide a framework that can test for any given node presenting significantly different behavior across different values of the covariates. Further, we find that although a significant portion of the graphon literature focuses on block-model approximations of the graphon, in our setting full nonparametric inference is possible and computationally tractable. We illustrate our approach using a set of brain network observations from multiple individuals.

C1282: **Continuously updated indirect inference**
Presenter: **Maria Kyriacou**, University of Southampton, United Kingdom
Co-authors: Peter CB Phillips, Francesca Rossi

Spatial units are often heterogeneous as they vary in many of their observed characteristics and so the assumption of homoskedasticity may not hold in practice. In the presence of unobserved heterogeneity of the disturbance term, standard methods based on the (quasi-)likelihood function produce, in general, inconsistent estimates of both the spatial parameter and the exogenous regressors coefficients. There is an evident lack of estimation methods that account for the presence of heteroskedasticity of unknown form. A robust generalized methods of moments estimator as well as a modified likelihood method have been previously proposed to address this issue. We propose an indirect inference methodology which relies on a simple Ordinary Least Squares (OLS) procedure as its starting point, which is computationally simple, robust to unobserved heterogeneity, allows for a general range of weight matrix structures and has excellent finite sample performance. Our proposed Continuously Updated Indirect Inference (CUII) estimator is derived using a binding function with a continuously-updated diagonal variance-covariance matrix. Simulation results reveal that our proposed estimator is effective in reducing both bias and MSE compared to competitor estimators.

C1526: **Nonparametric trend ordinary kriging method with applications to air quality data**
Presenter: **Lifeng Yang**, University of Southampton, United Kingdom
Co-authors: Zudi Lu

The three commonly used linear kriging methods, i.e. simple -, ordinary- and universal kriging, offer the property of best linear unbiased predictor. However, these uncomplicated spatial prediction and mapping procedures, which make them appealing to many practitioners, are subject to criticism owing mainly to their over-simplified linear regression trend structure, often introducing misspecifications to its spatial function. Instead of linear trend structure of data we propose a Nonparametric-Trend Ordinary Kriging (NTOK) method, to overcome this structural drawback by delegating the estimation of spatial trend to a nonparametric function. Combining its result with the current ordinary kriging method, the new NTOK acts as an improved alternative to the existing linear methods. Asymptotic justification for this estimation procedure is developed. Empirical applications of the above methods to air quality in the UK are compared to show the improvement of the proposed NTOK prediction.

CO562 Room G2 ECONOMETRIC METHODS FOR SPORT MODELLING AND FORECASTING Chair: Luca De Angelis

C0315: **Forecasting tennis betting odds by artificial neural networks**
Presenter: **Vincenzo Candila**, University of Salerno, Italy

In sports literature, published betting odds are considered the most accurate source of probability forecasts. However, due to the presence of the longshot bias and bookmaker's over-round, these betting odds do not represent true bookmaker expectations about the outcome of the event under consideration. Artificial neural networks (ANNs) are employed to forecast betting odds in tennis betting market, starting from variables observable at the beginning of the matches. The ANNs are capable of handling a variety of input variables, contrary to standard approaches in the context of sport outcome forecasting. Moreover, the forecasted betting odds by ANNs are odds directly related to the probability of winning, such that there is no bookmaker's over-round and no longshot bias. In terms of probabilities, the proposed ANN model provides forecasting performances generally superior to the benchmarks considered. Moreover, forecasting betting odds by ANNs generally allows us to achieve positive returns. For robustness purposes, the analysis is repeated considering different datasets consisting of matches randomly selected.

C0465: **Modelling performance variability and teammates' interactions in basketball**
Presenter: **Paola Zuccolotto**, University of Brescia, Italy
Co-authors: Marica Manisera

Basketball players' performance measurement is of critical importance for a broad spectrum of decisions related to training and game strategy. Despite this recognized central role, the main part of the studies on this topic focus on performance level measurement, neglecting other important characteristics, such as variability. Shooting performance variability is modeled with a Markov Switching dynamic, assuming the existence of two alternating performance regimes. Then, the relationships between each player's variability and the lineup composition is modeled as an ARIMA process with covariates and described with network analysis tools, in order to extrapolate positive and negative interactions between teammates, helping the coach to decide the best substitution during the game.

C0776: **Informational efficiency and price reactions in exchange betting markets**
Presenter: **Luca De Angelis**, University of Bologna, Italy
Co-authors: Giovanni Angelini
The degree of efficiency of exchange betting markets is investigated. Using event study analysis on high-frequency data, we examine the reaction of prices to events and the arrival of major news. In particular, we measure the post-event jumps of in-play odds and we analyse their dynamic behaviour. We test for informational efficiency in football exchange betting markets in three different directions: (i) we model and forecast the price reaction to news events (i.e. goals, red cards), (ii) we test whether price jumps create systematic bias which can be exploited to set a profitable betting strategy, (iii) we focus on possible over/under-reaction by investigating the main drivers which may create deviations from the expected intensity of jumps. To do so, we consider a dataset comprising prices collected every ten seconds from Betfair Exchange for all the English Premiership matches played in the 5 Seasons from 2009/2010 to 2013/2014.

C1316: **Picking scores: Forecasting low probability events**
Presenter: **James Reade**, University of Reading, United Kingdom
Co-authors: Carl Singleton, Alasdair Brown
In sport, and football in particular, there is a great level of satisfaction taken in correctly guessing what the score will be in advance of a game taking place. Part of this satisfaction must be drawn from just how difficult it is to predict exact scores; these are low-probability events. Furthermore, it is believed that agents are unable to distinguish between low and tiny probability events, and this forms one explanation for observed departures from price efficiency in gambling markets. We investigate scoreline forecasts from a range of sources; bookmaker prices, user tipsters, experts, and statistical models. We rank the various forecasts by a range of scoring rules both relative and absolute, and quantify the extent to which known biases are existent in these different types of forecast.

CO376 Room I2 SEMI- AND NONPARAMETRIC METHODS FOR NONLINEAR REGRESSION Chair: Harry Haupt

C0982: **Nonlinear quantile regression-based modeling of hedonic housing prices**
Presenter: **Markus Fritsch**, University of Passau, Germany
Co-authors: Harry Haupt, Joachim Schnurbus
Many applications of statistical real estate appraisal methods involve the following challenges: identifying the relevant characteristics of a property, estimating the shadow prices (marginal market valuation) of characteristics and estimating the prices of bundles (of characteristics) not observed. State-of-the-art hedonic housing price analysis comprises modeling price functions nonlinearly, accounting for complex spatial association structures (horizontal market segmentation), and allowing for varying functional relationships across the conditional price distribution (vertical market segmentation). We discuss two general classes of nonlinear quantile regression models which meet these criteria but pursue different avenues to simultaneously address the challenges outlined above. Due to the underlying assumptions, the inference obtained from both model classes differs analytically and – more importantly – leads to different economic interpretations. The methods are illustrated by applying them to data generating processes with various degrees of functional and spatial complexity in a Monte Carlo study and to geo-referenced urban housing price data.

C0998: **Mixed kernel estimation of counterfactual distributions for Munich rent survey**
Presenter: **Joachim Schnurbus**, University of Passau, Germany
Co-authors: Goeran Kauermann
The tremendous increase in rent prices for apartments in larger cities is a problem in most countries. A general question in this respect is whether increments in the rent are merely a matter of rising demand that is exploited or whether an increasing quality of apartments also contributes to the increase in the apartment rent. To tackle this question we provide a counterfactual distribution-based decomposition of several current releases of rent surveys from Munich, Germany, that allows to disentangle the rent increase over time into two effects. First, the rent increase caused by an improvement of the flats and second, the increase due to inflation and demand. A novel nonparametric kernel estimator for mixed continuous and discrete covariates is proposed for estimating the counterfactual distribution. Application to the Munich rent market indicates that the majority of the rent increase seems not to be justified by improved flat characteristics.

C1076: **Additive semiparametric framework for land use regression**
Presenter: **Svenia Behm**, University of Passau, Germany
Co-authors: Markus Fritsch, Harry Haupt
Existing land use regression (LUR) approaches usually employ parametric assumptions to model the conditional distribution of air pollutant measurements. We propose a flexible data-driven additive semiparametric framework for modeling the annual mean nitrogen dioxide concentration across Germany which rests on the crucial assumption of additivity. Through exploratory analysis, we find considerable spatial variation and nonlinearities in the data and, therefore, model the spatial characteristics via bivariate splines and the structural characteristics via univaritate splines and in linear additive form. Our specification allows us to account for local heterogeneity, potential nonlinearities and spatial anisotropy in a flexible, data-driven way – while avoiding imposing parametric assumptions a priori (i.e., without looking at the data). In- and out-of-sample metrics support the proposed model. Additive semiparametric models are a promising choice to analyse and predict the conditional distribution of air pollutant concentration. A straightforward extension of our approach is to model several characteristics of the conditional pollutant distribution such as quantiles or expectiles.

C1706: **Improved asymptotics for nonlinear quantile regression**
Presenter: **Harry Haupt**, University of Passau, Germany
Based on recent results on the asymptotics for nonlinear quantile regressions under dependence and heterogeneity, some improved assumptions are proposed and possible extensions are discussed.

CO258 Room M2 BEHAVIORAL FINANCIAL MACROECONOMICS Chair: Christian Proano

C0345: **A toxic cocktail: Low interest rates and banks' search-for-yield behavior**
Presenter: **Benjamin Lojak**, University of Bamberg, Germany
Co-authors: Christian Proano, Tomasz Makarewicz
The relationship between expansionary monetary policy and banks' risk taking behavior is investigated. We study a model in which risk averse banks credit firms and also manage a portfolio consisting of a risky and a risk-free asset. When banks fund firms they take into account their solvency and potential gains from outside investment strategies. We show that low policy rates induce banks to search-for-yield through the softening of lending standards, incentivizes firms to take on riskier positions and increases asset price volatility that renders the financial markets more unstable.

C0351: **Financial frictions and wages**
Presenter: **Britta Gehrke**, University of Erlangen-Nuremberg, Germany
Co-authors: Hamzeh Arabzadeh, Almut Balleer
The interaction between financial frictions and wages is analyzed. We use a large data set for Germany for 2006 to 2014 that combines administrative data on workers and wages with detailed information on firms' balance sheets. Controlling for firm characteristics and time fixed effects, we find that higher leverage (as a measure for financial frictions) implies on average lower wages. We build a theoretical model with labor market frictions and monitoring costs in the financial market. We show that wages react differently to financial frictions depending on whether and how they affect the relative costs of wages and hiring and the surplus of the job. We further show how employment volatility depends on these different mechanisms and document how higher employment volatility can be related to less rather than more rigid wages. Our empirical results then identify these different mechanisms in the data.

C0361: **Animal spirits in a NKM with financial intermediation and a stock market**
Presenter: **Naira Kotb**, Otto-Friedrich-Universitat Bamberg, Germany
Co-authors: Christian Proano
A macro-finance interaction model is presented which integrates together a NKM with bounded rationality, an agent-based stock market and a banking sector. In the model, financial intermediation leads to the existence of two, rather than one, interest rate: (1) the policy rate, which is also the rate paid by banks to households on deposits, and (2) the rate paid by firms to banks on loans. The later constitutes of the first plus a spread. The spread is a key variable in the model. Households' savings are diversified among bank deposits and stock purchases. It is found that, banking intensifies animal spirits and amplifies shocks. The intensity of amplification is much higher when the spread is closely tied to the output gap than when it is tied to the stock prices. It is also found that the effect of households' purchase of stocks on the stability of the real sector (and the stock market) depends on the choice of parameters that relate the spread to the output gap and the stock prices.

C1678: **Animal spirits, risk premia and monetary policy at the zero lower bound**
Presenter: **Christian Proano**, University of Bamberg, Germany
Co-authors: Benjamin Lojak
A stylized macroeconomic model is set up to analyze the risk-related effects of monetary policy under boundedly rational perceptions both in normal times, as well as in periods where the zero lower bound (ZLB) binds. In our model financial market participants use alternative heuristics to assess the risk premium over the policy rate in accordance to an "implicit Taylor rule" that measures the stance of conventional monetary policy and which serves as an informative instrument during times when the funds rate is constrained by the ZLB. In such a case, conventional monetary policy is totally exhausted so that the central bank is forced to move to unconventional types of policy. We propose alternative monetary policy measures out of the liquidity trap effective under the assumed form of bounded rationality.

CO434 Room N2 FINANCIAL TIME SERIES ECONOMETRICS Chair: Peter Exterkate

C0344: **Jump testing with the pre-averaged bipower variation and subsampling estimation of the asymptotic variance matrix**
Presenter: **Bezirgen Veliyev**, Aarhus University, Denmark
Co-authors: Kim Christensen, Nopporn Thamrongrat
A noise-robust extension of the bipower variation-based jump test is proposed, which is based on pre-averaging and can implemented on high-frequency data that are perturbed by microstructure noise. The main hurdle in this context is to derive a consistent estimator of the asymptotic covariance matrix of the pre-averaged bipower variation, which is asymptotically jump-robust and has good finite sample properties, such as being positive semi-definite and well-conditioned. We propose such an estimator based on a subsampling approach. Simulations show that the proposed test has size control under a variety of noise structures, while it has excellent power under the alternative. In the empirical application, we recover jump dates from real data.

C0429: **Variance estimation in the presence of self-excited jumps**
Presenter: **Simon Kwok**, University of Sydney, Australia
Robust inference on the volatility process has become an important topic in high-frequency financial econometrics. There are methods for disentangling the variance component from jumps in jump diffusion models, but not all of them are robust to more realistic forms of jump dynamics. For example, bipower variation tends to overestimate the integrated variance when jumps are self-excited. The robustness of different estimators of integrated variance is studied and compared. It is found that the nonparametric threshold method delivers a more precise estimate than the multipower variation approach in the presence of autocorrelated jumps.

C0843: **The taming of the two: Simulation-based asset pricing with multi-period disasters and two consumption goods**
Presenter: **Jantje Soenksen**, Eberhard Karls University Tuebingen, Germany
A novel approach is proposed to facilitate the estimation of the preference parameters of a two consumption good C-CAPM that accounts for multi-period disasters, partial government defaults, and the possible destruction of the stock of the durable good. The maximum likelihood estimation of the disaster process parameters requires a cross-country panel of historical consumption data and international business cycle dates. The estimation of the risk aversion coefficient and the intertemporal elasticity of substitution (IES) is facilitated by the simulated method of moments. The results show that the empirical equity premium can be explained with economically plausible and quite precise risk aversion and IES estimates. This conclusion withstands a battery of robustness checks.

C0991: **Loss function derived expected shortfall backtests under estimation error**
Presenter: **Sander Barendse**, Erasmus University Rotterdam, Netherlands
The purpose is to investigate estimation error effects on expected shortfall (ES) backtests that are based on first order conditions of a recently introduced joint consistent loss function for Value-at-Risk (VaR) and ES. We show that the asymptotic covariance of the test statistics contain additional terms when estimation error is present, and provide explicit expressions for these terms, which are functions of the model specification. We develop a bootstrap procedure to correct for estimation error. In Monte-Carlo experiments we observe that the robust backtests based on bootstrap critical values have correct size properties, whereas the uncorrected backtests overreject considerably. Finally, we compare the ES backtests studied with competing backtests for several data generating processes. We find that no backtest consistently outranks other backtests in terms of power.

CO070 Room O2 ECOSTA JOURNAL: HIGH FREQUENCY DATA Chair: Alain Hecq

C0230: **High frequency linear time series models and mixed frequency data**
Presenter: **Manfred Deistler**, Vienna University of Technology, Austria
The focus is on the identifiability of the parameters of high frequency multivariate ARMA type models from mixed frequency time series data. For the VAR case, we demonstrate identifiability for generic parameter values using the population second moments of the observations. We display a constructive algorithm for the parameter values and establish the continuity of the mapping attaching the high frequency parameters to these populations second moments. These structural results are obtained using two alternative tools: extended Yule Walker equations and blocking of the output process. The cases of stock and flow variables, as well as of general linear transformations of high frequency data, are treated. We discuss how our constructive identifiability results can be used for parameter estimation. In a next step we show that the results on generic identifiability can be extended to the VARMA case, provided that the MA order is smaller than or equal to the AR order. However, in the case where the MA order exceeds the AR order, and in particular in the VMA case, results are completely different. Then, when the innovation covariance matrix is non-singular, "typically" non-identifiability occurs not even local identifiability. This is because, e.g., in the VMA case, as opposed to the VAR case, the not directly observed auto-covariances of the output can vary "freely".Finally, we discuss modeling by generalized linear dynamic factor models in the mixed frequency case.

C0829: **Modeling high-frequency trading volume**
Presenter: **Eduardo Rossi**, University of Pavia, Italy
Co-authors: Paolo Santucci de Magistris, Leopoldo Catania
Trading volume can be measured instantaneously for each trade or cumulated for a given time interval (time aggregates). The latter implies that for longer time intervals the trading volume is an increasing process. In high-frequency trading, this data seem to be preferred to tick-by-tick level data as it dispenses with certain pitfalls in econometric modeling, such as the irregular spacing of time spells. For short time intervals and less liquid stocks cumulated trading volume series contain a high proportion of zero observations. Further, the cumulated trading volumes series show overdispersion and intraday periodicities. Since we do not apply any transformation to the cumulated trading volumes, these have to be treated as realizations of non-negative integer random variables. When we consider long time span, cumulated trading volumes series can be characterized by trends and heterogeneity across time. The aim is to propose a new approach to the modeling of cumulated trading volumes series based on mixtures of discrete time integer-valued processes. The resulting process has a closed-form conditional density which can also be specified with time-varying parameters to accommodate the evolving features of the observed series.

C1211: **Mixed time aggregation of multivariate linear processes**
Presenter: **Michael Thornton**, University of York, United Kingdom
The time aggregation of vector linear processes: (i) containing mixed stock-flow data; and, (ii) aggregated at mixed frequencies is explored, showing how the parameters of the underlying model translate into those of the equivalent model of the aggregate. Based on manipulations of a general state-space form, the results may be applied to a wide range of linear ARMAX processes, including the discrete representation of a continuous time process, and may be iterated to model multiple frequencies or aggregation schemes.

E0251: **The beta-adjusted covariance estimator**
Presenter: **Kirill Dragun**, VUB, Belgium
Co-authors: Kris Boudt, Steven Vanduffel
Stock return covariance estimation is proposed to be improved by imposing the equality between the covariance matrix-implied stock-ETF covariance, and the estimated stock-ETF pairwise covariance. The proposed beta adjusted covariance estimation iteratively projects the realized covariance on an improved covariance respecting the constraints. The simulation study confirms that the proposed estimator efficiently deals with biased approximations by traditional estimators caused by asynchronous trading data and significantly improves accuracy of the estimated covariances.

CC651 Room E2 CONTRIBUTIONS IN FINANCIAL ECONOMETRICS I Chair: Esther Ruiz

C1549: **Statistical inferences for price staleness**
Presenter: **Aleksey Kolokolov**, Alliance Manchester Business School, United Kingdom
Co-authors: Davide Pirino, Giulia Livieri
Asset transaction prices sampled at high frequency are much staler than one might expect, in the sense that they frequently lack new updates showing zero returns. We propose a theoretical framework that hinges on the existence of a latent continuous-time stochastic process p_t valued in the open interval (0,1), which represents, at any point in time, the probability of occurrence of a zero return. Using a standard infill asymptotics design, we develop an inferential theory for testing, non-parametrically, the null hypothesis that p_t is constant over one day. Under the alternative, which encompasses a semimartingale model for p_t, we prove that the integrated volatility of the probability of staleness can be consistently estimated. Empirically, on a large dataset of NYSE stocks, we provide evidence that the null of constant probability of staleness is fairly rejected and that the integrated volatility of p_t is mainly determined by transaction volume, bid-ask spread and realized volatility.

C1642: **Recent credit risk and bubble behavior in the corporate energy sector**
Presenter: **Isabel Figuerola-Ferretti**, Universidad Pontificia Comillas, Spain
The relationship between oil and credit risk in the energy sector over the last two oil price crises is analyzed. We measure credit risk in energy corporations using CDS spreads and assess whether credit risk in energy companies exhibited departures from random walk behavior. By using the multiple bubble methodology proposed, we detect two main mildly explosive periods in CDS prices: a predominant mild explosive period just before the global financial crisis and another important explosive period after the recent 2014 crude oil price collapse. We relate the dated bubble episodes seen in CDS prices with the time series behavior of crude oil prices and with the level of corporate debt measured by different corporate leverage measures. Results show that the 2015 episode of mild explosivity reported for CDS prices in energy corporates is associated with an abrupt increase in debt levels and debt equity ratios following the taper tantrum in 2013 and the subsequent 2014 crude oil sell off.

C1398: **Trends everywhere: The case of hedge fund styles**
Presenter: **Charles Chevalier**, Universite Paris Dauphine, France
Co-authors: Serge Darolles
The aim is to investigate empirically whether time-series momentum returns can explain the performance of hedge funds in the cross-section. Following the trend of the literature, a volatility adjusted time-series momentum signal is applied on a daily basis across a large set of futures, covering the major asset classes. We build a hierarchical set of trend factors: the full version TREND can be split in summable factors across two dimensions, the horizon of the signals and the traded asset class. We show that Managed Futures, Global Macro and Fund of Hedge Funds strategies can be partly explained by a TREND exposure, whereas Equity Market Neutral and Quantitative Directional are only exposed to long term trend factors. Moreover, a TREND exposure is a significant determinant of hedge funds returns at the aggregate level, as well as at the fund

level. Finally, funds with high TREND beta outperform by 41 basis points of alpha the funds with low Trend beta. These results prove useful when managing the risk of a portfolio of hedge funds strategies, since assessment of the Trend exposure is easier. Another contribution is related to the understanding of the CTA space, composed of pure trend funds as well as funds that do not exhibit any TREND exposure.

C1719: **Models for realised volatility**
Presenter: **Dario Palumbo**, University of Cambridge, Italy
Co-authors: Andrew Harvey
A statistical framework is set for modeling realised volatility (RV) using DCS/GAS. It shows how a preliminary analysis on FTSE, based on fitting a linear Gaussian model to logRV confirms a two component specification and at the same time reveals a weekly pattern in RV. It also yields an interesting comparison with the HAR model, which is a simple way of accounting for long memory in volatility. Fitting the two component specification with leverage and a day of the week component is then carried out directly on RV with a Generalised Beta of the second kind (GB2) conditional distribution - equivalent to the estimation of a model for logRV with an Exponential Generalised Beta of the second kind (EGB2) conditional distribution, of which the normal distribution is a limiting case. The preliminary analysis of logRV also indicates heteroscedasticity in the residuals. The relationship between the GB2 and EGB2 distributions suggests that this heteroscedasticity may be due to a dynamic tail index in the GB2 model, and the DCS model is extended to allow for this possibility. Ultimately the forecasting power of the DCS model is compared with the HAR revealing similar forecasting performance besides its higher descriptive power.

CG624 Room H2 CONTRIBUTIONS IN ECONOMETRIC ANALYSIS OF THE BUSINESS CYCLE Chair: Caterina Liberati

C1605: **Modeling of economic and financial conditions for nowcasting and forecasting recessions: A unified approach**
Presenter: **Cem Cakmakli**, Koc University, Turkey
Co-authors: Hamza Demircan, Sumru Altug
A unified framework is proposed for the joint estimation of the indexes that can broadly capture economic and financial conditions together with their cyclical regimes of recession and expansion. Specifically, we utilize a dynamic factor model together with (Markov) regime-switching model parameters that exploit the temporal link between the cyclical behavior of economic and financial factors. This is achieved by constructing the cycle in the financial factor using the cycle in the economic factor together with phase shifts. The resulting framework allows the financial cycle to potentially lead/lag the business cycle systematically and exploits the information in economic and financial variables for estimation of both economic and financial conditions as well as their cyclical behavior efficiently. We examine the potential of the model using a mixed frequency ragged-edge dataset for Turkey. Comparison with conventional competitors reveals that the proposed specification provides precise estimates of economic and financial conditions and it delivers quite accurate probabilities of recessions. We further conduct a recursive real-time exercise of now/forecasting business cycle turning points. The results show convincing evidence of superior predictive power of our specification by signaling oncoming recessions (expansions) as early as 3.5 (3.4) months ahead of the actual realization.

C1261: **The switching skewness over the business cycle**
Presenter: **Stephane Lhuissier**, Banque de France, France
Motivated by the analysis of the evolution of the distribution of macroeconomic time series data over time, the aim is to develop and apply a Gibbs-sampler for autoregressive time series subject to regime switches in the tails of the distribution. More specifically, we consider the skew-normal distribution, in which the shape parameter is allowed to change over time according to a Markov-switching process. As an empirical illustration, we analyse the distribution of the growth rates of postwar U.S. real GDP, and find periodic shifts between a left and right-skewed distribution regime, with the former corresponding closely to NBER recession dates. Hence, more theorizing is needed to better understand the interaction between variation in tails and the business cycle.

C1551: **On the construction of composite economic indicators: The case of a new EU member state**
Presenter: **Boriss Siliverstovs**, Bank of Latvia, Latvia
Composite economic indicators (CEIs) for business cycle monitoring for a transition economy (Latvia) are constructed. The data are characterised by rather large revisions of official statistics (GDP growth) and time series properties of national economic indicators are dominated by the rather small number of observations pertaining to the great financial crisis. The lessons from our exercise is that when constructing composite economic indicators the informative content of data transformations of underlying individual economic/financial indicators varies with business cycle phases. We advocate that instead of focusing on one version of composite economic indicator based on a pre-selected set of transformations of individual time series, several versions of CEIs need to be monitored in order to be able to judge economic outlook in a credible and timely manner.

C1671: **Dominant U.S. manufacturing sectors: A factor model analysis**
Presenter: **Soroosh Soofi Siavash**, Bank of Lithuania, Lithuania
The focus is on the issue of identifying observed time series variables which serve as proxies for the factors underlying sectoral comovement. In the studies using factor model analysis, an approximate dynamic factor model fit sectoral data well. In multisector models with the input-output linkages, the macroeconomic fluctuations are viewed being primarily a result of shocks specific to the sectors which have an important role in supplying products to other sectors, or are of a great size. We use a factor method to investigate whether any observed time series variable of an individual sector serves as a factor proxy in large sectoral panels. We show that the method can identify the factors with a probability approaching unity when $N, T \to \infty$ even if the factors are relatively weak. In an application of the method to sectoral industrial production growth rates, we find that; (a) growth rates of a few heavy machinery and electrical equipment sectors serve as proxy for a factor, and (b) the sectors identified appear to have a key role in supplying capital products in U.S. economy, but have a moderate role in supplying intermediate products to others, and are of a moderate size.

Saturday 15.12.2018 08:45 - 10:05 Parallel Session F – CFE-CMStatistics

EI009 Room A0 ADVANCES IN EXTREME VALUE ANALYSIS Chair: Anna Kiriliouk

E0158: **The proportional tail framework for extreme quantile regression**
Presenter: **Clement Dombry**, Universite de Franche Comte, France
Co-authors: Benjamin Bobbia, Davit Varron
Extreme quantile regression is a long-standing issue in extreme value theory. The goal is to predict the quantile of order $1-p$ of the response $Y \in \mathbb{R}$ given covariates $X \in \mathbb{R}^d$ when $p = p(n)$ goes to zero as the sample size n goes to infinity. It has many applications, for instance in the context of risk management to assess the Value at Risk of the daily log-return of an asset given covariates that account for the market situation. The purpose is to present new results for extreme quantile regression in the proportional tail framework. This is closely related to the framework of heteroscedastic extremes, where the extremes of independent non-identically random variables are considered - the extremes depends on time through the so-called skedasis function σ. The main assumptions are that the response variable Y is heavy tailed and that the conditional tail function $\bar{F}_{Y|X=x}$ of Y given $X = x$ is asymptotically proportional to the unconditional one $\bar{F}_Y$. We present an analysis of the proportional tail framework based on coupling techniques and focus on properties of estimators of the extreme value index, the skedasis function and the extreme conditional quantiles.

E0159: **A nonparametric estimator of the extremal index**
Presenter: **Juan Juan Cai**, Delft University of Technology, Netherlands
Co-authors: Andrea Krajina
The extremal index is a number in the unit interval determining the amount of tail dependence in a sequence of stationary random variables. It connects the standard extreme value theory of an iid sample to the case where the independence assumption no longer holds. It describes the clustering behavior of exceedances of a high threshold. We show that the extremal index is determined by the stable tail dependence function. We illustrate this link on theoretical examples and develop a nonparametric estimator of the extremal index. We prove that the estimator is consistent and asymptotically normal under some mixing conditions. The simulation study shows that the estimator has good finite sample properties and with the real-data example we provide an interesting application.

E0575: **Testing the multivariate regular variation model**
Presenter: **Chen Zhou**, Erasmus University Rotterdam, Netherlands
Co-authors: Fan Yang, Chen Zhou, John Einmahl
A test for the multivariate regular variation model is proposed. The approach is based on testing whether the extreme value indices of the radial component conditional on the angular component falling in different subsets are at the same level. Combining the test on the constancy across different conditional extreme value indices with testing the regular variation of the radial component, we obtain the test for testing multivariate regular variation. Simulation studies demonstrate the good performance of the proposed tests. We apply this test to examine two datasets used in previous studies that are assumed to follow the multivariate regular variation model.

EO550 Room Aula C RECENT ADVANCES IN HIGH-DIMENSIONAL STATISTICS Chair: Yin Xia

E0818: **Clustering of high-dimensional Gaussian mixtures with EM algorithm and its optimality**
Presenter: **Jing Ma**, Fred Hutch Cancer Research Center, United States
Co-authors: Linjun Zhang, Tony Cai
Unsupervised learning is an important problem in statistics and machine learning with a wide range of applications. CHIME is presented, a procedure for clustering of high-dimensional Gaussian mixtures that is based on the EM algorithm and a direct estimation method for the sparse discriminant vector. Both theoretical and numerical properties of CHIME are investigated. We establish the optimal rate of convergence for the excess mis-clustering error and show that CHIME is minimax rate optimal. In addition, the optimality of the proposed estimator of the discriminant vector is established. The technical tools developed for the high-dimensional setting can also be used to establish the optimality of the clustering of Gaussian mixtures in the conventional low-dimensional setting. The merit of CHIME is illustrated in both simulated and real data settings.

E0820: **Tests for principal eigenvalues and eigenvectors**
Presenter: **Xinghua Zheng**, HKUST, China
Co-authors: Jianqing Fan, Yingying Li, Ningning Xia
CLTs are established for the principal eigenvalues and eigenvectors under a large factor model setting. As an application, we develop two-sample tests for difference in either the principal eigenvalues or principal eigenvectors. In particular, these tests can be used to detect structural breaks in large factor models. While there exist such tests, they can not distinguish between individual eigenvalues and/or eigenvectors. Our tests provide unique insights into the source of structural breaks.

E0827: **High-dimensional minimum variance portfolio estimation based on high-frequency data**
Presenter: **Yingying Li**, Hong Kong University of Science and Technology, Hong Kong
Co-authors: Tony Cai, Jianchang Hu, Xinghua Zheng
The aim is to study the estimation of high-dimensional minimum variance portfolio (MVP) based on high frequency returns which can exhibit heteroskedasticity and possibly be contaminated by microstructure noise. Under certain sparsity assumptions on the precision matrix, we propose an estimator of MVP and prove that our portfolio asymptotically achieves the minimum variance in a sharp sense. In addition, we introduce consistent estimators of the minimum variance, which provide reference targets. Simulation and empirical studies demonstrate that our proposed portfolio performs favorably.

E0859: **Pre-processing with orthogonal decompositions for high-dimensional explanatory variables**
Presenter: **Cheng Yong Tang**, Temple University, United States
It is well known that strong correlations between explanatory variables are problematic for high-dimensional regularized regression methods. Due to the violation of the irrepresentable condition, the popular lasso method may suffer from false inclusions of non-contributing variables. We propose preprocessing orthogonal decompositions (PROD) for the explanatory variables in high-dimensional regressions. The PROD procedure is constructed based upon a generic orthogonal decomposition of the design matrix. We investigate in detail three specific cases of the PROD: one by the conventional principal component analysis, one by a novel optimization incorporating the impact from the response variable, and one by random projections. We recognize that the PROD can be flexibly adapted taking multiple objectives into consideration such as avoiding increasing the variance of the resulting estimator while alleviating strong correlations between the explanatory variables. Extensive numerical studies with simulations and data analysis show the promising performance of the PROD in improving the performance of high-dimensional penalized regression. Our theoretical analysis also confirms its effect and benefit for high-dimensional regularized regression methods.

EO226 Room F1 CLUSTERING COMPLEX DATA: A BAYESIAN PERSPECTIVE Chair: Gary Rosner

E0780: **Clustering and predicting recurrent blood donations via donors' covariates**
Presenter: **Alessandra Guglielmi**, Politecnico di Milano, Italy
Blood is an important resource in global healthcare and therefore an efficient blood supply chain is required. Predicting arrivals of blood donors is fundamental since it allows for better planning of donations sessions. With the goal of characterizing behaviors of donors, we analyze gap times between consecutive blood donations. In order to take into account population heterogeneity we adopt a Bayesian model for clustering. In such a context, defining the model boils down to assign the prior for the random partition itself and to flexibly assign the cluster-specific distribution, since, conditionally on the partition, data are assumed iid within each cluster and independent between clusters. In particular, we drive the prior knowledge on the random partition by increasing the probability that two donors with similar covariates belong to the same cluster. The resulting model is a covariate-dependent nonparametric prior, thus departing from the standard exchangeable assumption. Specifically, we modify the prior on the partition prescribed by the class of normalized completely random measures by including in the prior a term that takes into account the distance between covariates. We fit our model to a large dataset provided by AVIS (Italian Volunteer Blood-donors Association), which is the largest provider of blood donations in Italy.

E0816: **Discovering interactions using covariate informed random partition models**
Presenter: **Fernando Quintana**, Pontificia Universidad Catolica de Chile, Chile
Co-authors: Garritt Page, Gary Rosner
Combination chemotherapy treatment regimens created for patients diagnosed with childhood acute lymphoblastic leukemia have had great success in improving cure rates. Unfortunately, patients prescribed these types of treatment regimens have displayed susceptibility to the onset of osteonecrosis. Some have suggested that this is due to pharmacokinetic interaction between two agents in the treatment regimen (asparaginase and dexamethasone) and other physiological variables. Determining which physiological variables to consider when searching for interactions in scenarios like these, minus a priori guidance, has proved to be a challenging problem, particularly if interactions influence the response distribution in ways beyond shifts in expectation or dispersion only. We propose an exploratory technique that is able to discover associations between covariates and responses in a very general way. The procedure connects covariates to responses very flexibly through dependent random partition prior distributions, and then employs machine learning techniques to highlight potential associations found in each cluster. We apply the method to data produced from a study dedicated to learning which physiological predictors influence severity of osteonecrosis multiplicatively.

E1031: **Normalized almost sure finite point processes for mixture models**
Presenter: **Raffaele Argiento**, University of Torino, Italy
Modelling via finite mixtures is one of the most fruitful Bayesian approach, particularly useful for clustering when there is unobserved heterogeneity in the data. The most popular algorithm under this approach is the reversible jump MCMC that can be nontrivial to design, especially in high-dimensional spaces. We will show how nonparametric methods can be transferred into the parametric framework. We first introduce a class of almost sure finite discrete random probability measures obtained by normalization of finite point processes. Then, we use the new class as mixing measure of a mixture model and derive its posterior characterization. The resulting class encompasses the popular finite Dirichlet mixture model. In order to compute posterior statistics, we propose an alternative to the reversible jump: borrowing notation from the nonparametric Bayesian literature, we set up a conditional MCMC algorithm based on the posterior characterization of the unnormalized point process. To discuss the performance of our algorithm and the flexibility of the model, we illustrate some examples on simulated and real data.

E1081: **Bayesian spatio-temporal clustering for areal data**
Presenter: **Annalisa Cadonna**, WU, Vienna University of Economics and Business, Austria
Co-authors: Alessandra Guglielmi, Andrea Cremaschi
Recent availability of mobile data provides researchers with datasets which are collected over time and on a huge spatial grid. The goal is the development of appropriate models and efficient algorithms for clustering of large to huge spatio-temporal datasets. Specific focus is placed on their potential application to clustering regions based on population density dynamics. In fact, large scale quantitative information on population density dynamics is of great interest to urban planners and city managers. We analyze high-dimensional areal data describing the use over time of the mobile-phone network in this area. The goal is to identify and cluster sub-regions of the metropolitan area of Milan which share similar characteristics along time in terms of population density dynamics, and to allow for the clustering to vary over time. Moreover, we would like to be able to detect isolated activities taking place in specific locations and times within the metropolitan area. To reach our goal, we perform Bayesian spatio-temporal clustering using a non-parametric approach based on a time-varying Dirichlet process. Preliminary results show time varying clustering, interpretable in terms of population density dynamic, such as weekly daily work activities, commuting, and big isolated events.

EC636 Room Aula Magna CONTRIBUTIONS IN HIGH-DIMENSIONAL STATISTICS Chair: Natalia A Stepanova

E1595: **Cluster-robust standard errors for linear regression models with many controls**
Presenter: **Riccardo D Adamo**, University College London, United Kingdom
The linear regression model is widely used in empirical economics to estimate the structural/treatment effect of some variable on an outcome of interest. Researchers often include a large set of regressors in order to control for observed and unobserved confounders. We develop inference methods for linear regression models with many controls and clustering. We show that inference based on the usual cluster-robust standard errors is invalid in general when the number of controls is a non-vanishing fraction of the sample size. We then propose a new clustered standard errors formula that is robust to the inclusion of many controls and allows us to carry out valid inference in a variety of high-dimensional linear regression models, including multi-way fixed effects panel data models and the semiparametric partially linear model. Monte Carlo evidence supports our theoretical results and shows that our proposed variance estimator performs well in finite samples.

E1581: **Revealing the joint mechanisms in traditional data linked with big data**
Presenter: **Niek de Schipper**, Tilburg University, Netherlands
Co-authors: Katrijn Van Deun
Recent technological advances have made it possible to study human behavior by linking novel types of data to more traditional types of psychological data, for example linking psychological questionnaire data with genetic risk scores. Revealing the variables that are linked throughout these traditional and novel types of data gives crucial insight in the complex interplay between the multiple factors that determine human behavior, e.g., the concerted action of genes and environment in the emergence of depression. Little or no theory is available on the link between such traditional and novel types of data, the latter usually consisting of a huge number of variables. The challenge is to select in an automated way those variables that are linked throughout the different blocks and this eludes current available methods for data analysis. To fill the methodological gap, we present a novel data integration method.

E1575: **Asymptotically minimax predictive density for sparse Poisson sequence model with different sample sizes**
Presenter: **Ryoya Kaneko**, The University of Tokyo, Japan
Co-authors: Keisuke Yano, Fumiyasu Komaki
There is growing demand for high-dimensional sparse count data. Sparsity in count data implies zero-inflation, that is, the situations where there exists an excess of zeros. In handling high-dimensional sparse count data, there often appears inhomogeneity in sample sizes of coordinates. To manage such sparse count data with different sample sizes, we introduce Poisson sequence models with different sample sizes under sparsity constraints on the parameter space, and consider predictive density estimation under Kullback-Leibler loss from a decision-theoretical point of view. We propose a Bayes predictive density that attains exact asymptotic minimax risk over sparse parameter space in Poisson sequence models with different sample sizes. The proposed predictive density has the merit of being adaptive to an unknown sparsity. We also apply the proposed method to real-world datasets and discuss its practical effectiveness.

E1532: **Semi-sparse PCA**
Presenter: **Nickolay Trendafilov**, Open University, United Kingdom
Co-authors: Lars Elden
It is well known that the classical exploratory factor analysis (EFA) of data with more observations than variables has several types of indeterminacy. We study the factor indeterminacy and show some new aspects of this problem by considering EFA as a specifiic data matrix decomposition. We adopt a new approach to the EFA estimation and achieve a new characterisation of the factor indeterminacy problem. A new alternative model is proposed, which gives determinate factors and can be seen as a semi-sparse principal component analysis (PCA). An alternating algorithm is developed, where in each step a Procrustes problem is solved. It is demonstrated that the new model/algorithm can act as a specific sparse PCA and as a low-rank-plus-sparse matrix decomposition. Numerical examples with several large data sets illustrate the versatility of the new model, and the performance and behaviour of its algorithmic implementation.

EC637 Room C1 CONTRIBUTIONS IN BAYESIAN METHODS Chair: Botond Szabo

E1429: **Bayes-Gaussian aggregation of a single set of forecasts**
Presenter: **Ville Satopaa**, INSEAD, France
A supra-Bayesian aggregator is developed that inputs a decision-maker's (DM) prior distribution of a continuous outcome and then updates that belief based on experts' point predictions. Given that the aggregator only inputs the DM's belief and the experts' predictions, it can be applied to an isolated prediction task without any past data. The underlying probability model is parametric and captures the experts' bias, accuracy, and dependence. An objective prior is developed for these behavioral parameters, and the resulting posterior is shown to be proper. The posterior is estimated with an off-the-shelf numerical procedure that scales well in the number of experts and does not require tuning. Its use is illustrated on real-world point predictions of human body mass and different economic indicators.

E1479: **Quantifying the causal effect of speed cameras using two-stage Bayesian bootstrap**
Presenter: **Prajamitra Bhuyan**, Imperial College London, United Kingdom
A causal doubly-robust (DR) approach combines propensity score (PS) and outcome regression (OR) models to give an average treatment effect estimator that is consistent under correct specification of either of the two component models. In this set-up, standard Bayesian methods are difficult to apply because restricted moment models do not imply fully specified likelihood functions. To avoid this difficulty, the existing methods are restricted to utilize full Bayesian features and involve frequentist estimates of the propensity score. As a result, these methods inherited some of the deficiencies involved in the frequentist DR approach under misspecification of component models and the estimate becomes biased. A two-stage Bayesian bootstrap approach is proposed which allows incorporation of prior information and uncertainty quantification associated with both OR and PS model. Simulations show that the approach performs well under various sources of misspecification of the outcome regression or propensity score models. The proposed method is used to quantify the effect of speed cameras on road traffic collisions in British cities.

E1645: **A Jacobian approach for the incidental parameter problem**
Presenter: **Guangjie Li**, Cardiff University, United Kingdom
The information orthogonal reparameterization method for the incidental parameter problem is studied and extended. We found that the properties of the estimators for the common parameters depend on the function on the right-hand side of the differential equation used to find the reparameterization. When the function is not linear in the incidental parameters, the Jacobian from the original incidental parameters to the new incidental parameters can be used as a bias-reducing rather than a bias-removing prior as in the static panel logit and probit models. When the function is linear, it is possible to obtain consistent estimators of the common parameters as in the panel linear autoregressive models with strictly exogenous regressors and predetermined regressors. When the regressors are strictly exogenous, though the information orthogonal reparameterization does not exist, the Jacobian can still be found. When the regressors are predetermined, though the Jacobian cannot be found, a related moment condition can be used to obtain the consistent estimators.

E1449: **Bayesian mixture item response modeling in the presence of noncompliers**
Presenter: **Kensuke Okada**, The University of Tokyo, Japan
Detecting noncompliers who do not follow questionnaire instructions is one of the major challenges in survey research. We propose a Bayesian mixture item response theory modeling approach to model and detect these noncompliers based on their answers to the questionnaire items. For this purpose, previously proposed latent variable models for response styles and for reverse-coded items were extended as a mixture model for both compliers and noncompliers. In order to fit the Bayesian model to the data, the Hamiltonian Monte Carlo algorithm was used to efficiently draw random samples from the joint posterior distribution. The proposed method was applied to a secondary dataset of psychological surveys. Our survey contained several instructional manipulation check items that served as indicators of noncompliance. The results showed that the proposed modeling approach provides better predictive fit and better interpretability than ordinary models that do not consider the effects of noncompliance. Our findings highlight the importance of taking the existence of noncompliers into account, particularly in this age of online surveys.

EC638 Room E1 CONTRIBUTIONS IN NON- AND SEMI-PARAMETRIC METHODS Chair: Philippe Lambert

E1201: **Multiplication-combination tests for incomplete paired data**
Presenter: **Lubna Amro**, Institute of Statistics, Ulm University, Germany
Co-authors: Markus Pauly, Frank Konietschke
Statistical procedures are considered for hypothesis testing of real valued functionals of matched pairs with missing values. In order to improve the accuracy of existing methods, we propose a novel multiplication combination procedure. Dividing the observed data into dependent (completely observed) pairs and independent (incompletely observed) components, it is based on combining separate results of adequate tests for the two sub datasets. The methods can be applied for parametric as well as semi- and nonparametric models and make efficient use of all available data. In particular, the approaches are flexible and can be used to test different hypotheses in various models of interest. This is exemplified by a detailed study of mean- as well as rank-based approaches. Extensive simulations show that the proposed procedures are more accurate than existing competitors. A real data set illustrates the application of the methods.

E1362: **Estimating a semiparametric additive model for discrete choice data using backfitting algorithm**
Presenter: **Patricia Gelin Doctolero**, University of the Philippines Diliman, Philippines
Co-authors: Erniel Barrios, Joseph Ryan Lansangan
Discrete choice models are estimated with the assumption of existence of a link function. To relax this assumption, a semiparametric additive model of the utility function is proposed. Quasi likelihood estimation is embedded into the backfitting algorithm and used in estimating the utility function. A simulation study is developed to evaluate the performance of the fitted model based on misclassification rate. The proposed model performed well in cases of linear and nonlinear nonparametric functions given all alternatives have balanced data allocation. Moreover, results showed that the model is robust to different magnitudes of misspecification error, and increases in sample size lead to slight increase in the predictive ability. Misclassification rates in validation data are also generally smaller than when using the usual generalized additive model.

E1435: **Adaptive estimation of semi-parametric partially linear predictive regression under heteroskedasticity**
Presenter: **Hsein Kew**, Monash University, Australia
Adaptive estimation is considered in semiparametric partially linear predictive regression model with unconditional heteroscedasticity of an unknown form. We develop an adaptive semiparametric estimator weighted by a non-parametric variance estimator. The weighted estimator is shown to deliver potentially large asymptotic efficiency gains over the conventional unweighted estimator. Monte Carlo simulations confirm this theoretical result. We implement the proposed estimation method by studying the in-sample predictability of US future stock returns using the commonly used financial variables as predictors.

E1465: **Screening biomarkers associated with individual treatment effect**
Presenter: **Shonosuke Sugasawa**, University of Tokyo, Japan
The development of molecular diagnostic tools to achieve precision medicine requires accurate screening biomarkers associated with individual treatment effect. Although several effective data analytic strategies have been proposed for this purpose, they have limitations when it comes to flexibly capturing the complex relationships between clinical outcome and possibly high-dimensional covariates. We employ semiparametric hierarchical mixture modeling and propose an effective method for screening biomarkers associated with individual treatment effect. We apply the proposed method together with some existing methods to simulated data set and real trial data.

EC644 Room H1 CONTRIBUTIONS IN APPLIED STATISTICS II Chair: Concepcion Ausin

E1318: **Estimating the dependence of sensitive responses of mixed types in randomized response models**
Presenter: **Thomas Chan**, Hong Kong University of Science and Technology, Hong Kong
Co-authors: Mike So, Amanda Chu
The aim is to estimate the dependence of sensitive responses of multiple types obtained from randomized response techniques. Although only the respondents know which question, either sensitive or unrelated, they answer, by dividing the whole sample into two parts, we propose to estimate the summary statistics based on the method of moment. This approach is applicable to mixed response type data, namely having dichotomous and Likert sensitive responses in the survey. With the method of moment estimates, we calculate the conditional mean of continuous sensitive responses given a discrete response and the partial correlations among continuous sensitive responses. As a medical application, we study the dependence structure among the responses of a survey concerning health and pressure on college students by our proposed approach.

E1587: **Bayesian hierarchical vine copula models for the analysis of glacier discharge**
Presenter: **Concepcion Ausin**, Universidad Carlos III de Madrid, Spain
Co-authors: Mario Gomez, Carmen Dominguez
Glacier discharge is the loss of liquid water produced by melting ice. The aim is to analyze the relationship among the glacier discharge and other meteorological variables such as temperature, humidity, solar radiation and precipitation. We propose a Bayesian hierarchical vine copula model where we assume that the dependence structure among variables is different in each season, although governed by common hyperparameters. Further, we also assume a hierarchical structure in each particular season such that the relationship among variables is different for the cases of positive and/or zero precipitation and/or discharge, but with common hyperparameters. Bayesian inference is implemented and compared using both ABC and MCMC techniques. The proposed methodology is applied to a real data set of glacier discharge measured in King George Island in the Antarctica.

E1653: **A Bayesian mixed multinomial logit model for partially microsimulated data on labor supply**
Presenter: **Consuelo Nava**, University of Aosta Valley, Italy
Co-authors: Cinzia Carota
The determinants of labor choices of females within couples are studied in the presence of partially microsimulated data which induce choice sets not identical for all agents under examination. Using traditional random utility models, two main assumptions are violated: the independence of irrelevant alternative and the homogeneity of available choice sets across decision makers. As an alternative, a Bayesian mixed multinomial logit model suitably adapted to describe labor choices made by females within couples is proposed. The constructed specification of random effects allows us to deal with both non-homogeneous choice sets and individual heterogeneity, also taking advantages of different (parametric and nonparametric, frequentist and Bayesian) clustering techniques. By comparing the results obtained under the proposed approach to the ones obtained considering a model without random effects, the relative explanatory ability of these models in the above-described scenario is discussed.

E1456: **Spatial mixed model for areal data on the simplex**
Presenter: **Agnese Maria Di Brisco**, University of Milano Bicocca, Italy
Co-authors: Sonia Migliorati
On March the fourth, elections took place in Italy for the two Chambers of Parliament. Many newspapers emphasized the victory of the 5 Star

Movement (5SM) and its unprecedented dominance in most of the southern regions of Italy. The aim is to analyze the electoral results through a rigorous statistical model to evaluate the presence and the possible impact of spatial structures. The response variable is the percentage of votes got by the 5SM in each electoral district. To cope with a bounded continuous outcome lying in the open interval (0,1), a mixture regression model is proposed based on a special mixture of two betas (referred to as flexible beta) sharing the same precision parameter but displaying two distinct component means subject to an inequality constraint. Advantages of this model are its many theoretical properties which are reflected in its computational tractability. Furthermore, the special mixture structure is designed to represent a wide range of phenomena (bimodality, heavy tails and outlying observations). The model is further extended by accounting for spatial correlation through random effects. Intensive simulation studies are performed to evaluate the fit of the proposed regression model. Inferential issues are dealt with by a (Bayesian) Hamiltonian Monte Carlo algorithm.

EC642 Room I1 CONTRIBUTIONS IN METHODOLOGICAL STATISTICS Chair: Yvik Swan

E1157: **Simex approach for bivariate random-effects meta-analysis of diagnostic accuracy studies**
Presenter: **Annamaria Guolo**, University of Padova, Italy

Meta-analysis represents a widely accepted approach for assessing the accuracy of a diagnostic test in distinguishing between diseased and nondiseased patients. The common bivariate random-effects meta-analysis model for inference on sensitivity and specificity of a test has a hierarchical structure describing the within-study sampling variability and the between-study variability arising from differences due, for example, to patients' characteristics. In addition, the model accounts for measurement error affecting the study specific estimation of sensitivity and specificity. Standard likelihood methods routinely used for inference are prone to several drawbacks, including the risk of unreliable conclusions in case of small sample size and substantial convergence issues. Simex, a simulation-based technique developed as correction strategy within the measurement error literature, is proposed as an alternative. The improvements of inferential conclusions based on Simex over the likelihood approach, mainly in terms of empirical coverage probabilities of confidence intervals, are shown under different scenarios, including varying sample size and correlation between sensitivity and specificity and in case of deviations from the normality assumption for the random-effects distribution. From a computational point of view, the application of Simex is appealing as it is neither involved nor suffering from the convergence issues affecting likelihood-based solutions.

E1469: **A wider class of estimators of positive normal means, individual and simultaneity**
Presenter: **Genso-Yuan Tsung Watanabe-Chang**, Mejiro University, Japan
Co-authors: Nobuo Shinozaki

While estimating a positive normal mean, θ, when variance is unknown, it has been shown that the Bayes estimator is a minimax and admissible estimator based on uniform prior on [0, ∞) of θ under squared error loss. We propose a generalized Bayes estimator based on gamma prior distribution for θ. The proposed generalized Bayes estimator includes the previous estimator and is an admissible estimator of positive normal mean. Based on the proposed generalized Bayes estimator, we consider the Stein-type estimator for estimation p unknown non-negative normal means, simultaneously, and give a sufficient condition for proposed Stein-type estimators dominate the generalized Bayes estimators under sum of squared error loss.

E1240: **Stochastic orders of transformed distributions**
Presenter: **Tommaso Lando**, VSB Technical University of Ostrava, Czech Republic

The aim is to compare, in terms of stochastic orders, pairs of distributions that can be obtained by the composition of a baseline distribution with a transformation (probability distortion) function. It can be seen that this is related to a comparison of transformed random variables. This approach may have application to parametric models and other fields.

E0883: **Discounted lifetime cost of post-retirement long-term care and annuity benefits under Markov mortality-morbidity models**
Presenter: **Colin Ramsay**, University of Nebraska-Lincoln, United States
Co-authors: Victor Oguledo

Suppose for individuals with health risk parameter θ, their health state can be modeled as continuous time Markov processes, $\{Z(t,\theta), t \geq 0\}$, with state space $S = \{1,2,\ldots,m\}$. Assume higher values of θ and S denote less healthy individuals and worse health states, respectively. There is one death state: m. Individuals get random health shocks then seek long term care. Let $\alpha > 0$ be the level quality of care received, with higher values of α denoting higher quality of care. Let $C_j(t,\theta,\alpha)$ denote the cost of care at time t for individuals in state j with risk parameter θ and quality of care α, with $C_j(t,\theta,\alpha)$ being non-decreasing in t, θ, and α. We assume the transition intensities of $Z(t,\theta)$ are time homogeneous and depend on the quality of care being provided. Using publicly available estimates of $C_j(t,\theta,\alpha)$ and transition intensities, we construct a time homogeneous Markov chain with transition intensities depending only on θ and α according to a quasi-frailty model. Transition probabilities are found using uniformization algorithms. We then determine the expected discounted lifetime cost of post-retirement long-term care and annuity benefits for various values of θ and α. For a given lifetime budget constraint, we determine the optimal value of α for given θ using the Nelder-Mead simplex algorithm.

EC640 Room L1 CONTRIBUTIONS IN MULTIVARIATE STATISTICS Chair: Luis Angel Garcia-Escudero

E1457: **Central limit theorem for Betti numbers in stochastic block models for random graphs**
Presenter: **Yang Di**, University of Nottingham, United Kingdom
Co-authors: Andrew Wood, Huiling Le, Karthik Bharath

In recent years there has been growing interest within statistics in topological aspects of random objects, one important direction being topological data analysis and the key concept of persistent homology. One interesting theoretical contribution is a central limit theorem (CLT) for Betti numbers in random clique complexes in Erdos-Renyi (ER) random graphs. The clique complex of an undirected graph is the simplicial complex formed by the sets of vertices in the cliques of the graph, a clique being a subgraph in which all edges are present. For each dimension, there is a certain range of probabilities where the Betti number is non-zero asymptotically almost surely in ER models. We consider a generalisation of the previous result to a CLT for Betti numbers in stochastic block model random graphs. The content of the result will be explained and the main ideas in the proof will be briefly outlined.

E1594: **Weighted energy score**
Presenter: **Xiaochun Meng**, University of Sussex, United Kingdom
Co-authors: James Taylor

Multivariate probabilistic forecasting is particularly intriguing and challenging due to its inherently complex nature and computational difficulty. For some applications, such as financial market risk assessment, a specific region is often of more importance than the whole distribution. To emphasise the region of interest for multivariate distributions, we propose the weighted energy score by generalising the existing energy score via threshold and quantile weight functions. The proposed weighted energy score is proper and provides useful insight into the evaluation of multivariate probabilistic forecasts. We use financial data to provide empirical support for the proposed weighted energy score.

E1633: **Bayesian minimax estimation for means in k sample problems**
Presenter: **Ryo Imai**, University of Tokyo, Japan

The simultaneous estimation of means of k multivariate normal populations is considered when one suspects that the k means are nearly equal. As an alternative to the preliminary test estimator based on the test statistics for testing hypothesis of equal means, we derive Bayesian and minimax estimators which shrink individual sample means toward a pooled mean estimator given under the hypothesis. It is shown that both the preliminary test estimator and the Bayesian minimax shrinkage estimators are further improved by shrinking the pooled mean estimator. The performance of the proposed shrinkage estimators is investigated by simulation. We will also discuss individual estimation of a mean of one sample and Efron-Morris type estimation of a mean matrix in the matrix normal model.

E1616: **MANOVA in block compound symmetry setting**
Presenter: **Ivan Zezula**, P.J. Safarik University, Slovakia

A multivariate linear normal model with block exchangeable covariance structure is considered. Such a structure can be naturally implied by the design of the experiment. Its main advantage is strongly reduced number of the second order parameters, which allows mean testing even with a small sample size. The problem of simultaneous testing of equality of mean values of several populations sharing the covariance structure will be considered.

EC634 Room M1 CONTRIBUTIONS IN FUNCTIONAL DATA ANALYSIS Chair: Sonja Greven

E1460: **Constrained LiNGAM approach for tensor data**
Presenter: **Ippei Takasawa**, Doshisha University, Japan
Co-authors: Kensuke Tanioka, Hiroshi Yadohisa

Independent Component Analysis (ICA) for three-way data has been proposed to analyze fMRI data in the domain of cognitive science. However, ICA cannot estimate causal relations. LiNGAM is one of the methods that can be applied to estimate causal relations among variables. Some extensions of LiNGAM and LiNGAM for tensor data have been proposed. Circumstances in which causal relations change based on the group of individuals and depending on the period during which data is observed when causal relations are estimated for three-way data do exist. In such circumstances, it is advantageous to reveal each groups distinct causal relations; however, current LiNGAM for tensor data cannot make these estimations. To overcome this problem, we propose a constrained LiNGAM approach for tensor data as an extension of LiNGAM for tensor data that enables both common causal relations and each groups distinct causal relations to be revealed. Using this method, common causal relations and the groups distinct relations can be estimated and interpreted. In particular this method is based on the concept of the three-way modeling and estimation of causal relations using LiNGAM. Furthermore, the proposed method is applied to real-world data and the result is evaluated.

E1681: **Handling missing scalar and functional data in integrative analysis with applications to mental health research**
Presenter: **Adam Ciarleglio**, George Washington University, United States

In mental health research, the number of clinical trials and observational studies that include multimodal neuroimaging is growing rapidly. Often, the goal is to integrate both the clinical and imaging data in order to address a specific research question. Functional data analytic tools for analyzing such data do exist and can perform well, but most/all methods assume that the data being analyzed are complete. Analyses that use these tools can be undermined by the fact that some proportion of both the clinical and imaging data may be missing for some study participants, often leading to data sets where only a small number of subjects have data available for all variables. We present approaches for imputation of missing scalar and functional data when the goal is to fit a scalar-on-function regression model for the purpose of either (1) estimating the association between a scalar outcome and a scalar or functional predictor or (2) developing a predictive model. We present results from a simulation study showing the performance of various imputation approaches with respect to fidelity to the observed data, estimation of the parameters of interest, and prediction. The proposed approaches are also illustrated using data from a placebo-controlled clinical trial assessing the effect of SSRI in subjects with major depressive disorder.

E0576: **Bandwidth selection of recursive nonparametric relative regression for independent functional data**
Presenter: **Yousri Slaoui**, University of Poitiers, France

New kernel regression estimators are proposed based on the minimization of the mean squared relative error. We study the properties of the proposed recursive estimators and compare them with the recursive estimators based on the minimization of the mean squared error. It turns out that, with an adequate choice of the parameters, the proposed estimators outperformed the recursive estimators based on the minimization of the mean squared error in some specific situations as the presence of outliers or when the response of the model is usually positive. We corroborate these theoretical results through a real chemometric dataset.

E1623: **Assessing diversity over time via functional data analysis and related tools**
Presenter: **Fabrizio Maturo**, University G. d Annunzio of Chieti Pescara, Italy
Co-authors: Francesca Fortuna, Tonio Di Battista

The problems of measuring and monitoring diversity have been widely discussed in recent decades; nevertheless, both theoretical perspectives and empirical results appear conflicting and inconsistent in ecology, management, business administration, social science, and other disciplines that deal with this concept that is multidimensional and multivariate in nature. The aim is to reconsider the problem of assessing diversity from a statistical perspective for solving some inconsistency of the classical diversity indices. We start from the concept of diversity profile and Hill's number integral function, which are monotone parametric functions useful to represent the concept of diversity of a specific community and develops a procedure for treating these curves in a functional context. Exploiting Hill's numbers, functional data analysis, and the monotone smoothing approach via B-splines, different functional instruments are proposed for assessing diversity changes over time and comparing the state of variety with respect to a benchmark. The main purpose is to provide human resources specialists, scholars, and ecologists with additional tools to improve the understanding of the dynamics of diversity within organizations, ecological communities, or other statistical samples where the role of diversity is worthy to be inspected.

EC643 Room P1 CONTRIBUTIONS IN STATISTICAL MODELLING Chair: Ori Davidov

E1369: **Adaptive maximum-likelihood-type estimation for discretely and noisily observed diffusion processes**
Presenter: **Shogo Nakakita**, Osaka University, Japan
Co-authors: Masayuki Uchida
An ergodic diffusion process is considered which is defined by the following stochastic differential equation and parametric inference with discrete and noisy observation. While the estimation with short-term high-frequency and noisy observation has been one of the central topics in the context of financial data analysis, the focus is on the long-term observation scheme to estimate the parametric drift term of the stochastic differential equation. To extract the state of the latent process, we compose the sequence of local means with respect to large numbers of partitions of observation. With this sequence, we construct two quasi-likelihood functions for diffusion and drift parameters respectively. It is possible to optimise these quasi-likelihoods separately, and this adaptive procedure has an advantage over the existent simultaneous one from the viewpoint of computational burden. We show that both the estimators have asymptotic properties such as consistency, asymptotic normality with different convergence rate and asymptotic independence.

E1397: **Multiple imputation and selection of ordinal level-2 predictors in multilevel models**
Presenter: **Leonardo Grilli**, University of Florence, Italy
Co-authors: Carla Rampichini, Omar Paccagnella, Maria Francesca Marino
A strategy is devised to handle ordinal level-2 predictors of a two-level random effect model in a setting characterized by two nontrivial issues: (i) level-2 predictors are severely affected by missingness; (ii) there is redundancy in both the number of predictors and the number of categories of their measurement scale. We tackle the first issue by considering a multiple imputation strategy based on information at both level-1 and level-2. We tackle the second issue by means of regularization techniques for ordinal predictors, also accounting for the multilevel data structure. The motivation arises from a case study at the University of Padua about the relationship between student ratings of a course and several characteristics of the course, including teacher feelings (ordinal predictors) and practices (binary predictors) collected by a specific survey with nearly half missing respondents.

E1597: **Multi-omics integrated analysis by means of graphical models**
Presenter: **Ilaria Bussoli**, University of Padova, Italy
Thanks to the advances in technology and bioinformatics of the last decade, a large amount of biological data coming from various experiments in metabolomics, genomics and proteomics is available. However, as it is the case with omics disciplines, the complex information content of such experiments introduces a challenge of its own, hence forming biologically relevant conclusions requires specialized forms of data analysis. One of the many problems in this area is how to integrate or model the information provided by differential gene expression and differential gene co-expression under different phenotypic subsets with the one delivered by transcriptome variations and DNA variant. To solve this, these biological records and the topology of the affected biological pathways are incorporated and modelled through conditional Gaussian regression models and chain graph models for mixed variables (both continuous and discrete). Topological properties of collapsibility and decomposability of the graph underlying each biological pathway are assessed to reduce dimensionality. On the obtained sub-regressions, initial testing on relevance of DNA variants (described as binary variables) on differential co-expression of genes is conducted. A direct visualisation of the direction and amplification of biological signals is implemented on the interested biological pathways.

E1500: **A general framework for prediction in multidimensional smoothing**
Presenter: **Alba Carballo**, Universidad Carlos III de Madrid, Spain
Co-authors: Maria Durban, Dae-Jin Lee
There are many situations in which prediction of new observations in the context of smoothing regression is needed (smooth time series, longitudinal models, etc.), but somehow this topic has been overlooked over the years. We propose a general framework for out-of-sample prediction in multidimensional smoothing. We explain how to construct basis and penalty matrices in a two-dimensional P-spline setting and show how prediction is carried out under different points of view: penalized regression, smooth mixed models and Gaussian process regression. We show the differences between the properties of the methodology used, and propose the use of constrained penalized splines to overcome the coherence problems that arise in the two dimensional case. One important application of the methods proposed is the forecasting of mortality rates from two-dimensional life's tables. We use data from the Human Mortality Data Base to illustrate our methodology.

EC639 Room Q1 CONTRIBUTIONS IN ROBUST STATISTICS Chair: Tim Verdonck

E0380: **Saddlepoint approximations for the distribution of some robust estimators of the variogram**
Presenter: **Alfonso Garcia-Perez**, UNED, Spain
The estimation of the variogram is a very important issue in geostatistics where a random variable is observed at some fixed locations. Matheron's estimator is the classical variogram estimator used in spatial statistical applications. There are also some robust versions, but their distributions (or some approximations of them) have not been studied in detail. Besides, although the sample size in geostatistics is not usually small, because the estimation of the variogram is made for each lag h, the number of observations used in each of these estimations, could be small. Hence, a saddlepoint approximation for the distribution of the variogram estimator should be suitable. We obtain a von Mises plus saddlepoint approximation for the distribution of these estimators, assuming that the sequence of the observations verifies the intrinsic stationarity property and that they follow a model distribution near to the normal; specifically a contaminated normal model. The accuracy of these approximations is illustrated with some Monte Carlo experiments. The approximation so obtained is used to analyze the robust properties of several variogram estimators. We shall also use it to test the variogram model and to analyze the required independence of the transformed observations used in the saddlepoint approximation.

E1325: **Robust regression with compositional covariates**
Presenter: **Aditya Mishra**, Flatiron Institute, Simons Foundation, United States
Co-authors: Christian Lorenz Mueller
Recent advances in low-cost metagenomic and amplicon sequencing techniques enable routine sampling of environmental and host-associated microbial communities across different habitats. The data produced by these large-scale surveys typically comprise relative abundances (or compositions) of microbial taxa at different taxonomic levels. To investigate the dependency of additional covariate measurements such as metabolites or host phenotypes on the microbial compositions we introduce a general robust regression framework for compositional data. We propose a novel log-contrast regression model with mean shift parameters that allows the identification of sample outliers and maintains sub-compositional coherence with respect to the associated phylogenetic tree. The model is estimated using a sparse penalized regression approach that simultaneously enforces sparsity in mean shift and covariate parameters. We demonstrate the superiority of our approach using a wide range of synthetic simulation scenarios and infer novel associations between body mass index measurements and human gut microbes on a large public collection of human gut microbiome data.

E1552: **Robust measurement errors method**
Presenter: **Fatemah Alqallaf**, Kuwait University, Kuwait
Co-authors: Claudio Agostinelli
The Simulation-Extrapolation methods (SIMEX) is widely used in presence of measurement errors. Unknown parameters and regression coefficients are estimate using algorithms based on generalized least squares, and the method, as a whole, is very sensitive to outliers. To overcome this problem, we propose a robust technique based on robust regression methods, such as MM-estimators. We focus our attention on the estimation of the correlation coefficient and on the coefficient of determination of the regression. Illustrations based on Monte Carlo simulations and real examples are provided.

E1531: **Depth for curve data and applications**
Presenter: **Myriam Vimond**, ENSAI, France
Co-authors: Myriam Vimond, Pavlo Mozharovskyi, Pierre Lafaye de Micheaux
Statistical data depth is defined as a function that determines centrality of an arbitrary point with respect to a data cloud or to a probability measure. During the last decades, this seminal idea of data depth evolved to a powerful machinery proving to be useful in various fields of science. Recently, extending the notion of data depth to the functional setting attracted a lot of attention among theoretical and applied statisticians. We go further and suggest a notion of data depth suitable for data represented as curves, or trajectories, which inherits both Euclidean-geometry and functional properties while overcoming certain limitations of the previous approaches. We show that our curve's depth satisfies theoretical requirements of general depth functions that are meaningful for trajectories. We apply our methodology to diffusion tensor brain images and also to pattern recognition of hand written digits and letters.

EG004 Room Aula 4 CONTRIBUTIONS IN BOOTSTAP FOR TIME SERIES Chair: Soumendra Lahiri

E1232: **Bootstrap-based inference for sparse high-dimensional time series models**
Presenter: **Jonas Krampe**, University of Mannheim, Germany
Co-authors: Jens-Peter Kreiss, Efstathios Paparoditis
Fitting sparse models to high dimensional time series is an important area of statistical inference. We consider sparse vector autoregressive models and develop appropriate bootstrap methods to infer properties of such processes, like the construction of confidence intervals and of tests for individual or for groups of model parameters. The bootstrap methodology generates pseudo time series using a model-based bootstrap procedure which involves an estimated, sparsified version of the underlying vector autoregressive model. Inference is performed using so-called de-sparsified or de-biased estimators of the autoregressive model parameters. We derive the asymptotic distribution of such estimators in the time series context and establish asymptotic validity of the bootstrap procedure proposed for estimation and, appropriately modified, for testing purposes. In particular, we focus on testing that a group of autoregressive coefficients equals zero. The theoretical results are complemented by simulations which investigate the finite sample performance of the bootstrap methodology proposed. A real-life data application is also presented.

E1561: **A forecasting-EVT method**
Presenter: **Clara Cordeiro**, CEAUL and FCT, UALg, Portugal
Co-authors: Manuela Neves
Statistical analysis of extreme values has been in the spotlight in studies of flooding events, heatwaves, hurricanes, sea level rise, among many others. This type of events has promoted the development of special statistical methodologies for its study, understanding and control, whenever possible. These events present unique statistical challenges and require characterizing adequately the tail of the distribution of the quantity of interest. Extreme value theory (EVT) is the branch of the statistics that has been used to model such type of data. Well-known forecasting methods do not capture so well these events what is reflected in its extrapolation. A forecasting-EVT method, which applies a forecasting procedure to the time series and models the residuals through an adequate EVT distribution is proposed. Moreover, the extreme distribution of residuals is estimated and bootstrap estimators of the shape parameter will be used to improve the modelling of the tail behaviour of the residuals distribution. This is a preliminary work presenting a contribution to model extremes of a time series.

E1360: **Sieve-based test for cointegration**
Presenter: **Jan Andrew Reforsado**, University of the Philippines Los Banos, Philippines
Co-authors: Erniel Barrios, Joseph Ryan Lansangan
Cointegration testing is an important aspect of modeling in nonstationary time series data to avoid the possibility of observing spurious relationship among variables. Existing tests for cointegration usually exhibit less optimal behavior specially for short time series data. A vector error correction model is estimated through the backfitting algorithm, the fitted model is used in replicating the data through sieve bootstrap. The empirical distribution of eigenvalues from the lagged error correction matrix generated from the data replicates are used in testing for cointegration. Simulation study shows that the proposed nonparametric test yields size and power at least comparable to some well-known tests for cointegration.

E1381: **Unit root test in a semiparametric model**
Presenter: **Sarah Bernadette Aracid**, University of the Philippines Los Banos, Philippines
Co-authors: Erniel Barrios, Joseph Ryan Lansangan
Presence of unit root in time series data is implicated in the persistent effect of random shocks in the behavior of a model, leading most unit root tests to be incorrectly-sized or have low power or both. A nonparametric test for the presence of unit root is proposed. To better understand the instance where unit root occurs, hence, mitigate the possible problem of present unit root tests, it is assumed that another time series x_t possibly affect the target time series y_t in addition to the autocorrelation dynamics. A nonparametric effect of x_t can spare the autocorrelation structure from further contaminations, hence, the test can characterize presence of unit roots in y_t easily. Simulation study showed that the proposed test yields better size and power compared to some popular tests for unit root.

EG569 Room Aula B CONTRIBUTIONS IN COVARIANCE MATRICES Chair: Dongchu Sun

E1195: **Existence and uniqueness of maximum likelihood estimators of Kronecker product covariances**
Presenter: **Satoshi Kuriki**, The Institute of Statistical Mathematics, Japan
Co-authors: Mathias Drton
Consider iid samples from a vector-valued Gaussian distribution. When the covariance matrix of the Gaussian distribution has no structure, the MLE uniquely exists with probability one if and only the sample size n is equal to or greater than the size of the covariance matrix. However, this is not the case where the covariance matrix has a structure and is described with a fewer number of parameters. We consider the case where the covariance matrix is the Kronecker product of two matrices ($m_1 \times m_1$ and $m_2 \times m_2$ matrices). We show that the existence and the uniqueness of the MLE are characterized by a rank of an $m_1 \times m_2 \times n$ tensor. In particular, when $n = 2$, the problem can be reduced by Kronecker's canonical form of two matrices. The tensor rank is given explicitly as the solution of an integer programming. The Groebner basis computation is also useful when m_1 and m_2 are small.

E1422: **Lugsail lag windows and their application to MCMC**
Presenter: **Dootika Vats**, University of Warwick, United Kingdom
Co-authors: James Flegal
Lag windows are commonly used in the time series, steady state simulation, and Markov chain Monte Carlo literature to estimate the long range variances of estimators arising from correlated data. We propose a new lugsail lag window specially designed to yield biased from above estimators for the long range variances. We use this lag window for batch means and spectral variance estimators in Markov chain Monte Carlo simulations and establish conditions ensuring strong consistency and mean square consistency. Further, we calculate the bias and variance of lugsail estimators and demonstrate there is little loss compared to other estimators. Finally, we study the finite sample properties of lugsail estimators in various examples.

E1512: **Methods for improving background error covariance matrix rebuild in data assimilation**
Presenter: **Sibo Cheng**, Electricite de France, France
Co-authors: Jean-Philippe Argaud, Didier Lucor, Angelique Poncot, Bertrand Iooss
A recurrent obstacle in data assimilation is the lack of information for background error covariance modeling. In the case of meteorology, the background covariance is often estimated from an observations ensemble or forecast differences. However, for many industrial fields, the modeling remains highly empirical relying on some form of expertise and physical constraints enforcement in the absence of historical observations/predictions. The Desroziers *a posteriori* tuning algorithm of 2001 is well known in variational methods for adjusting the ratio between background and observation error covariance matrices. We have developed two novel sequential adaptive methods: **CUTE**(Covariance Updating iTerative mEthod) and **PUB**(Partially Updating BLUE method) for building background error covariance matrices in order to improve the assimilation result under the assumption of a good knowledge of the observation error covariances. We have compared these two methods with the Desroziers approach in a twin fluid mechanics experiment framework together with a linear observation operator. The optimality in terms of assimilation errors is similar for all three methods. However, experiments show that the two new methods own a non-negligible advantage concerning correlation rebuild under the hypothesis that the background error is dominant over the one of observations.

E1612: **Likelihood ratio test for the double level compound symmetric structure**
Presenter: **Vusi Bilankulu**, University of Pretoria, South Africa
Co-authors: Andriette Bekker, Carlos Coelho
It is shown how from an adequate decomposition of the null hypothesis of the double level block compound symmetric covariance structure it is easy to obtain the expression for the likelihood ratio test (l.r.t.) statistic to test this hypothesis, as well as the general expression for its moments. From this expression it is then possible to identify the structure of the exact distribution and to obtain the characteristic function of the negative logarithm of the l.r.t. statistic. Then, from an adequate decomposition of this characteristic function it is possible to identify the distribution components that will be left untouched and those that will have to be asymptotically approximated. Based on this identification, near-exact distributions are then built. They are shown to be asymptotic both for increasing sample sizes, as well as for increasing numbers of variables and increasing numbers of groups of variables, while common asymptotic distributions are only asymptotic for sample size. Moreover, these near-exact distributions show very good performances for very small samples, enabling us to obtain very good approximations for situations where being p the number of variables involved and n the sample size, p/n approaches 1 from below.

EG163 Room Aula A CONTRIBUTIONS IN TIME SERIES I Chair: Holger Dette

E1308: **Estimating long run effects in models with cross sectional dependence**
Presenter: **Jan Ditzen**, Heriot-Watt University, United Kingdom
It is shown how to estimate long run coefficients in a dynamic panel with heterogeneous coefficients and common factors and a large number of observations over cross-sectional units and time periods. The common factors cause cross-sectional dependence which is approximated by cross-sectional averages. Heterogeneity of the coefficients is accounted by taking the unweighted averages of the unit specific estimates. Three different models to estimate long run coefficients are considered, a simple dynamic model, an error correction model and an ARDL model. It is explained how to estimate all three models and estimation results are compared by simulation. Further emphasis is put on estimating the standard errors of the long run coefficients. Estimated standard errors obtained by the delta method and bootstrapped standard errors are compared.

E1346: **Testing for long memory in panel random-coefficient AR(1) data**
Presenter: **Anne Philippe**, Nantes, France
Co-authors: Remigijus Leipus, Vytaute Pilipauskaite, Donatas Surgailis
It is well known that random-coefficient AR(1) process can have long memory depending on the index β of the tail distribution function of the random coefficient, if it is a regularly varying function at unity. We discuss the estimation of β from panel data comprising N random-coefficient AR(1) series, each of length T. The estimator of β is constructed as a version of the tail index estimator applied to sample lag 1 autocorrelations of individual time series. Its asymptotic normality is derived under certain conditions on N, T and some parameters of our statistical model. Based on this result, we construct a statistical procedure to test if the panel random-coefficient AR(1) data exhibit long memory. A simulation study illustrates finite-sample performance of the introduced estimator and testing procedure.

E1519: **Estimating non-stationary common factor: Implications for risk sharing**
Presenter: **Pilar Poncela**, JRC, Italy
Co-authors: Esther Ruiz, Francisco Corona
The finite sample properties of alternative factor extraction procedures in the context of non-stationary Dynamic Factor Models (DFMs) are analyzed and compared. On top of considering procedures already available in the literature, we extend the hybrid method based on the combination of principal components and Kalman filter and smoothing algorithms to non-stationary models. We show that, if the idiosyncratic noises are non-stationary, procedures based on extracting the factors using the nonstationary original series work better than those based on differenced variables. We apply the methodology to the analysis of cross-border risk sharing fitting non-stationary DFM to aggregate Gross Domestic Product and Consumption of a set of 21 industrialized countries from the Organization for Economic Co-operation and Development (OECD). The goal is to check if international risk sharing is a short or long-run issue.

E1602: **Multiscale asymptotics and stationarity test for stable locally stationary processes**
Presenter: **Alessandro Cardinali**, University of Plymouth, United Kingdom
Three main contributions to the study of multiscale locally stationary processes are provided. We first establish, assuming a (possibly nongaussian) stable marginal process distribution, the asymptotic independence and weak convergence for the multiscale periodogram and some related functionals. We then use this theoretical framework to propose a nonparametric approximation method for the unknown distribution function of unobservable LS innovations. We address this task by introducing a new method to produce pseudo innovations whose empirical CDF can be used to approximate the theoretical CDF of unobservable innovations. Finally, we use the above frameworks to derive a nonparametric bootstrap stationarity test for multiscale LS processes. The finite sample properties of this test are assessed through simulations showing that our method

successfully controls rejection rates for processes having either Gaussian or nongaussian innovations. An empirical analysis based on exchange rates returns shows that, when used on a rolling window, our test can also help to successfully identify well known economic shocks.

EG053 Room B1 CONTRIBUTIONS IN LATENT VARIABLE MODELS AND GRAPHICAL MODELS Chair: Wicher Bergsma

E0700: **Predicting trends of institutional confidence through a hidden Markov model with survey weights and missing responses**
Presenter: **Fulvia Pennoni**, University of Milano-Bicocca, Italy
Co-authors: Ewa Genge

A statistical methodology is proposed for the analysis of a latent concept which is fluctuating over time such as the perceived trust towards financial and political institutions. We conceive confidence as a mental unobservable feature of each person which is related to the observed time-varying and time-fixed covariates. We model the uncertainty in the response through a hidden Markov models, and we account for longitudinal survey weights, as well as missing responses when survey data are available. We estimate the model parameters by a weighted log-likelihood which is maximized by the expectation-maximization algorithm in order to find hidden clusters of people with the same perceptions towards the institutions. We allocate each individual according to the Viterbi algorithm applied to the posterior probabilities. By considering the Polish society we find four hidden groups of Poles: discouraged, with no opinion, with selective trust and with fully trust towards institutions. We predict an increasing tendency to choose the institutions to support.

E1611: **Scale-invariant estimations for the factor analysis model based on its geometric structure**
Presenter: **Michiko Okudo**, The University of Tokyo, Japan
Co-authors: Fumiyasu Komaki

Factor analysis is an important tool of multivariate analysis, especially in psychology. Factor analysis models are latent variable models, and the observation is divided into two parts, "common factor" and "specific factor". The maximum likelihood estimation of these models can be difficult when the dimension of common factor space is high and Bayesian estimation methods have been studied. We propose new priors for factor analysis models which are invariant under scalings of observations. These priors take advantage of the model manifold's geometric structure.

E1483: **Identifiability of discrete Bayesian network with a latent source**
Presenter: **Hiroaki Naito**, Doshisha University, Japan
Co-authors: Hisayuki Hara

Identifiability of discrete graphical models defined by directed acyclic graphs with a latent source is discussed. In this case, identifiability of parameters is not trivial. The problem reduces to whether the parametrization map is generically finite-to-one or not. It is well known in algebra that there exist computational algebraic algorithms to detect if the parametrization map is finite-to-one or not. However, the computational cost of these algorithms is quite high even for moderate-sized models. As a preceding study, for Gaussian graphical models with a latent source, some useful sufficient conditions have been derived. On the other hand, for discrete graphical models with a latent source, the identifiability of all models with up to four observable variables has been investigated. We apply the results for the Gaussian model and derive some sufficient condition for binary Bayesian network models to be identifiable even for larger models with more than five observable variables. We also provide a useful algorithm to detect the identifiability of a given model within polynomial time.

E1628: **A novel computationally tractable algorithm for discovering probabilistic graphical models in high-dimensional data**
Presenter: **Edith Alice Kovacs**, University of Debrecen, Hungary
Co-authors: Noemi Horvath, Roland Molontay

Revealing the dependence structure in high dimensional distributions has attracted a lot of research interest recently. We introduce a procedure that aims to reduce redundancy between random variables by exploiting conditional independences using only low-dimensional marginal probability distributions in the approximation task. The basis is a previous method and its generalization - the cherry tree. The first model uses only two-dimensional marginal probability distributions in the approximation of the high dimensional probability distributions, while the latter uses typically three-, four- or five-dimensional marginals. It has been shown that the cherry tree probability distributions give better approximation than the Chow-Liu algorithm does. However, fitting cherry tree probability distributions to the sample data is computationally much more expensive, in high dimensions (greater than 50) it may even be intractable. A computationally more tractable method for modeling probability distributions by cherry trees will be presented. We prove that the new algorithm also provides a better approximation than Chow-Liu does. We also give some numerical results to illustrate the effectiveness of our new method.

EG515 Room D1 CONTRIBUTIONS IN NONPARAMETRIC REGRESSION Chair: Juan-Carlos Pardo-Fernandez

E1305: **An independence test for a nonparametric random effects meta-regression model**
Presenter: **Daniel Gaigall**, Leibniz University Hannover, Germany

A nonparametric generalization of a random effects meta-regression model frequently used in life sciences is considered. For testing goodness-of-fit for the regression function or testing independence of input and between study variation noise, We apply the Hoeffding-Blum-Kiefer-Rosenblatt independence criterion to pairs of input and residuals. It turns out that the test statistic has the well-known distribution free limiting null distribution of the classical criterion. Related quantiles are available as critical values. Properties of the test statistic under alternatives are pointed out as well. A permutation procedure is a second option to obtain critical values. Simulations investigate size and power of both tests for small and moderate sample sizes. Application to real data from clinical trials illustrates how the tests work in practice.

E1326: **The de-biased group lasso estimation for varying coefficient models**
Presenter: **Toshio Honda**, Hitotsubashi University, Japan

There has been a lot of attention on the de-biased or de-sparsified Lasso since it was proposed in 2014. The Lasso is very useful in variable selection and obtaining initial estimators for other methods in high-dimensional settings. However, it is well-known that the Lasso produces biased estimators. Therefore several authors simultaneously proposed the de-biased Lasso to fix this drawback and carry out statistical inferences based on the de-biased Lasso estimators.The de-biased Lasso procedures need good estimators of high-dimensional precision matrices for bias correction. Thus the research is almost limited to linear regression models with some restrictive assumptions or generalized linear models with stringent assumptions. To our knowledge, there are a few papers on linear regression models with group structure, but no result on structured nonparametric regression models such as varying coefficient models. We apply the de-biased group Lasso to varying coefficient models and closely examine the theoretical properties and the effects of approximation errors involved in nonparametric regression.

E1566: **Nonparametric regression estimation in chain graph models**
Presenter: **Mate Baranyi**, Institute of Mathematics, Budapest University of Technology and Economics, Hungary
Co-authors: Marianna Bolla

It is known that a regression graph with a chordal graph for the context variables can be oriented to be Markov equivalent to a DAG on the same skeleton if and only if it does not contain any chordless collision path in four nodes. Constructions for such a DAG and applications of

linear, linearized, and logistic regression for prediction along the paths were intensively studied. A nonparametric regression method, using local averaging estimators, is introduced for prediction based on a complete sample. The method makes it possible to perform nonparametric regressions recursively along the DAG, irrespective of the type of the context and response variables. Hence, predictions for the response variables of new-coming cases can be done in the possession of the values of their context variables only. The technique can be extended to undirected graphical models on a chordal graph, where the prediction goes from the separators to the residuals of the cliques ordered in a junction tree structure. We prove consistency under very general conditions on the distribution and the selected kernel. An application to sociological data is also presented.

E1294: **Faster computationally approximation for comparing the error distributions in nonparametric regression**
Presenter: **Gustavo Rivas**, National University of Asuncion, Paraguay
Co-authors: Maria Dolores Jimenez-Gamero
Several procedures have been proposed for testing the equality of error distributions in two or more nonparametric regression models. We deal with methods based on comparing estimators of the cumulative distribution function (CDF) of the errors in each population to an estimator of the common CDF under the null hypothesis. The null distribution of the associated test statistics has been approximated by means of a smooth bootstrap (SB) estimator. The proposal is to approximate their null distribution through a weighted bootstrap. It is shown that it produces a consistent estimator. The finite sample performance of this approximation is assessed by means of a simulation study, where it is also compared to the SB. From a computational point of view, the proposed approximation is more efficient than the one provided by the SB.

EG145 Room G1 CONTRIBUTIONS IN SAMPLING AND DESIGN OF EXPERIMENTS Chair: Helmut Waldl

E1419: **Statistical properties of sub-cohort selection when testing interactions in biomarker studies**
Presenter: **Leslie McClure**, Drexel University, United States
Co-authors: Leann Long, Stephanie Tison, Suzanne Judd, George Howard, Mary Cushman
The Reasons for Geographic And Racial Differences in Stroke (REGARDS) study is a cohort of over 30,000 participants concerned with understanding racial and regional disparities in cardiovascular disease risk factors and stroke in the US. When examining the impact of biomarkers on cardiovascular disease risk factors, and particularly the differential effect of biomarkers by race, it is not financially feasible to assay these biomarkers in all participants. One strategy is to measure the biomarkers in a sub-sample of the cohort, but it isn't clear how to best choose the sample. We assessed different approaches to selecting the sub-cohort, in order to maximize power to detect interactions, while minimizing bias and maximizing the coverage of the 95% confidence interval. We simulated sub-samples of $n = 4000$, with characteristics of the participants based on the REGARDS cohort, to estimate these operating characteristics. We considered 3 assumptions: simple random sampling, sampling with equal allocation across race, and sampling with equal allocation across sex-race groups, and compared the statistical properties observed to those when using the full cohort. We will discuss the results of our simulations and practical implications for implementing the sampling and subsequent analyses.

E1298: **On the appropriate rank-based sampling scheme for estimating the inequality indices**
Presenter: **Najmeh Nakhaei Rad**, Mashhad Branch, Islamic Azad University, Mashhad,Iran, Iran
Co-authors: Mahdi Salehi
Estimating the inequality indices is a great interdisciplinary topic between statisticians and economists. Moreover, various indices have been proposed in order to reflect different properties of the income inequality. The simple random sampling (SRS) plan is employed as the most commonly tool for collecting the income data, while it has been shown that the ranked set sampling (RSS) is significantly better than SRS when the population follows the well-known income distributions. However, there are some simplified versions for RSS which can be examined for this purpose as well. We answer to this question: "among the SRS, RSS and its competitors, which scheme results a more appropriate estimator for each income inequality index?". A simulation study as well as analyzing a real data set, yields a reasonable result for this question.

E1361: **A semiparametric mixed analysis of covariance model for a crossover design with carryover effects**
Presenter: **Leonard Allan Almero**, University of the Philippines Los Banos, Philippines
Co-authors: Erniel Barrios, Joseph Ryan Lansangan
A semiparametric mixed analysis of covariance model for a crossover design with carryover effects is postulated. The responses are adjusted for covariate effect through a nonparametric function of the covariates. To estimate the model, a hybrid of restricted maximum likelihood estimation and smoothing splines regression is imbedded into a backfitting algorithm. A bootstrap-based test is then developed for testing differences in treatment means for fixed effects and/or significance of variance components for random effects. Simulation studies indicate that for a random effects model, the bootstrap-based test for variance components is correctly-sized. The test is also powerful and relatively robust to the hypothesized magnitude of variance. For a fixed effects model, the bootstrap-based test performs relatively better than the ordinary analysis of covariance when under 10% mean effect differences. Furthermore, for a model with either a fixed effects or random effects, the test remains advantageous over ANCOVA in the presence of misclassification error and in non-normal error on unbalanced data.

E1637: **Misspecified covariance structure and optimal designs for prediction**
Presenter: **Helmut Waldl**, Johannes Kepler University Linz, Austria
Modeling spatial or spatio-temporal data requires the choice of a (spatio)-temporal covariance function. Assumptions such as isotropy, stationarity or separability are usually used to make parameter estimation and prediction easier. Second-order stationarity, for instance, makes it possible to parametrize a covariance function with just a few parameters that may determine even high dimensional covariance matrices totally. In doing so, we will never use the correct covariance matrix for prediction. Especially when we are looking for good or optimal designs for prediction with respect to an arbitrary criterion, a misspecified covariance matrix may have severe impact on the quality of prediction of the seemingly optimal design. We compare the performance of different commonly used covariance structures for spatial models based on simulated data. Surprisingly even correlation functions enjoying the good reputation of being robust against model misspecification yield suboptimal prediction.

CO542 Room Q2 CRYPTOCURRENCY Chair: Ostap Okhrin

C0450: **A cross section of expected cryptocurrency returns based on continuous betas**
Presenter: **Tony Klein**, Queen's University Belfast, United Kingdom
Co-authors: Simon Trimborn, Thomas Walther, Christoph Wegener
The aim is to analyze the application of recent advances in asset pricing to a cross section of over 1000 cryptocurrencies and a novel data set of 5-min intra-day prices. By utilizing high-frequency price data, we derive continuous and jump betas which are applied to further dissect the market risk premiums and its transmission channels. When comparing the two betas and the standard CAPM, we conjecture that most of the market premium can be explained by the discontinuous components given the nature of the cryptocurrency market. As explosive behavior is a fairly common occurrence in cryptocurrency markets, we particularly focus on drawing inference on the betas under explosiveness.

C0524: **Discover regional and size effects in global bitcoin blockchain via sparse-group network autoregressive modeling**
Presenter: **Simon Trimborn**, National University of Singapore, Singapore

Bitcoin blockchain has been continuously growing to a global network with millions of accounts since its creation in 2009. The dynamics of the transaction interactions reflects virtual funds movements of the Bitcoin economy. It also provides insight into the inherent risk in the Bitcoin network at a global level. We propose a Sparse-Group Network AutoRegressive (SGNAR) model to describe the essential dynamic dependence structure of Bitcoin. Our study considers up-to-date Bitcoin blockchain, from February 2012 to July 2017, with all the transactions classified into 60 groups according to region and transaction size. We develop a regularized estimator for large-dimensional dynamic network with two-layer sparsity, which enables discovering active groups with influential impact on the global Bitcoin transactions and demonstrate dynamic evolution phases of the Bitcoin network. In particular, large investors from North America and medium sized users from Europe influence the network in the last year, while previously no network connectivity was observed. It follows the inherent risk, defined as the risk of the Bitcoin network to fail, shrank lately compared to all years up to 2015.

C0565: **Estimating higher distribution moments with high frequency data**
Presenter: **Manuel Schmid**, TU Dresden, Germany

In standard return modelling approaches, returns are often assumed to follow a normal distribution. This assumption implies zero skewness as well as a zero excess kurtosis. Both of these implications do not correspond to empirical observation and eventually lead to problems e.g. in financial risk management. On the other side, the typical non-parametric estimation of these values requires a huge amount of data to be reliable. For this reason, it is advisable to exploit the availability of high frequency data and construct estimators in the fashion of the well-known realized variance. A previous estimation approach is extended to non-martingale price processes. On the basis of Monte Carlo simulations, we show that our estimators are unbiased and consistent when the underlying price process can be modelled as a stochastic volatility jump diffusion process. Distribution properties of the estimators are discussed.

C1440: **On cryptocurrency**
Presenter: **Ostap Okhrin**, Dresden University of Technology, Germany

What is cryptocurrency? What does the Block-Chain mean? Is the BitCoin the same as Block-Chain? Will we have in the future true Money? Is cryptocurrency something only for criminals? Does it make sense to invest in cryptocurrency? We make a small journey from the origins of cryptocurrency to its current state. We will do mining together with the participants, and we will create our first BitCoin wallet.

CC650 Room H2 CONTRIBUTIONS IN COMPUTATIONAL ECONOMETRICS Chair: Christophe Croux

C1458: **On the computation of discrete mixtures of continuous distributions: Theoretical stability of algorithms**
Presenter: **Joana Leite**, Coimbra Business School | ISCAC - IPC and CMUC, Portugal
Co-authors: Jose Carlos Dias, Joao Pedro Nunes

The importance of having good algorithms to compute certain values of interest in statistics and in other areas of knowledge cannot be disputed. Algorithms implemented in many software are usually unknown to the users and can subsist for many years, even though some problems arise with their use, either CPU time or accuracy and precision. This is the case of some cumulative distribution functions (cdf), namely discrete mixtures of continuous distributions, as noncentral t and chi-square distributions. For the latter, the benchmark is the previous algorithms, based on the use of recurrence relations of special functions. The theoretical stability of these relations is not guaranteed when applied in both directions. In this setting and reporting new developments, the necessary adjustments in the algorithms are made and accuracy tests are carried out.

C1615: **A minimum-distance estimator for the calibration of simulation models**
Presenter: **Mario Martinoli**, University of Insubria, Italy
Co-authors: Raffaello Seri

The aim is to propose a new minimum-distance estimator for the calibration of simulation models on real data exploiting a nonparametric smoothing step. Consider a simulation model with parameter space Θ such that, for any $\theta \in \Theta$, the model can be used to simulate a time series $z^M(\theta)$ of length M. We want to calibrate it using an observed time series y^N of length N. We define $D\left(y^N, z^M(\theta)\right)$ as the distance between the series. When $N, M \to \infty$, under suitable assumptions of ergodicity, $D\left(y^N, z^M(\theta)\right) \to f(\theta)$, where $\theta^\star = \arg\min_{\theta\in\Theta} f(\theta)$ is the pseudo-true value of the model. For $\{\theta_i, i = 1, \ldots, P\}$, a finite subset of Θ, one can simulate $z^M(\theta_i)$ and the noisy measurements $D\left(y^N, z^M(\theta_i)\right)$ can be used to nonparametrically estimate the function f as $\widehat{f}$, where $\widehat{\theta} = \arg\min_{\theta\in\Theta} \widehat{f}(\theta)$ is an estimator of $\theta^\star$. We investigate the asymptotic properties of this estimator and we provide empirical evidence of its performance.

C1505: **GME discrete model with exogenous spatial effect variable for the job satisfaction**
Presenter: **Enrico Ciavolino**, University of Salento, Italy
Co-authors: Rossella Bernardini Papalia, Maurizio Carpita, Esteban Fernandez Vazquez

A discrete choice model with exogenous variables is proposed which takes into account spatial effect based on the Generalized Maximum Entropy (GME) estimator. The proposed model analyzes data coming from a survey devoted to measure the job satisfaction in the Italian social cooperatives, carried out on about 4000 workers over at the Italian provincial level. The objective is to describe and analyse the links between various work quality factors, taking into account the spatial effects and considering as outcome variable a discrete Likert scale. The GME is considered as estimator to deal with endogeneity and as flexible method to define the relationships between the variables of the model.

C1630: **Estimating bond risk premia via sequential learning**
Presenter: **Tomasz Dubiel-Teleszynski**, London School of Economics, United Kingdom
Co-authors: Konstantinos Kalogeropoulos, Nikolaos Karouzakis

A Bayesian learning framework for the estimation and predictability of bond risk premia under a dynamic term structure model is implemented. We develop a sequential process for investors who learn about parameters, state variables and model uncertainty, when new information arrives. We account for model uncertainty by implementing an analysis of the time-varying parameters, in particular those driving the market price of risk specification and assess the efficiency and economic importance of risk restrictions from a forecasting perspective. The methodology improves the numerical behavior of estimation and addresses the issues of the instability of parameters and the ill-behaved likelihood functions. It allows for statistical and economically plausible parameter estimation when it comes to out-of-sample bond return predictability. The estimates are capable of capturing the stylized facts of the yield curve behavior, such as the violation of the expectation hypothesis, through the predictability of excess returns, and the persistence of interest rates and provide better forecasts of bond excess returns, improving investor utility.

CG111 Room O1 CONTRIBUTIONS IN COPULAS AND APPLICATIONS Chair: Yarema Okhrin

C1624: **A vine-copula extension for the HAR-RV model**
Presenter: **Martin Magris**, Tampere University of Technology, Finland
The heterogeneous auto-regressive model of realized volatility (HAR-RV) is revised to account for non-linearity in the volatility variables, namely daily, weekly and monthly volatility components. The additive structure between these terms, although appealing for interpretation, estimation and inference, is linked to (i) a structural autoregressive hypothesis on the nature three components, and (ii) requires the classic set of hypotheses for the OLS estimation. With real high-frequency intraday data, the OLS hypotheses are not always met, while alternatives to the AR structure can be explored. We abandon the additive framework by modelling the joint distribution of the volatility components via Vine copulas. From the preliminary backtesting analyses, our model shows an increased forecasting accuracy (in terms of Diebold-Mariano test) and an improved description of volatility dynamics in terms of quantile exceedances (hit ratios) with respect to the standard HAR-RV model.

C1348: **Portfolio optimization based on forecasting models using vine copulas: An empirical assessment for the financial crisis**
Presenter: **Andreas Stephan**, Jonkoping University, International Business School, Sweden
Co-authors: Maziar Sahamkhadam
Vine copulas are employed for modeling the symmetric and asymmetric dependency structure and forecasting of financial returns. AR-GARCH models are used for filtering out the residuals. Asset allocation is performed during the 2007-2010 financial crisis and different portfolio strategies were tested including maximum reward-to-risk ratio (SR), minimum variance (MV) and minimum conditional Value-at-Risk (CVaR). Regular, drawable and canonical vine copulas were specified including Clayton, Frank, Joe and mixed copula. Both in-sample and out-of-sample analyses of portfolio performances were conducted. The out-of-sample portfolio back-testing showed that vine copulas reduce portfolio risk more than simple copulas. The results of the VaR back-testing and risk-adjusted performance showed improvement in forecasting of the downside risk for all portfolio strategies obtained from using mixed copula families, implying time-varying tail dependence of stock market returns. Copula families which capture symmetric tail dependence (Frank) and upper tail dependence (Joe) lead to higher terminal values of portfolios over the financial crisis.

C1335: **Tail dependence in the Australian electricity market: Evidences from the vine copula and the dependence-switching copula**
Presenter: **Shixuan Wang**, University of Reading, United Kingdom
Co-authors: Nicholas Apergis, Giray Gozgor, Chi Keung Marco Lau
The tail dependence of the Australian Electricity Markets (AEM) is investigated by using two copula methods: the vine copula (perspective: multivariate and static) and the dependence-switching copula (perspective: bivariate and dynamic). The findings from the vine copula approach indicate that the best multivariate dependence structure of the AEM exactly matches the geographical connectedness of the States, which implies that the infrastructure connectedness of transmission wires determines the dependence structure in the AEM. In addition, we observe that most pairs have a symmetric tail dependence indicated by the students t-copula, except in the case of the Gumbel copula and in relevance to the pair of Victoria and Tasmania, suggesting only the presence of right tail dependence. Based on the information criteria, we find the evidence of the dependence-switching copula for four pairs of states. For those pairs, the dependence-switching copula revealed the time-varying dependence structure, asymmetric tail dependence, and longer sojourn time in the certain regimes. Our findings provide valuable insights for market participants and policymakers.

C1202: **Nonparametric dynamic copula modelling to analyze dependence structures between domestic indexes**
Presenter: **Jone Ascorbebeitia**, University of the Basque Country, Spain
Co-authors: Eva Ferreira, Susan Orbe
The analysis between portfolio comovements is of great interest in economics and finance in order to have as much as possible power over the risk. The time varying asset dynamics make more difficult to control those comovements and require much more accuracy and more sophisticated estimation models able to capture the dynamics. Unfortunately, financial variables are non-gaussian distributed and present more complicated structures. So, to measure the dependence between them it is necessary to consider dependence measures beyond Pearson's linear correlation. To overcome this fact, we suggest the use of time-varying copulas to analyze the relationship between European domestic index dynamics. As it is already known, the use of copulas allows us to model the dependence better than elliptic distributions do. In this context, we propose a nonparametric time-varying dependence estimator based on Kendall's tau to analyze the dependence between index dynamics and we derive its asymptotic properties. A simulation study investigates the performance of the estimator and compares with the performance of other dependence estimation methods existing in the literature. Finally, we provide a statistic to test for Kendall's tau significance.

CG549 Room A2 CONTRIBUTIONS IN FINANCIAL MODELLING AND FORECASTING Chair: Jiri Witzany

C1319: **Forecasting conditional covariance matrices in high-dimensional data using generalised dynamic factor model**
Presenter: **Carlos Trucios**, Sao Paulo School of Economics - FGV, Brazil
Co-authors: Joao Henrique Goncalves Mazzeu
The generalised dynamic factor model with infinite-dimensional factor space is used to develop a new procedure to estimate and forecast the conditional covariance matrix in high-dimensional data. The performance of the procedure is evaluated via Monte Carlo experiments and the results show good finite sample properties. The new procedure is used to construct minimum variance portfolios in moderate and high-dimensional real datasets. The results reveal a better out-of-sample portfolio performance when compared with alternative procedures.

C1477: **On the properties of Λ-quantiles**
Presenter: **Fabio Bellini**, University of Milano-Bicocca, Italy
Co-authors: Ilaria Peri
The aim is to study the properties of a family of risk measures recently introduced in the financial literature under the name of Λ-Value at Risk, that provides an interesting generalization of the usual quantiles. We provide an axiomatic foundation for Λ-quantiles, similar in spirit to analogous axiomatizations of the usual quantiles. Under mild technical conditions, we characterize the class of scoring functions that are strictly consistent with the Λ-quantile interval, generalizing the corresponding class of GPL functions that elicits the usual quantiles. Finally, we discuss financial applications of Λ-quantiles forecasting by means of regression techniques.

C1511: **Economic policy uncertainty as an indicator of abrupt movements in the US stock market**
Presenter: **Paraskevi Tzika**, University of Macedonia, Greece
Co-authors: Theologos Pantelidis
A two regime switching model is developed as an attempt to relate expected US stock market returns to deviations from fundamentals and to Economic Policy Uncertainty (EPU). The analysis is based on monthly data that cover the period from January 1900 to December 2017 and the EPU index is used as an explanatory variable. The findings suggest that the US stock market spends most of the time in a low-volatility regime, periodically switching to a high-volatility regime during periods of financial instability. In an attempt to examine the forecasting ability of the

model, out-of-sample probabilities of a crash and a boom are estimated recursively. The results provide evidence that our model is able to depict periods of abrupt movements in the US stock market. Finally, the estimated model and the associated probabilities of a crash and a boom are used to develop and evaluate trading strategies, in order to analyse the financial usefulness of the model.

C1713: **Sequential Gibbs particle filter algorithm with an application to stochastic volatility and jumps estimation**
Presenter: **Jiri Witzany**, University of Economics in Prague, Czech Republic
Co-authors: Milan Ficura

The aim is to propose and test a novel particle filter method called sequential Gibbs particle filter which allows the estimation of complex latent state variable models with unknown parameters. The framework is applied to a stochastic volatility model with independent jumps in returns and volatility. The implementation is based on a novel design of adapted proposal densities making convergence of the model relatively efficient as verified on a testing dataset. The empirical study applies the algorithm to estimate stochastic volatility with jumps in returns and volatility model based on the Prague stock exchange returns. The results indicate surprisingly weak jump in returns components and a relatively strong jump in volatility components with jumps in volatility appearing at the beginning of crisis periods.

CG125 Room B2 CONTRIBUTIONS IN PORTFOLIO OPTIMIZATION II Chair: Zinoviy Landsman

C1338: **Portfolio optimization based on multivariate GARCH copula models**
Presenter: **Maziar Sahamkhadam**, Linnaeus University, Sweden

Multivariate GARCH models in combination with vine copula are tested to forecast one-step ahead stock index returns and compute optimal portfolio weights. The common marginal distributions in multivariate GARCH models are multivariate normal and Student t, which are not able to capture the tail dependency structure required in modeling fat-tailed financial returns. To model tail behavior, the Clayton canonical vine copula is used, which captures the lower asymmetric tail dependence with uniform marginal obtained from dynamic conditional correlation and generalized orthogonal GARCH models. Moreover, we test extreme value theory in modeling the downside risk in multivariate GARCH-Copula forecasting settings. Three portfolio strategies including minimum Conditional Value-at-Risk (CVaR), maximum reward-to-risk (SR) and Markowitzs mean-variance (MV) are obtained and back-tested over an out-of sample period, which contains financial crisis. The results show out-performance based on DCCGARCH and GOGARCH models comparing to univariate GARCH. In particular, the results of VaR back-testing and risk-adjusted performance show improvement in forecasting the downside risk for CVaR portfolio strategy obtained by GOGARCH canonical Clayton vine model with semi-parametric marginal distribution (based on EVT). This model also leads to a better economic performance for SR portfolio strategy over a long-horizon investment period.

C1357: **Bond portfolio optimization using regime switching dynamic Nelson Siegel models**
Presenter: **Takeshi Kobayashi**, NUCB Business School, Japan
Co-authors: Naoki Makimoto

Markowitz's approach of portfolio selection is applied to US government bond portfolios. We use dynamic Nelson Siegel yield curve modes to estimate expected value of portfolios, and variances, and covariances of portfolios. We also extend the dynamic Nelson Siegel model to a Markov switching latent variable model that allows for discrete changes in the stochastic process followed by the interest rates. We consider switching the following four parameters: (i) a loading parameter, (ii) time-varying volatility, (iii) an unconditional mean and (iv) a matrix of the autoregressive process coefficient. We derive a model of a myopic single-period investment strategy in a transaction cost conscious mean-variance framework with dynamic factor models under regime switches. We compare the investment performance of these models with that of the standard dynamic Nelson and Siegel model and find that regime switching dynamic factor models are useful in constructing bond portfolios which realize higher risk adjusted return during yield curve regime shifts.

C1517: **Portfolio diversification in the spectral domain**
Presenter: **Martin Hronec**, Charles University in Prague, Faculty of Social Sciences, Czech Republic

When investors risk preferences differ across time horizons, the diversified portfolio needs to be immune not only against shocks aggregated across time horizons but also at specific time horizons. We apply spectral analysis into diversification-based portfolio selection models. By replacing the covariance matrix estimates with the cross-spectrum based ones, we restrict the optimization problems in these models to the desired frequency band, allowing an investor to target specific time horizons. Further, we generalize for an investor facing risk constraints at different frequencies by including the shape of the cross-spectrum into the optimization problem. We provide several numerical and empirical examples that show investors may benefit by considering not only diversification across aggregate risk sources but also across different frequencies.

C1609: **Dynamically optimal multi-period mean-variance portfolio selection with transaction costs and no-shorting constraint**
Presenter: **Zi Ye**, Nanyang Technological University, Singapore
Co-authors: Chi Seng Pun

A mean-variance portfolio selection problem is considered in multi-period with proportional transaction costs under no-shorting constraint. By adopting dynamic programming and duality theory, we derive the analytical solution of the optimal investment policy, which is a piecewise function, and find the boundaries of buying region, no-transaction region, selling region and selling-all region. In addition, with the developed concept of time consistency in efficiency (TCIE), our analysis shows this policy is always TCIE, which means the investor has no incentive to change idea in the whole investment time horizon, besides we find no shorting constraint is a sufficient condition of TCIE for a no transaction model. Furthermore, we use the efficient frontier to illustrate the policy, it shows the allocation of the initial wealth and the proportional transaction cost rates would affect the investor's behavior. The results have implications in the financial market that contains friction rates which should be taken into account in investment problems.

CG018 Room C2 CONTRIBUTIONS IN STRUCTURAL BREAKS Chair: Arnaud Dufays

C1342: **Multiple structural breaks in cointegrating regressions: A model selection approach**
Presenter: **Alexander Schmidt**, University of Hohenheim, Germany
Co-authors: Karsten Schweikert

A new comprehensive treatment of structural change in cointegrating regressions is proposed. First, we consider a setting with fixed breakpoint candidates and show that a modified adaptive lasso estimator can consistently estimate structural breaks in the intercept and slope coefficient of a cointegrating regression. Second, we extend our approach to a diverging amount of breakpoint candidates and provide simulation evidence that timing and magnitude of structural breaks are estimated consistently. Third, we use the adaptive lasso estimation to design new tests for cointegration in the presence of multiple structural breaks, derive the asymptotic distribution of our test statistics and show that the proposed tests have power against the null of no cointegration. Finally, we use our new methodology to study the effects of structural breaks on the long-run PPP relationship.

C1389: **Selective linear segmentation for detecting relevant parameter changes**
Presenter: **Arnaud Dufays**, Namur University, Belgium
Co-authors: Elysee Houndetoungan
Change-point processes are one flexible approach to model long time series. We propose a method to uncover which model parameter changes when a change-point is detected. When the number of break points is small, an exhaustive search based on a consistent criterion is used to select the best set of parameters that change over time. In the other situation, we use a penalized likelihood approach to reduce the number of models to consider, and we prove that the penalty function will lead to a consistent selection of the true model. Estimation in such a case is carried out via the deterministic annealing expectation-minimisation algorithm. Interestingly, the method accounts for model selection uncertainty and provides a probability of selecting a specific set of covariates. Monte Carlo simulations highlight that the method works well in small and large samples for many time series models. An application on hedge funds returns shows how we can exploit the framework.

C1464: **Modelling long memory and structural breaks in count data**
Presenter: **Mawuli Segnon**, University of Munster, Germany
An integer valued negative binomial Markov switching Multifractal (NegBin-MSM) model is developed by adapting the MSM process for count data setting. We provide the statistical properties of the NegBin-MSM process and demonstrate its capacity to reproduce overdispersion, long memory and structural breaks that characterize count data. We show via Monte Carlo simulation that the maximum likelihood estimator is consistent and asymptotically efficient. An empirical application with financial transaction data illustrates the practical importance of the model.

C1558: **Forecasting commodity prices in a data-rich, unstable environment**
Presenter: **Anastasia Allayioti**, University of Warwick, United Kingdom
Co-authors: Fabrizio Venditti
Recent research has shown that commodity prices exhibit substantial co-movement, which can be captured by few common factors, broadly related to global demand for commodity shocks, which are pervasive across all commodity prices, and idiosyncratic (commodity-specific) supply shocks. A separate literature has stressed how the composition of underlying structural shocks that drive commodity prices has changed over time, potentially resulting in unstable unconditional correlations. These two findings suggest that (i) forecast accuracy for the price of a given commodity could benefit from the information contained in other commodity prices and that (ii) dealing with potential structural breaks could also improve forecast accuracy. We investigate the merits of constructing forecasts for key commodity prices from models that use large information sets and deal with structural breaks. We consider large TVP-VARs, TVP dynamic factor models (TVP-DFMs) and TVP hierarchical dynamic factor models (TVP-HDFMs), which impose the presence of specific blocks on the factor model structure of commodity prices. Given that standard estimation methods for small-dimensional models fail in a data-rich environment, we adopt non-parametric kernel-based methods and forgetting factor techniques. A distinct contribution of these methods is that, unlike most of previous ones, we evaluate both point and density forecasts.

CG377 Room D2 CONTRIBUTIONS IN FINANCIAL ECONOMETRICS II Chair: Eduardo Rossi

C1722: **Brexit: Tracking and disentangling the sentiment towards leaving the EU**
Presenter: **Gabriel Martos**, Fundacion Universidad Torcuato Di Tella, Argentina
Co-authors: Miguel de Carvalho
On 23 June 2016 the UK held a referendum so to decide whether to stay or leave the European Union. The uncertainty surrounding the outcome of this referendum had major consequences in terms of public policy, investment decisions, and currency markets. We discuss some subtleties entailed in smoothing and disentangling poll data at the light of the problem of tracking the dynamics of the intention to Brexit, and propose a multivariate singular spectrum analysis method that produces trendlines on the unit simplex. The trendline yield via multivariate singular spectrum analysis is shown to bear a resemblance with that of local polynomial smoothing, and singular spectrum analysis presents the nice feature of disentangling directly the dynamics into components that can be interpreted as changes in public opinion or sampling error. Merits and disadvantages of some different approaches to obtain smooth trendlines on the unit simplex are contrasted, both in terms of local polynomial smoothing and of multivariate singular spectrum analysis.

C1503: **Multivariate automated circulant SSA**
Presenter: **Eva Senra**, Universidad de Alcala, Spain
Co-authors: Juan Bogalo, Pilar Poncela
Circulant Singular Spectrum Analysis (CSSA) is an automated version of SSA that allows to extract the unobserved components associated to any frequency in a time series in an automated way. We generalize the technique to a multivariate setup and automatize it in the same way by the use of block circulant matrices applied to a new multivariate trajectory matrix. With the multivariate extension (M-CSSA) we can decompose the estimated signal at any frequency into the sum of M orthogonal components that will allow to characterize the main sources of the fluctuations and obtain more robust signals for each individual time series. We also specify a time series factor model for an estimated vector signal that allows to obtain the common factors in the frequency domain from the information obtained with M-CSSA. Finally, we illustrate the application of the technique to a group of series.

C1672: **On distribution of order books**
Presenter: **Martin Smid**, Institute of Information Theory and Automation, Czech Republic
Order-books of limit order markets follow a complicated dynamics with an infinite dimensional state space. Thus, even under simplifying assumptions, their distribution is not analytically tractable. However, the distribution can be described in two steps: a formula for a conditional distribution of order books given the history of trade and quote data exist while the distribution of the trade and quote process may be described recursively. Using these expressions and ergodicity of the process, which is guaranteed under reasonable conditions, parameters of the model may be estimated by maximum likelihood, which is demonstrated on data from several US stock markets.

C1638: **Identification of overdetermined structural VAR models**
Presenter: **Francesco Cordoni**, Scuola Normale Superiore, Italy
Co-authors: Fulvio Corsi
When the number of observed macroeconomic variables is larger than the number of structural shocks driving the economy, the associated structural VAR system is said to be overdetermined. We propose an identification method for overdetermined structural VAR models by first pretesting for the number of shocks and then combining a collapsing procedure with a PML approach. We show the consistency of the proposed combined procedure and examine its finite sample properties with Monte Carlo simulations. The empirical application of the proposed scheme on U.S. data allows to identify the low dimensional system of structural shocks driving the U.S. economy.

CG095 Room E2 CONTRIBUTIONS IN ASSET PRICING Chair: Francesco Violante

C1472: **Tales of sentiment driven tails**
Presenter: **Jozef Barunik**, UTIA AV CR vvi, Czech Republic
Co-authors: Cathy Yi-Hsuan Chen, Wolfgang Karl Haerdle

The link between investor sentiment and asset valuation is the subject of considerable debate in the profession. The aim is to abandon the classical asset pricing that relies on expected utility, and introduce a dynamic quantile model for asset pricing, in which the agent maximizes stream of the future quantile utilities instead. Using the model, we empirically investigate if investor sentiment distilled from textual mining analysis can price tails of the return distributions. On the panel of 100 stocks, we document influence of aggregate investors sentiment on future conditional quantiles of the return distributions. Aggregate sentiment explains cross-section of tails even after controlling for popular factors used in the literature, as well as firm-specific sentiment and volatility.

C1502: **Tail risks, asset prices, and investment horizons**
Presenter: **Matej Nevrla**, Charles University, Czech Republic
Co-authors: Jozef Barunik

The aim is to examine how extreme market risks are priced in the cross-section of asset returns at various horizons. Based on the frequency decomposition of covariance between indicator functions, we define the quantile cross-spectral beta of an asset capturing tail-specific as well as horizon-, or frequency-specific risks. Further, we work with two notions of frequency-specific extreme market risks. First, we define tail market risk that captures dependence between extremely low market as well as asset returns. Second, extreme market volatility risk is characterized by dependence between extremely high increments of market volatility and extremely low asset return. Empirical findings based on the datasets with long enough history, 30 Fama-French Industry portfolios, and 25 Fama-French portfolios sorted on size and book-to-market support our intuition. Results suggest that both frequency-specific tail market risk and extreme volatility risks are significantly priced and our five-factor model provides an improvement over specifications considered by previous literature.

C1506: **Dynamic quantile models, rational inattention, and asset prices**
Presenter: **Lukas Vacha**, Univerzita Karlova, Fakulta socialnich ved, Czech Republic
Co-authors: Jozef Barunik

The aim is to study asset pricing under uncertainty with agents having quantile preferences, and limited information processing capacity. Abandoning the classical asset pricing that relies on expected utility we introduce a dynamic quantile model for asset pricing, in which the agent maximizes stream of future quantile utilities instead. In addition, an agent cannot acquire all information about future states of her portfolio freely. In contrast to the rational expectation models, the agent has a limited amount of attention since the information she obtains is costly. In our model, the agent maximizes stream of her future quantile utilities according to her quantile utility preferences subject to information costs constraints. Our results show that there is a significant benefit when a standard expected utility is expanded into quantile preference utilities.

C1495: **Dynamic quantile model for bond pricing**
Presenter: **Frantisek Cech**, UTIA AV CR vvi, Czech Republic
Co-authors: Jozef Barunik

A dynamic quantile model is introduced for bond pricing with an agent who values securities by maximizing the quantile level of her utility function. The transition from traditional to quantile preferences allows us to study the pricing of the term structure of interest rates by economic agents differing in their levels of risk aversion. Moreover, the framework is robust to fat tails commonly observed in the empirical data. In the application, we focus on the quantile pricing of the two, five, ten and thirty years US and German government bonds. For the analysis, we use flexible quantile regression framework which is applied over highly liquid bond futures contract from the Chicago Board of Trade and EUREX exchanges.

CG067 Room F2 CONTRIBUTIONS IN FINANCIAL MARKETS Chair: Teruo Nakatsuma

C0265: **Market efficiency: Saudi stock exchange**
Presenter: **Hans-Philipp Otto**, EBS Universitat fur Wirtschaft und Recht, Germany

Saudi Arabia is facing an array of major social, economic, and structural changes striving for transformation from an oil dependent country into a high tech and knowledge based economy with a thriving private sector driven by the threat of budget deficits while at the same time shaping the future by undertaking major lighthouse development projects. The Saudi Stock Exchange amplifies its efforts in the enhancement of the structural and regulatory improvement making the Saudi Stock Market attractive to foreign investors and prepared for the Aramco IPO. The aim is, firstly, to provide a thorough overview of the literature in the overlap of the weak form market efficiency hypothesis testing and the Saudi Stock Market and, secondly, to deepen this overlap by adding the dimension of the subindices to the assessment while, thirdly, applying a well-balanced and exhaustive set of parametric and nonparametric methods to the complete timeframe with 15 subindices from 2007 to 2017 to answer the research questions whether the TASI and its subindices are weak form efficient and whether the degree of efficiency does evolve over time by examining segmented datasets on possible evolutions over time. It can be concluded that the TASI as well as all subindices are not weak form market efficient but do not exhibit long range dependency. Interestingly, the efficiency improves during the time of the financial crisis and alienates during the period of recovery from the financial crisis.

C1395: **Neighbors matter: Geographical distance and trade timing in the stock market**
Presenter: **Margarita Baltakiene**, Tampere University of Technology, Finland
Co-authors: Kestutis Baltakys, Hannu Karkkainen, Juho Kanniainen

In financial markets, investors are socially connected and have access to overlapping sources of information. Many investor groups have similar trading strategies because of the common information sources or the ability to share the knowledge with each other. As a result, neighboring investors may exchange information about their transactions in stock markets, leading to similar trading behavior. We find that pairwise trade timing similarities between investor pairs are negatively associated to the geographical distance between corresponding investor pairs. This suggests that local information transfer channels between neighboring individual investors are used in decision making. We also observe that differences in age and language moderate this association. The analysis is conducted using investor account level data from different regions of Finnish households.

C1454: **Liquidity in the FX market: Empirical evidence from an aggregator**
Presenter: **Milla Siikanen**, Tampere University of Technology, Finland
Co-authors: Juho Kanniainen, Ulrich Noegel

In foreign exchange (FX) trading, an aggregator is used to connecting traders with liquidity providers (LPs). In an aggregator, a trader receives a continuous stream of bid and ask quotes from a predefined set of LPs, and the difference between the best bid and ask prices over a set of liquidity streams is called an inside spread. We empirically study liquidity in an FX aggregator. We show that, on average, traders obtain a relatively tight spread already with four or five streams; the use of more streams yields a marginal benefit only. For given numbers of liquidity streams, we

determine the optimal combinations of streams minimizing the spread. The optimal combinations are obtained using a genetic algorithm. Our findings indicate that most of the traders could—at least in theory—reduce the average spread by more than half with the optimal combination of streams, and a trader could save up to $0.18 basis points per euro traded. However, traders may not be able to fully exploit the improvements in spreads because, in practice, the traders are not completely free to choose just any liquidity streams in the aggregator. On the other hand, if the traders changed their selected liquidity streams, the LPs would be likely to change their quoting behavior. In addition, we find that a model proposed for a liquidity aggregator in an earlier research fits our empirical data accurately, even under the quite simplistic assumptions of homogeneous LPs.

C1539: **Cross-asset contagion in the financial crisis: A Bayesian time-varying parameter approach**
Presenter: **Manuela Pedio**, Bocconi University, Italy
Co-authors: Erwin Hansen, Massimo Guidolin

Contagion mechanisms in the US financial markets are studied. The recent US subprime crisis provides us with one exogenous shock in a specific market (mortgage-backed securities) to measure contagion. We model the dynamic linkages among markets and allow for changes in this relationship to capture contagion. We look at how and to what extent a negative shock that initially occurred in the asset-backed security (ABS) low-quality market propagated to ABS higher grade, Treasury repos, Treasury note, corporate bond, and stock markets. We rely on dynamic time series models estimated with Bayesian methods. We estimate several specifications ranging from single-state vector autoregressive (VAR) models with constant parameters to fully flexible VAR models where the parameters may vary at each observation. We provide evidence of structural changes in the cross-asset relationships and therefore of contagion. Moreover, by observing the impulse response functions of the models, we conclude that contagion mainly occurred through the flight-to-liquidity, risk premium, and correlated information channels.

CG463 Room G2 CONTRIBUTIONS IN MACHINE LEARNING FOR TIME SERIES FORECASTING Chair: Harish Bhat

C1312: **L_2 boosting for high dimensional locally stationary time series**
Presenter: **Kashif Yousuf**, Columbia University, United States
Co-authors: Serena Ng

High dimensional time series analysis has attracted an increasing amount of attention in the econometrics literature in recent years. However, one of the main limiting assumptions made by most works on the topic is assuming stationarity and time invariant effects of the predictors. We study the theoretical properties of L_2 boosting for high dimensional time varying coefficient models, where the coefficients are modeled as smooth functions evolving over time and the predictors are locally stationary. We establish consistency of our procedures when using either componentwise local linear or local constant estimators as base learners. Dependence is quantified by functional dependence measures and the asymptotic properties of our methods depend on the moment conditions, the sparsity level, and the strength of dependence in the underlying processes, among other factors. Practical issues such as choosing the bandwidth for the base learners, and the number of boosting iterations are also addressed. Lastly finite sample performance of our procedures is shown through extensive simulation studies, and we include an application to macroeconomic forecasting.

C1371: **Calibrating rough volatility models: A convolutional neural network approach**
Presenter: **Henry Stone**, Imperial College London, United Kingdom

Convolutional neural networks are used to solve the classification problem of finding the Hölder exponent of two Gaussian processes: the well-known fractional Brownian motion and the rBergomi model, a recently proposed stock price model used in mathematical finance. We contextualise the latter as a calibration problem, thereby providing a very practical and useful application.

C1482: **Decomposition of high frequency Forex signals for copula based pairs trading strategy with support vector regression**
Presenter: **Carlin Chu**, The Open University of Hong Kong, Hong Kong
Co-authors: Po Kin Chan

The microstructure noise inherited in high frequency trading signal is a nuisance for effective modeling of short-term trends. The typical approach to tackle this issue is to remove the noisy data part by filtering or smoothing methods. Oscillating zig-zag patterns are considered as noises and smoothed out. However, this approach does not consider the possibility of utilizing the information hidden in the noise data. A proper usage of decomposition method to exploit the noise information is investigated. Empirical Mode Decomposition (EMD) is proposed to decompose the non-stationary trading signal into intrinsic mode functions (IMFs) and residual series for building a prediction model. The aim is to extend the use of copula-based Mispriced Index (MI) and Support Vector Regression (SVR) for pairs trading by incorporating IMFs on the model building stage. Properties of IMFs, selection of bivariate copulas and suitability of different SVR kernels are examined. The empirical results indicated that the use of IMFs can significantly improve the model accuracy in almost all settings. The subtle relationships among the EMD, copula functions and hyperparamters of SVR are discussed.

C1673: **Comparing linear and non-linear dynamic factor models for large macroeconomic datasets**
Presenter: **Alessandro Giovannelli**, University of Rome Tor Vergata, Italy

A non-linear extension for macroeconomic forecasting is proposed by using a large dataset based on a dynamic factor model (DFM). The main idea is to allow the factors to have a non-linear relationship to the input variables using the methods of (i) kernel and (ii) neural networks principal component analysis. We compare the empirical performances of these methods with (iii) the standard principal-component model introduced by Stock and Watson in 2002, conducting a pseudo forecasting exercise based on a Euro Area macroeconomic dataset composed by 834 monthly variables spanning the period January 1996 - September 2017. Using a rolling window for estimation and prediction, the results obtained from the empirical study suggest that (i) and (ii) have the same forecasting performances of (iii) for both Industrial Production and Inflation, but (i) significantly outperforms (iii) for the Unemployment Rate. Moreover, there is no difference with respect to previous results if we consider the pre-crisis period. However, during the crisis and subsequent recovery, we observe a slight improvement of (ii) with respect to (i) for Industrial Production and Inflation while (i) is the best model for Unemployment Rate.

CG016 Room I2 CONTRIBUTIONS IN FINANCIAL TIME SERIES Chair: Roxana Halbleib

C1375: **A residual bootstrap for conditional value-at-risk**
Presenter: **Alexander Heinemann**, Maastricht University, Netherlands
Co-authors: Eric Beutner, Stephan Smeekes

A fixed-design residual bootstrap method is proposed for the two-step estimator associated with the conditional Value-at-risk (VaR). The bootstrap's consistency is proven under mild assumptions for a general class of volatility models and bootstrap intervals are constructed for the conditional VaR to quantify the uncertainty induced by estimation. A large-scale simulation study is conducted revealing that the equal-tailed percentile interval based on the fixed-design residual bootstrap tends to fall short of its nominal value. In contrast, the reversed-tails interval based on the fixed-design residual bootstrap yields accurate coverage. In the simulation study we also consider the recursive-design bootstrap. It turns out that the recursive-design and the fixed-design bootstrap perform equally well in terms on average coverage. Yet in smaller samples the fixed-design scheme leads on average to shorter intervals. An empirical application illustrates the interval estimation using the fixed-design residual bootstrap.

C1712: **Forecasting stochastic volatility with realized volatility estimators and particle filters**
Presenter: **Milan Ficura**, University of Economics in Prague, Czech Republic
Co-authors: Jiri Witzany
SVJD-RV-Z class of models is developed, utilizing the realized variance for better estimation of the stochastic variances, and the non-parametric *Z*-estimator for more accurate estimation of price jumps. Several adapted particle filters, specifically designed for latent-state filtering in SVJD models, are derived, and a Sequential Gibbs Particle Filter (SGPF) algorithm is developed for the sequential learning of their parameters. In the empirical study, four SVJD models (with intraday data, self-exciting jumps in prices and volatility, as well as multiple volatility components) are applied for the task of realized volatility forecasting on the time series of 7 foreign exchange rates and 10 ETF/ETN securities in the daily, weekly and monthly forecast horizon. The performance of the SVJD models is compared with 3 GARCH models (GARCH, EGARCH and GJRGARCH), 15 HAR model specifications (HAR, AHAR, SHAR, HARJ and HARQ), and 15 Echo State Neural Network (ESN) based volatility models previously developed. The SVJD-RV-Z models with jumps in volatility and prices are shown to exhibit the highest out-sample predictive power, comparable to the best HAR and ESN model specifications.

C1553: **QML estimation of a stochastic volatility with leverage and size effect model**
Presenter: **Paolo Chirico**, University of Eastern Piedmont, Italy
If the use of the Kalman filter (KF) for simple, univariate and multivariate, stochastic volatility (SV) models is well known, the use of the filter for asymmetric SV model is not simple. As known, the auxiliary model used with the KF, the linear structural model of the log-squared returns, does not allow to consider the correlation between the current return innovation and the innovation on the next return volatility (leverage effect). On the other hand, the disturbances of the auxiliary model are correlated because of the presence of the size effect, the effect on the return volatility due to the magnitude of the previous return. Taking into account that correlation, a SV with leverage and size effect (SV-SLE) model is performed using the KF. The applications of the model to some financial time series show interesting results: in some cases (series), the volatility innovation is not significant so the SV-LSE model is very similar to the EGARCH model; in other cases, the SV-LSE model fits the series better than the EGARCH model.

C1554: **A general class of score-driven smoothers**
Presenter: **Giuseppe Buccheri**, Scuola Normale Superiore, Italy
Co-authors: Giacomo Bormetti, Fulvio Corsi, Fabrizio Lillo
It is first shown that, in the steady state, Kalman filter and smoother recursions can be re-parameterized in terms of the score of the conditional density and the Fisher matrix. Since in the new representation the predictive filter has the form of score-driven models, we introduce, by analogy, a score-driven update filter (SDU) and smoother (SDS). In this new framework, we recover smoothed estimates of observation-driven models, as well as assess filtering uncertainty and construct confidence bands. We test both empirically and through simulations the advantages of SDU and SDS over standard score-driven filters and exact simulation-based methods.

CG065 Room M2 CONTRIBUTIONS IN MACROECONOMIC POLICIES AND MACROECONOMETRICS Chair: Giorgio Primiceri

C1209: **International trade, exchange rate regimes, and financial crises**
Presenter: **Maria Santana Gallego**, University of the Balearic Islands, Spain
The main objective is to study the impact of different exchange rate regimes on international trade and to analyze their performance during crises. To this end, a gravity equation for bilateral trade is estimated for a sample of 191 countries over the period 1970-2016 by adding a set of regressors built from a de facto classification of exchange rate arrangements and the dates of recognized financial crises. Moreover, we differentiate between anchor currencies and direct and indirect exchange rate arrangements. The gravity model is consistently estimated by including three different types of high-dimensional fixed effects and using PPML estimates. The main empirical findings are: i) other intermediate exchange rate regimes, between completely fixed and completely flexible, promote flows of goods between countries; ii) results depend on the anchor currency and indirect arrangements do not have any significant impact on international trade; iii) systemic banking crises negatively affect trade flows between countries; and iv) the impact of the exchange rate regimes on trade during crisis depends on the anchor currency and whether the crisis takes place in the exporting or the importing country.

C1560: **The effects of conventional and unconventional monetary policy on exchange rates**
Presenter: **Barbara Rossi**, Universitat Pompeu Fabra and ICREA, Spain
Co-authors: Atsushi Inoue
What are the effects of monetary policy on exchange rates? And have unconventional monetary policies changed the way monetary policy is transmitted to international financial markets? According to conventional wisdom, expansionary monetary policy shocks in a country lead to that country's currency depreciation. We revisit the conventional wisdom during both conventional and unconventional monetary policy periods in the US by using a novel identification procedure that defines monetary policy shocks as changes in the whole yield curve due to unanticipated monetary policy moves and allows monetary policy shocks to differ depending on how they affect agents' expectations about the future path of interest rates as well as their perceived effects on the riskiness/uncertainty in the economy. Our empirical results show that: (i) a monetary policy easing leads to a depreciation of the country's spot nominal exchange rate in both conventional and unconventional periods; (ii) however, there is substantial heterogeneity in monetary policy shocks over time and their effects depend on the way they affect agents' expectations; (iii) we find favorable evidence to the overshooting hypothesis.

C0610: **Impulse response functions in DSGE models as a perturbation to the deterministic solution**
Presenter: **Viktors Ajevskis**, Bank of Latvia, Latvia
In the conventional perturbation approach to solve DSGE models, the dynamics of the deviation of solutions from the steady state after a shock hitting an economy represents an impulse response function (IRF). A method to construct the IRF as a deviation from a deterministic solution is proposed. In this framework, the deterministic solution is treated as a trend. The approach detects asymmetric reactions of an economy to shocks in different initial conditions. For example, in an economic downturn a negative shock might affect the economy more severe than in normal economic conditions. The method allows for constructing the IRF for highly non-linear DSGE models.

C1667: **Symmetry and separability in two-country cointegrated vector autoregressive processes**
Presenter: **Hans-Martin Krolzig**, University of Kent, United Kingdom
Co-authors: Reinhold Heinlein
Introducing a previous approach to time series econometrics, it is shown that the dynamics of symmetric linear possibly cointegrated two-country VAR models can be separated into two autonomous subsystems: the country averages and country differences, where the latter includes the exchange rate. Under symmetry, the cointegration rank of the two-country model is given by the sum of the two subsystems. The two economies are cotrending if the country-differences subsystem is stable. It is shown that separability carries over even under asymmetries in the form of difference in the size of the countries' economies, where asymptotically a small-open-economy separation into closed économique dominant and country-difference subsystems emerges. In the case of time-varying country weights, the two-country systems and the country-average-difference

representation are no longer isomorphic, but it is the latter that should be considered structural. The possibility of recursive structural VECM representations under symmetry is also evaluated. The derived conditions for symmetry and separability are easily testable and applied to a nine-dimensional quarterly cointegrated VAR model for the US and Euro Area in the post-Bretton-Woods era. We find evidence for symmetry in the long-run and with regard to the exchange rate dynamics.

CG539 Room N2 CONTRIBUTIONS IN LONG MEMORY Chair: Liudas Giraitis

C0903: **Macroeconomic forecasting with fractional factor models**
Presenter: **Tobias Hartl**, University of Regensburg, Germany
Instead of pre-differencing time series for the estimation of dynamic factor models, the use of models that incorporate common fractionally integrated unobserved components in levels is suggested. Three frameworks that allow for long-range dependence both in the common components and idiosyncratic errors are derived. In these models the factors either establish fractional cointegration relations, or they can be eliminated by taking non-integer differences. A two-stage estimator, that combines principal components and the Kalman filter, is proposed. The forecast performance is studied for a large macroeconomic dataset for the US, where we find that benefits from the fractional factor models can be substantial, as they outperform univariate autoregressions, principal components in integer differences, a combination of both and factor-augmented error-correction models.

C1409: **Long memory conditional heteroscedasticity in count data**
Presenter: **Manuel Stapper**, WWU Muenster, Germany
A new class of long memory integer-valued processes is introduced, which are adaptations of the well-known FIGARCH and HYGARCH processes to a count data setting. Statistical properties of the models are provided and it is shown via simulation that reasonable parameter estimates are easily obtained via conditional maximum likelihood estimation. An asymptotic test is derived and used to test for restrictions. To illustrate the practical importance of the models, an empirical application with financial transaction data is performed. For this purpose, high frequency data is collected and the number of price changes in 60-second intervals used as time series.

C1690: **On the long memory feature through temporal aggregation**
Presenter: **Aleksandr Pereverzin**, University of East Anglia, United Kingdom
One of the key features of empirical work with economic or financial time series is that the time series under consideration is often aggregated in time. The effect of temporal aggregation on time series which are characterized by a long memory dynamics is studied. The aim is to investigate if a long memory property of time series is invariant to the sampling frequency or aggregation scheme. We combine the time and frequency domain analysis and generalize the up to date theoretical knowledge about the temporal aggregation in discrete-time long memory ARFIMA processes. Monte Carlo simulation is conducted to validate the theoretical implications about the effects of temporal aggregation on long memory processes and estimating the memory parameter of aggregated series in the time and frequency domains. We concentrate our empirical analysis on the high frequency foreign exchange data. Several various tests are used to investigate the long memory dynamics of the foreign exchange rates absolute and squared returns series on the various levels of temporal aggregation. Our results have implications for financial risk management dealing with volatility modeling and forecasting.

C1246: **A new filter for long memory time series**
Presenter: **Adriana Cornea-Madeira**, University of York, United Kingdom
Co-authors: Joao Madeira
Macroeconomic and financial time series may have long memory (that is, are integrated of order larger than zero but smaller than one). In such series deviations from the long-run mean decline slower than exponential decay. We study how the properties of time series which display long memory are affected by the application of filters (such as the Hodrick-Prescott and Butterworth) which extract cyclical and trend components. Without relying on any model assumptions, we then propose a new filter designed to take into account for the possible presence of long memory and apply it to unemployment, current account and price-dividend ratio time series.

CG533 Room O2 CONTRIBUTIONS IN INFLATION Chair: Tomasz Lyziak

C1590: **Inflation comovements in advanced economies**
Presenter: **Luis J Alvarez**, Bank of Spain, Spain
Co-authors: Lola Gadea, Ana Gomez-Loscos
Although there is a vast literature on GDP comovement across countries, there is scant evidence on inflation comovements. To fill this gap we use a dataset on consumer price inflation that takes into account heterogeneity in price developments. We consider not headline inflation and also, but also core measures, energy and food prices. We analyze comovements across all advanced economies and across euro area countries. Synchronization is measured using the Moran-Stock-Watson spatial correlation index. We find that inflation comovements among advanced economies are quite relevant, but smaller than for GDP. Inflation comovements among euro area countries are higher than for advanced economies as a whole. Comovements have tended to increase over time, possibly reflecting the role of growing trade integration and the role of a common euro area monetary policy. Comovements among countries are highest for the energy component and lowest for core prices. We also consider bandpass filtered series to take into account that inflation has different drivers across frequency bands. We find that for high frequencies the degree of comovement is fairly low, whereas it is highest for the medium run, when the Phillips curve mechanism is expected to be strongest. Trend inflation also shows a sizable degree of comovememt, although it is lower than for long-run GDP fluctuations.

C1333: **Price convergence in the European Union: What has changed**
Presenter: **Aleksandra Halka**, Narodowy Bank Polski, Poland
Co-authors: Agnieszka Leszczynska-Paczesna
After a period of nominal convergence across EU countries, we have observed a significant weakening of the process in recent years. We explore this change. We use disaggregated price level indices to analyse the price convergence process in European countries in the period of 1999-2016. We estimate the beta and sigma convergence as well as distinguish the main drivers of this process. The results indicate that the cessation of the nominal convergence coincided with the outbreak of the global financial crisis in 2008. For some of the subaggregates in the consumption basket convergence simply declined, but for some other ones a statistically significant divergence seems to have come in place. We believe that the main factor contributing to the slowdown of the convergence process may be the slowdown of the real convergence after 2008, a development not lost on some other authors. This shift in paradigm while particularly pronounced for the old member states (EU15) was also discernible for some new members. Other factors that might have affected the convergence process include the decline in the speed of the deregulation among EU countries after the enlargement in and a downturn in the continued opening of the analysed economies.

C1387: **On the estimation of the Phillips curve for the Russian economy**
Presenter: **Andrey Zubarev**, Russian Presidential Academy of National Economy and Public Administration, Russia
The focus is on the estimation of the hybrid New-Keynesian Phillips curve with different kinds of price indices for the Russian economy. Three different measures of inflation are based on the following indices: the GDP deflator, the CPI, and the GDP deflator, net of exports. The main method of estimation is the continuously updating general method of moments (CUE), which has a smaller bias on finite samples and more valid values of the Hansen J-test for the overidentification compared to the standard GMM. The main result is that it is the dynamics of inflation calculated on the basis of the GDP deflator, net of exports, that is best described by the Phillips curve equation, and that the output gap is significant and has a positive sign which is consistent with the theory. Mayhap, this is due to the fact that the prices of imported and exported goods are not explicitly included into this measure of inflation. That is, it can be referred to as some "internal" inflation. This type of inflation is particularly new for the Phillips curve literature. Another important result is a slightly greater weight of forward-looking expectations in the formation of the inflation process.

C1428: **Sources of inflation comovements**
Presenter: **Karol Szafranek**, Warsaw School of Economics (Szkola Glowna Handlowa w Warszawie), Poland
The principal aim is to unravel the evolution of the inflation co-movements between economies and to quantify the sources of these fluctuations. To this end, in the first step a dynamic conditional correlation model is estimated enabling the quantification of the evolving CPI inflation correlation between countries in time. In the second step a Bayesian structural vector autoregression model is proposed to explain the variability of the estimated inflation correlation between country's inflation rate and global inflation. Preliminary results for the small open economy of Poland suggest that an increase in risk aversion and divergences in monetary policy decrease correlation of inflation, while an increase in the volatility of oil prices and country's openness increase this dependency measure. The influence of the business cycle synchronization index remains ambiguous.

CG093 Room P2 CONTRIBUTIONS IN BAYESIAN ECONOMETRICS Chair: Richard Gerlach

C0372: **A randomized missing data approach to robust filtering with applications to economics and finance**
Presenter: **Pawel Szerszen**, Federal Reserve Board of Governors, United States
Co-authors: Dobrislav Dobrev, Derek Hansen
A simple new approach is put forward to robust filtering of state-space models, motivated by the idea that the inclusion of only a small fraction of available highly precise measurements can still extract most of the attainable efficiency gains for filtering latent states, estimating model parameters, and producing out-of-sample forecasts. The new class of particle filters we develop aims to achieve a degree of robustness to outliers and model misspecification by purposely randomizing the subset of utilized highly precise but possibly misspecified or outlier contaminated data measurements, while treating the rest as if missing. The arising robustness-efficiency trade-off is controlled by varying the fraction of randomly utilized measurements or the incurred relative efficiency loss from such randomized utilization of the available measurements. As an empirical illustration, we consider popular state space models for inflation and equity returns with stochastic volatility and document favorable performance of our robust particle filter and density forecasts on both simulated and real data. More generally, our randomization approach makes it easy to robustly incorporate highly informative but possibly contaminated modern big data streams for improved state-space filtering and forecasting.

C1583: **Efficient Bayesian estimation of the stochastic volatility model with leverage**
Presenter: **Darjus Hosszejni**, WU Vienna University of Economics and Business, Austria
Co-authors: Gregor Kastner
The sampling efficiency of MCMC methods in Bayesian inference for stochastic volatility (SV) models is known to highly depend on the actual parameter values, and the effectiveness of samplers based on different parameterizations differs significantly. We derive novel samplers for the centered and the non-centered parameterizations of the practically highly relevant SV model with leverage, where the return process and the innovations of the volatility process are allowed to correlate. Moreover, based on the idea of ancillarity-sufficiency interweaving, we combine the resulting samplers in order to achieve superior sampling efficiency, irrespective of the baseline parameterization. The method is implemented using R and C++. Furthermore, we carry out an extensive comparison to already existing sampling methods for this model.

C1640: **Real-time forecasts of Henry Hub natural gas prices**
Presenter: **Arthur Thomas**, IFP Energies nouvelles-University of Nantes, France
Co-authors: Benoit Sevi
Using some hand-collected data from the monthly energy review dating back to 1997, a monthly real-time dataset is constructed to generate forecasts of natural gas prices as established at the Henry Hub. We compare the performance of a variety of models including state-of-the-art Bayesian and time-varying-parameters (TVP) models. Considering extensions to possible structural breaks and non-Gaussian distributions along with volatility models, we provide evidence that Bayesian and TVP models help in forecasting the real price of natural gas for horizons up to one year.

C1562: **A Bayesian inference approach to the inverse problems in the financial markets**
Presenter: **Yasushi Ota**, Okayama University of Science, Japan
A type of arbitrage model is explained. Financial derivatives are contracts wherein payment is derived from an underlying asset such as a stock, bond, commodity, interest, or exchange rate. Black and Sholes first found how to construct a dynamic portfolio of the derivative security, and by using Ito's lemma and the absence of arbitrage opportunities, the stochastic behavior of the derivative security is governed by the parabolic type partial differential equation and a suitable initial condition. Their approach is developed in probability theory, and the hedging and pricing theory of the derivative security is established as mathematical finance. However, as shown in deriving the Black-Scholes model, under the no-arbitrage property of the financial market, the real drift does not enter the equation. Taking this into account, we have derived the following new model. Next we illustrate our new mathematical approach, and finally, by using the numerical algorithm and MCMC method to our inverse problem, we confirm that using market prices of options with different strike prices enables us to identify the term structure of local volatility and real drift.

Saturday 15.12.2018 10:35 - 12:40 Parallel Session G – CFE-CMStatistics

EO482 Room Aula 5 RECENT ADVANCES IN THE ANALYSIS OF COMPLEX DATA Chair: Ci-Ren Jiang

E0307: **Inverse regression for multivariate functional data: Application to renewable energy forecast**
Presenter: **Ci-Ren Jiang**, Academia Sinica, Taiwan
Co-authors: Lu-Hung Chen
Inverse regression is an appealing dimension reduction method for regression models with multivariate covariates. Recently, it has been extended to the cases with functional or longitudinal covariates. However, the extensions simply focus on one single functional or longitudinal covariate. Motivated by a real application, we extend functional inverse regression to the cases with multiple functional covariates, whose domains could be different. The asymptotic properties of the proposed estimators are investigated for both functional and longitudinal cases. The computational issues are taken care with data binning, the fast Fourier transformation and random projections on a multi-core computation platform. In addition to simulation studies, the proposed approach is applied to predict the wind power capacity factor of the next day with the weather forecasts made today. Both demonstrate the good performance of our method.

E0860: **Use of multistate model for multiple endpoints in oncology clinical trials analysis and designs**
Presenter: **Chen Hu**, Johns Hopkins University, United States
In oncology clinical trials, disease progressions are most commonly captured through a series of sequentially observed events, such as cancer recurrence and deaths. The relationship between covariate (e.g., therapeutic intervention), recurrence, and death is often of interest, as it may provide key insights of optimal treatment decisions and future study designs. However such investigation is often complicated by the latency of disease progression leading to undetected or missing progression-related events. We consider a progressive multistate model with a frailty modeling the association between progression and death, and propose a semiparametric regression model for the joint distribution. An Expectation Maximization (EM) approach is used to derive the maximum likelihood estimators of covariate effects on both endpoints, the probability of missing progression event, as well as the parameters involved in the association. The asymptotic properties of the estimators are studied using theory of martingale and empirical process. We evaluate the utility of the proposed model for data analysis and study design based on both Monte Carlo simulations and real data examples.

E1035: **A computationally efficient algorithm for random effects selection in linear mixed models**
Presenter: **Mihye Ahn**, University of Nevada Reno, United States
The random effects selection has been received little attention in the literature. In linear mixed models, several methods for random effects selection have been proposed. However due to computationally intensive tasks, it is limited to apply the existing methods in practice. We propose two approximate methods of the moment-based method for random effects selection. The exact moment-based method has two challenging computation issues: nonlinear semidefinite programming and nonlinear programming with a linear inequality constraint. In particular, the most time-consuming step is the second computation to produce sparse solutions of the variance-covariance matrix of random effect factors. Since the objective function has up to fourth order terms and it makes the computation tedious, we suggest using a linear approximation to the penalized variance-covariance matrix. It reduces the objective function up to second order, and the quadratic programming can be easily implemented in some statistical software. By simulation studies, we show that the approximate methods also perform well and often outperform the exact method.

E1131: **Logistic regression augmented community detection**
Presenter: **Yunpeng Zhao**, Arizona State University, United States
Co-authors: Qing Pan, Chengan Du
When searching for gene pathways leading to specific disease outcomes, additional information on gene characteristics is often available that may facilitate to differentiate genes related to the disease from irrelevant background when connections involving both types of genes are observed and their relationships to the disease are unknown. We propose method to single out irrelevant background genes with the help of auxiliary information through a logistic regression, and cluster relevant genes into cohesive groups using the adjacency matrix. Expectation-maximization algorithm is modified to maximize a joint pseudo-likelihood assuming latent indicators for relevance to the disease and latent group memberships as well as Poisson or multinomial distributed link numbers within and between groups. A robust version allowing arbitrary linkage patterns within the background is further derived. Asymptotic consistency of label assignments under the stochastic blockmodel is proven. Superior performance and robustness in finite samples are observed in simulation studies. The proposed robust method identifies previously missed gene sets underlying autism related neurological diseases using diverse data sources including de novo mutations, gene expressions and protein-protein interactions.

E1227: **Weighted estimators of the complier average causal effect on restricted mean survival time**
Presenter: **Yun Li**, University of Michigan, United States
A major concern in any observational study is unmeasured confounding of the relationship between a treatment and outcome of interest. Instrumental variable (IV) analysis methods are able to control for unmeasured confounding. However, IV analysis methods developed for censored time-to-event data tend to rely on assumptions that may not be reasonable in many practical applications, making them unsuitable for use in observational studies. We develop weighted estimators of the complier average causal effect on the restricted mean survival time. The method is able to accommodate instrument-outcome confounding and adjust for covariate dependent censoring, making it particularly suited for causal inference from observational studies. We establish the asymptotic properties and derive easily implementable asymptotic variance estimators for the proposed estimators. Through simulation studies, we show that the proposed estimators tend to be more efficient than instrument propensity score matching based estimators or inverse probability of instrument weighted estimators. We apply our method to compare dialytic modality-specific survival for end stage renal disease patients using data from the United States Renal Data System.

EO588 Room Aula B MODERN APPROACHES TO HIGH DIMENSIONAL DATA ANALYSIS Chair: Jaroslaw Harezlak

E0542: **Hierarchical Bayesian models for integrating multimodal neuroimaging data**
Presenter: **Fengqing Zhang**, Drexel University, United States
The use of multimodal neuroimaging is a promising and recent approach to study complex brain disorders by utilizing complementary physical and physiological sensitivities. At the same time, however, the advent of multimodal neuroimaging has brought the need to analyze and integrate neuroimaging data with advanced statistical methods that can make full usage of their informational complexity. We aim to examine structural and functional brain changes specific to post-traumatic stress disorder (PTSD), a chronic and disabling anxiety disorder that can develop after a person is exposed to a traumatic event. Using data from the Philadelphia Neurodevelopmental Cohort (PNC) study, we identify three distinct groups, people with trauma exposure and no PTSD symptoms, people with trauma exposure and long-lasting PTSD symptoms as well as healthy controls. A large number of imaging features from different modalities including MRI, DTI, and resting-state fMRI are derived. We then develop hierarchical Bayesian models to combine heterogeneous data from multiple modalities and select predictive multimodal imaging signatures of PTSD.

E0629: **Predicting time-to-conversion to Alzheimer's disease using a longitudinal map of cortical thickness**
Presenter: **Ning Dai**, University of Minnesota, United States
Co-authors: Hakmook Kang, Galin Jones, Mark Fiecas

Prior studies have shown that cortical atrophy is associated with an increased risk of progression to clinical dementia. We take advantage of the longitudinal structural magnetic resonance imaging (MRI) data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) to investigate the relationship between conversion from mild cognitive impairment (MCI) to Alzheimer's disease (AD), and the dynamically changing cortical thickness over time and across the cortex of the brain. We develop a novel hierarchical Bayesian framework that embeds a spatial model on the cortical thickness effects within a survival model that predicts the timing of MCI-to-AD conversion. The proposed method allows for interpretation with respect to the temporal dynamics of imaging measurements, identifies the topographic patterns of anatomic regions where atrophy is associated with AD, and improves the performance of predicting AD onset by exploiting the spatial structure underlying the high-resolution observations over the cortical surface. Finally, we apply the proposed method to the longitudinal structural MRI data from ADNI to investigate how the impact of cortical thinning on the time to progress to AD varies across different regions of the brain.

E0751: **Two-sample tests for unweighted random graphs generated from latent space models**
Presenter: **Xixi Hu**, Indiana University Bloomington, United States
Co-authors: Michael Trosset

The problem of comparing graphs arises in many disciplines, including neuroscience, biology, and the social sciences. Treating graph comparison as a problem in statistical inference requires assuming a probability model that generates random graphs, e.g., a stochastic blockmodel or a latent space model in which each vertex is associated with a latent position and the probability of an edge between two vertices is a function of their latent positions. For latent space models, various methods have been suggested for testing the null hypothesis that two models with matched vertices are identical up to isometry. Previous work has emphasized the case in which one graph is generated by each model; it is not clear how to extend these methods to the case in which multiple graphs are generated by each model. If the edge probabilities are a known function of the Euclidean distance between the latent positions, then one can estimate two sets of common latent positions by metric multidimensional scaling (MDS) and construct a test statistic by Procrustes analysis. If the edge probability function is unknown but monotone, then one can use nonmetric MDS to construct scale-invariant representations of the common latent positions and proceed analogously. We investigate this procedure through simulations and apply our approach to real brain data, comparing the structural brain networks of subjects with autism to those of healthy controls.

E0849: **Multiple latent components clustering**
Presenter: **Stanislaw Wilczynski**, University of Wroclaw, Poland
Co-authors: Piotr Sobczyk, Malgorzata Bogdan, Julie Josse

In many scientific problems such as identification of genetic pathways based on gene expression data, one of the tasks is finding a lower dimensional subspace representing a collection of points from high-dimensional space. One of the simplest methods to achieve this is to use PCA. However, it is useful only if we assume that all points from the data come from the same lower dimensional subspace. In fact, in lots of cases a more general model is needed, which assumes that variables come from a mixture model and our high-dimensional space is a union of a few low dimensional subspaces. We propose a new method of finding the subspaces called Multiple Latent Components Clustering (MLCC). It is based on k-means algorithm, where clusters represent subspaces and the center of a cluster is a set of principal components. To estimate the number of clusters, modified version of Bayesian Information Criterion is used, which takes into account a prior distribution on a number of clusters. In each of the iterative steps of the algorithm, the number of principal components in a single cluster is estimated using Penalized Semi-integrated Likelihood (PESEL) method and the similarity between data point and cluster is measured by BIC. The algorithm is implemented in R package 'Varclust'. We will present the results of the comparison of our algorithm with other variable clustering methods, as well as results of real data analysis. We will point out the main differences and advantages of MLCC.

E0856: **Covariate assisted principal regression for covariance matrix outcomes**
Presenter: **Yi Zhao**, Johns Hopkins Bloomberg School of Public Health, United States

Modeling variances in data has been an important topic in many fields, including in financial and neuroimaging analysis. We consider the problem of regressing covariance matrices on a vector covariates, collected from each observational unit. The main aim is to uncover the variation in the covariance matrices across units that are explained by the covariates. The Covariate Assisted Principal (CAP) regression is introduced, an optimization-based method for identifying the components predicted by (generalized) linear models of the covariates. We develop computationally efficient algorithms to jointly search the projection directions and regression coefficients, and we establish the asymptotic properties. Using extensive simulation studies, the method shows higher accuracy and robustness in coefficient estimation than competing methods. Applied to a resting-state functional magnetic resonance imaging study, the approach identifies the human brain network changes associated with age and sex.

EO338 Room Aula Magna LARGE SCALE STATISTICAL INFERENCE: METHODOLOGY AND APPLICATIONS Chair: Tetyana Pavlenko

E0389: **On simultaneous confidence interval estimation for the difference of paired mean vectors in high-dimensional settings**
Presenter: **Masashi Hyodo**, Osaka Prefecture University, Japan
Co-authors: Hiroki Watanabe

To test whether two populations have the same mean vector in a high-dimensional setting, an unbiased estimator of the squared Euclidean distance between the mean vectors has been previously derived, and the asymptotic normality of this estimator has been proved under local assumptions about the mean vectors. These results are extended without assumptions about the mean vectors. In addition, asymptotic normality is established in a class of general statistics which includes important statistics under general moment conditions that cover both the previous moment condition and elliptical distributional assumption. These asymptotic results are applied to the construction of simultaneous intervals for all pair-wise differences between mean vectors of $k > 2$ groups. The finite-sample and dimension performance of the proposed methods is also studied via Monte Carlo simulations. The methodology is illustrated using microarray data.

E0453: **Testing independence in high-dimensional data: ρV-coefficient based approach**
Presenter: **Takahiro Nishiyama**, Senshu University, Japan
Co-authors: Masashi Hyodo, Tatjana Pavlenko

The problem of testing mutual independence of k high-dimensional random vectors is considered when the data are multivariate normal and $k \geq 2$ is a fixed integer. For this purpose, we focus on the vector correlation coefficient, ρV and propose an extension of its classical estimator which is constructed to correct potential sources of inconsistency related to the high dimensionality. Building on the proposed estimator of ρV, we derive the new test statistic and study its limiting behavior in a general high-dimensional asymptotic framework which allows the vector's dimensionality arbitrarily exceeds the sample size. Specifically, we show that the asymptotic distribution of the test statistic under the main hypothesis of independence is standard normal and that the proposed test is size and power consistent. Using our statistics, we further construct the step-down multiple comparison procedure for the simultaneous test for independence. Accuracy of the proposed tests in finite samples is shown through simulations for a variety of high-dimensional scenarios.

E0534: **Detection of non-null effects in linear models for sparse mixtures**
Presenter: **Annika Tillander**, Linkoping University, Sweden
Co-authors: Tatjana Pavlenko

For a linear classifier to be successful in a high-dimensional setting it is often needed to select a subset of features. This is a challenging task when the informative features are rare and weak. Accounting for the relation between features, given the sparse structure, can enhance the chances. It been shown that a block-diagonal approximation of the inverse covariance matrix lead to an additive classifier with good classification accuracy. This call for block-wise feature selection. A measure of information strength and a threshold is required. For single feature selection the Higher Criticism is a well-known thresholding method that is optimally adaptive i.e. performs well without knowledge of the sparsity and weakness parameters. It is shown how this method can be extended to handle thresholding for blocks of features. However, popular method it has limitations and will be compared to other goodness-of-fit tests based on sup-functionals of weighted empirical process for thresholding. The relevance and benefits in high-dimensional classification is demonstrated using both simulation and real data.

E0651: **Bayesian predictive inference in decomposable graphs using sequential Monte Carlo samplers**
Presenter: **Tetyana Pavlenko**, KTH Royal Institute of Technology, Sweden
Co-authors: Felix Rios

Bayesian predictive inference in the class of decomposable graphical models is considered within the classification framework. We present a multi-class graphical Bayesian predictive classifier that incorporates the uncertainty in the model determination into the standard Bayesian formalism. For each class, the dependence structure underlying the observed features is represented by a set of decomposable Gaussian graphical models. Emphasis is then placed on the Bayesian model averaging which takes full account of the class-specific model uncertainty by averaging over the posterior graph model probabilities. Even though the decomposability assumption severely reduces the model space, the size of the class of decomposable models is still immense, rendering the explicit Bayesian averaging over all the models infeasible. To address this issue, we consider the particle Gibbs strategy for posterior sampling from decomposable graphical models which utilize the Christmas tree algorithm as proposal kernel. We also derive the strong hyper Markov law which we call the hyper normal Wishart law that allows to perform the resultant Bayesian inference locally. The proposed predictive graphical classifier reveals superior performance compared to the ordinary Bayesian predictive rule that does not account for the model uncertainty, as well as to a number of out-of-the-box classifiers.

E0722: **Optimal recovery of sparse additive signals**
Presenter: **Natalia A Stepanova**, Carleton University, Canada
Co-authors: Cristina Butucea

The problem of exact and almost full recovery of an unknown multivariate signal f observed in a d-dimensional Gaussian white noise model is considered. We assume that f is smooth and has an additive sparse structure determined by the parameter s, the number of nonzero univariate signals contributing to f. We also assume that the dimension d increases and that the parameter s remains "small" relative to d. With these assumptions, the recovery problem becomes that of determining which sparse additive components of f are nonzero. The latter may be viewed as the problem of variable selection in high dimensions. We give conditions under which exact and almost full variable selections are possible and, in both regimes, we propose the best possible (in the asymptotically minimax sense) variable selection procedures. The proposed procedures are adaptive in the parameter s.

EO052 Room A1 ADVANCES IN LATENT VARIABLE MODELS FOR COMPLEX DATA Chair: Silvia Cagnone

E0560: **Model averaging weighted estimators for latent variable models in a longitudinal data setting**
Presenter: **Vassilis Vasdekis**, Athens University of Economics and Business/Research Center, Greece
Co-authors: Kostas Florios, Dimitris Rizopoulos, Irini Moustaki

In the latent variable setting for longitudinal data, different sets of weighted average estimators for model parameters are proposed. All of them are based on ideas coming from the composite likelihood. The first two are produced by minimizing the total variance of the resulting estimator. The weights are matrices either unrestricted or diagonal. The third estimator uses univariate weights. For their construction we propose a sequence of models each of which provides estimates for model parameters and a composite likelihood information criterion. The latter is used to evaluate the weights. The estimator seems to perform better than the other two estimators providing variance components with high coverage, especially when the number of time points at which measurements are obtained is large, regardless of the size of the cluster size.

E0821: **Generalized linear latent variable models in the analysis of ecological data**
Presenter: **Sara Taskinen**, University of Jyvaskyla, Finland
Co-authors: Jenni Niku

Very high-dimensional multivariate abundance data, which consist of records (counts, presence-absences, biomass) of a large number of interacting species at a set of units or sites, are very common in ecological studies. When analysing such multivariate abundance data the interest is often in visualisation of correlation patterns across taxa, hypothesis testing of environmental effects and making predictions for abundances. In several recent studies, a model-based joint analysis is shown to be a promising method for studying non-normal multivariate abundance data. One particular approach is the use of generalized linear latent variable models. These are constructed by fitting generalized linear models to each species, while including latent variables to account for residual correlation between species, for example, due to unmeasured covariates. Notice that in other fields, e.g in psychometrics and social sciences such models have been popular for a long time, however models used in such fields are not often suitable for ecological data. Some latest developments in the field of GLLVMs applied to ecological studies are discussed. The theory is illustrated using examples.

E0826: **A regularized estimation approach for the three-parameter logistic model**
Presenter: **Michela Battauz**, University of Udine, Italy
Co-authors: Ruggero Bellio

The three-parameter logistic model is an item response theory model used with dichotomous items. It is well known that the parameters of the model are weekly identifiable and that the maximization of the likelihood, which is performed using numerical algorithms, is prone to convergence issues. We propose the use of a penalized likelihood for the estimation of the item parameters. Two main approaches are explored. The first approach is based on the inclusion of a penalty term on the guessing parameters in the likelihood function. In particular, as penalty we consider the normal density or a ridge-type penalty. The tuning parameters are selected through cross validation. Model-based shrinkage estimation constitutes the second approach explored, which is pursued employing the bias-reduction methodology. The performance of the methods is investigated through simulation studies and a real data example. All the methods lead to shrinkage of the guessing parameter estimates, showing the usual bias-variance tradeoff of regularized methods. The bias-reduction method presents a smaller amount of shrinkage.

E0857: **On the use of latent variables to extend Gaussian mixture models**
Presenter: **Geoffrey McLachlan**, University of Queensland, Australia

Hiding behind the structure of finite mixture models are the latent variables that define the component labels in the conceptualization of a mixture distribution as applying in the case where the observed random variable is selected from one of the component distributions with (prior) probabilities specified by the mixing proportions. The inclusion of additional latent variables can further extend the flexibility of finite mixture distributions to model complex data. These additional latent variables include the latent factors in the case of mixtures of factor models and their deep versions and also the latent skewing variables in the case of mixtures of skew-symmetric component distributions. Various examples are presented to illustrate the improvement in model fit by adopting this use of latent variables.

E0911: **An algebraic estimator for large spectral matrices**
Presenter: **Matteo Farne**, University of Bologna, Italy
Co-authors: Matteo Barigozzi

A method is presented to estimate a large p-dimensional spectral matrix assuming that the data follow a dynamic factor model with a sparse residual. In specific, we apply a nuclear norm plus l_1 norm heuristics to any kernel input estimate at each frequency. We assume that the latent eigenvalues scale to p^{α}, $\alpha \in [0,1]$, and the sparsity degree scales to p^{δ}, with $\delta \leq \frac{1}{2}$ and $\delta \leq \alpha$. We prove that the algebraic recovery of latent rank and sparsity patterns is guaranteed if the smallest latent eigenvalue λ_r and the minimum residual nonzero entry in absolute value $\min_S$ are large enough across frequencies. The identifiability of the underlying matrix recovery problem requires the absolute convergence of latent and residual filters and a limited discrepancy among the eigenvectors of the factorial coefficients and the sparsity patterns of the residual coefficients across lags. The consistency of the input is derived via an appropriate weak dependence assumption both on factors and residuals. The recovery quality directly depends on the ratio $\frac{p^{\alpha}}{\sqrt{T}}$, where T is the sample length, and the magnitude of T is required to be $p^{3\delta}$ or larger. In a wide simulation study, we stress the crucial role of λ_r and $\min_S$ across frequencies, highlighting the conditions which cause our method to fail.

EO178 Room Aula A MARKOV SWITCHING REGRESSION AND HIDDEN MARKOV MODELS Chair: Thomas Kneib

E0547: **Modeling non-stationary operational risk: A smooth-transition distributional regression approach**
Presenter: **Julien Hambuckers**, University of Liege, Belgium
Co-authors: Thomas Kneib

The distribution of operational losses is particularly challenging to model, due to the high probability of extremes and the existence of time-varying structural dependencies. In particular, operational loss severity distribution is often concerned with changes in regulations, business cycles or financial crises that affect the dependence structure with potential predictors. To help accounting for this empirical feature, we introduce smooth-transition (ST) Generalized Pareto (GP). In this time-varying regression model, the parameters of the GP distribution are related to explanatory variables through a regression function, which depends itself on a time-varying predictor of structural changes. First, we discuss the computational challenges associated to this class of models. Then, we propose several estimation strategies and investigate their finite sample properties in a simulation study. Eventually, we use our findings to study the time-varying dependence structure of monthly operational risks with market volatility and past extreme events.

E0842: **Coupled Markov-switching regression models with application to electronic health record data**
Presenter: **Jennifer Pohle**, Bielefeld University, Germany
Co-authors: Roland Langrock, Ruth King, Mihaela van der Schaar

Hidden Markov models (HMMs) are time series models which assume the observations to depend on an underlying unobserved Markov chain with finitely many states. They have been applied in many different areas, for instance in speech recognition, finance, medicine, and ecology. In the case of multivariate time series, within a basic HMM formulation, the variables would be expected to evolve synchronously in the sense that they are driven by the same underlying state sequence. However, in some applications, e.g. in medicine, the observed variables do not necessarily evolve in lockstep, although they may be correlated. Coupled hidden Markov models overcome this limitation by assuming separate but correlated state sequences to underlie the different variables observed, hence coupling the state processes of multiple HMMs. However, the observations often depend not only on the underlying state, but also on external factors. Therefore, we extend coupled HMMs to allow for covariates in the observation processes, which leads to the flexible class of coupled Markov-switching regression models. We apply this method to electronic health record data collected for 842 patients within the medical intensive care unit at the University of California in Los Angeles.

E0863: **An analysis of a hot hand effect in professional darts using state-space models**
Presenter: **Marius Oetting**, Bielefeld University, Germany
Co-authors: Roland Langrock, Christian Deutscher, Vianey Leos-Barajas

The hot hand hypothesis is investigated in professional darts in a near-ideal setting with minimal to no interaction between players. Considering almost one year of tournament data, corresponding to 167,492 dart throws in total, state-space models are used to investigate serial dependence in throwing performance. In these models, a latent state process serves as a proxy for a player's underlying ability, and autoregressive processes are used to model how this process evolves over time. The results show a strong but short-lived serial dependence in the latent state process, thus providing evidence for the existence of the hot hand.

E0897: **Joint modeling of multi-scale time series data using hierarchical hidden Markov models**
Presenter: **Timo Adam**, Bielefeld University, Germany
Co-authors: Vianey Leos-Barajas, Roland Langrock

Hidden Markov models are prevalent in ecology and economics, where they are widely used to model time series data subject to state-switching over time. A basic hidden Markov model comprises an observed state-dependent process that is driven by a hidden state process, the latter of which is typically linked to behavioral modes of an animal (such as resting, foraging or traveling) or economic market regimes (such as recessions or periods of economics growth). To allow for meaningful inference, observations need to be equally spaced in time (or otherwise regularly sampled). However, in animal movement modeling, telemetry sensors often collect data from the same individual at different scales. Typical examples are step lengths obtained from GPS tags every hour, dive depths obtained from time-depth recorders once per dive or accelerations obtained from accelerometers several times per second. Similarly, in economics, stock market data are often collected at a daily (or even finer) scale, whereas macroeconomic indicators are typically observed on a monthly, quarterly or yearly basis. To account for differing temporal resolutions across multiple variables as well as to allow for joint inference at multiple scales, we consider hierarchical hidden Markov models, where the observations are regarded as stemming from several, connected hidden state processes, each of which operates at the scale at which the corresponding variables were observed.

E0290: **Time-specific clustering via rectangular latent Markov models, with an analysis of the well-being of nations**
Presenter: **Alessio Farcomeni**, Sapienza - University of Rome, Italy
Co-authors: Gordon Anderson, Maria Grazia Pittau, Roberto Zelli
In longitudinal model-based clustering methods the number of groups is usually fixed over time, apart from (mostly) heuristic approaches. We propose a latent Markov model admitting variation in the number of latent states. The consequence is that (i) subjects can switch from one group to another at each time period and (ii) the number of groups can change at each time period. Clusters can merge, split, or be re-arranged. For a fixed sequence of the number of groups, inference is carried out through maximum likelihood, using appropriate forward-backward recursions. A penalized likelihood form is introduced to simultaneously choose an optimal sequence for the number of groups and cluster subjects. The penalized likelihood is optimized through a novel expectation-maximization-Markov-Metropolis algorithm. The motivation arises from an analysis of the progress of well-being of nations, as measured by the three dimensions of the Human Development Index over the last 25 years. The main findings are that (i) transitions among nation clubs are scarce, and mostly linked to historical events (like dissolution of USSR or war in Syria) and (ii) there is mild evidence that the number of clubs has shrunk over time, where we have four clusters before 2005 and three afterwards. In a sense, nations are getting more and more polarized with respect to standards of well-being. R code is available at https://github.com/afarcome/LMrectangular.

EO344 Room Aula C OPTIMISATION FOR MACHINE LEARNING AND ONLINE METHODS Chair: Stephane Gaiffas

E1206: **On the optimality of the standard Hedge algorithm in the stochastic setting**
Presenter: **Jaouad Mourtada**, Ecole polytechnique, France
Co-authors: Stephane Gaiffas
Consider the standard problem of prediction with expert advice: at each step, a learner chooses a probability distribution on a set of M experts, then experts' losses are revealed, and the learner suffers the average loss of the experts under its chosen distribution. The goal is to control the learner's regret, defined as the difference between its cumulated loss and that of the best expert. Arguably the most well-known strategy is the exponential weights or Hedge algorithm, which has optimal $\sqrt{T \log M}$ regret in the worst-case, when experts' losses may be arbitrary and set by an adversary. Departing from the pessimistic worst-case analysis, recent work has sought to design algorithms combining minimax regret with improved guarantees on easier instances. We show that, surprisingly, the standard variant of Hedge with decreasing learning rate already exhibits such adaptivity. Indeed, we obtain a constant regret bound for Hedge in the stochastic case, when loss vectors are iid. Further, this bound has the optimal dependence on the number of experts and on the suboptimality gap between the leading expert and the rest, meaning that Hedge automatically adapts to this hardness parameter. Finally, we contrast this version of Hedge with the fixed learning rate and "doubling trick" variants, both of which fail to adapt to the stochastic case.

E1356: **Non-asymptotic analysis of local-SGD**
Presenter: **Aymeric Dieuleveut**, EPFL, Switzerland
Mini-batch stochastic gradient descent methods are the current state-of-the-art in large-scale distributed machine learning. Such methods are limited by the massive communication bottleneck they introduce. In this scenario, using a "local-SGD" model, where machines communicate their independent models frequently, might be preferred over the "large mini-batch" schemes to balance the computation-communication trade-off. A surprising lack of understanding of the behavior of these local-SGD methods is our primary motivation. We propose a simple non-asymptotic error analysis, which enables comparison at one extreme to one-shot averaging i.e., a single communication round among independent workers and at another to mini-batch averaging i.e., communicating at every step while underlining the computation-communication-performance tradeoffs. This is one of the first non-asymptotic analyses for local-SGD. We also give adaptive lower bounds on the frequency of communication, for local-SGD to perform optimally, i.e., as good as mini-batch averaging. Our results encompass ubiquitous algorithms like least-squares and logistic regression in the online as well as finite horizon setting. Moreover, they work well with large step-sizes and provide insights on the actual behavior of the stochastic process.

E1518: **Dual optimization for convex constrained objectives without the gradient-Lipschitz assumption**
Presenter: **Martin Bompaire**, Ecole polytechnique, France
Co-authors: Stephane Gaiffas, Emmanuel Bacry
The minimization of convex objectives coming from linear supervised learning problems, such as penalized generalized linear models, can be formulated as finite sums of convex functions. For such problems, a large set of stochastic first-order solvers based on the idea of variance reduction are available and combine both computational efficiency and sound theoretical guarantees (linear convergence rates). Such rates are obtained under both gradient-Lipschitz and strong convexity assumptions. Motivated by learning problems that do not meet the gradient-Lipschitz assumption, such as linear Poisson regression, we work under another smoothness assumption, and obtain a linear convergence rate for a shifted version of Stochastic Dual Coordinate Ascent (SDCA) that improves the current state-of-the-art. Our motivation for considering a solver working on the Fenchel-dual problem comes from the fact that such objectives include many linear constraints, that are easier to deal with in the dual. Our approach and theoretical findings are validated on several datasets, for Poisson regression and another objective coming from the negative log-likelihood of the Hawkes process, which is a family of models which proves extremely useful for the modeling of information propagation in social networks and causality inference.

E1548: **On the properties of sign estimators derived from hard thresholded lasso and hard thresholded basis pursuit**
Presenter: **Patrick Tardivel**, University of Wroclaw, Poland
In the high-dimensional linear model, when the number of observations is lower than the number of explanatory variables, we aim at estimating the sign of the model. It is well known that the irrepresentable condition is a necessary and "almost" sufficient condition to recover exactly the sign of the model with the lasso sign estimator. In a first step, we provide a new result about the irrepresentable condition: the probability to recover the sign of the model with the lasso sign estimator is smaller than 1/2 once the irrepresentable condition does not hold. Consequently, there is an issue to provide a sign estimator able to recover the sign of the model under a weaker assumption than the irrepresentable condition. In a second step, we show that sign estimators derived from hard thresholded lasso and hard thresholded basis pursuit only need identifiability condition to recover exactly the sign of the model. Because the identifiability condition is a weaker condition than the irrepresentable condition, these sign estimators are theoretically better than the lasso sign estimator. Finally, the irrepresentability and identifiability curves, function of the signal sparsity, show that the gap between the irrepresentable condition and the identifiability condition is huge. That is the reason why sign estimators derived from hard thresholded lasso and hard thresholded basis pursuit outperform the lasso sign estimator.

E1529: **A minimax near-optimal algorithm for adaptive rejection sampling**
Presenter: **Joseph C Lam**, Magdeburg University, Germany
Co-authors: Alexandra Carpentier, Gilles Blanchard, Juliette Achdou
Rejection sampling is a fundamental Monte-Carlo method. It is used to sample from distributions admitting a probability density function which can be evaluated exactly at any given point, albeit at a high computational cost. However, without proper tuning, this technique implies a high rejection rate. Several methods have been explored to cope with this problem, based on the principle of adaptively estimating the density by a simpler function, using the information of the previous samples. Most of them either rely on strong assumptions on the form of the density, or

do not offer any theoretical performance guarantee. We give the first theoretical lower bound for the problem of adaptive rejection sampling and introduce a new algorithm which guarantees a near-optimal rejection rate in a minimax sense.

E1727: **Learning with implicit regularization and sketching**
Presenter: **Luigi Carratino**, University of Genoa, Italy
Co-authors: Lorenzo Rosasco
Classically, regularization is achieved imposing explicit constraints on a data fit term, and optimization aspects are considered separately. We discuss how regularization can be controlled implicitly by an optimization method of choice. This latter approach has the advantage that training time controls at the same time statistical accuracy and time complexity of the obtained estimator. Moreover, we discuss how computations can be further reduced considering random projections. Our study bridges optimization and statistical studies.

EO619 Room B1 CAUSALITY: MODELING, REASONING, ESTIMATION AND PREDICTION I Chair: Vanessa Didelez

E1214: **Causal discovery with non-Gaussian data and latent variables**
Presenter: **Y Samuel Wang**, University of Chicago, United States
Estimating causal structure is considered from multivariate observational data, possibly with latent confounding. Specifically, we assume the data is generated by a linear structural equation model with non-Gaussian errors. We show that if the true structure corresponds to a bow-free acyclic path diagram, then the exact causal structure–not just an equivalence class–can be identified. In particular, we propose a constraint based algorithm, BANG (Bow-free Acyclic Non-Gaussian) which consistently estimates the underlying graph. If the maximum in-degree of the graph is bounded, then the number of tests required by BANG is bounded by a polynomial of the number of observed variables.

E0588: **Direct estimation of differences in causal graphs**
Presenter: **Yuhao Wang**, MIT, United States
Co-authors: Chandler Squires, Anastasiya Belyaeva, Caroline Uhler
The problem of estimating the differences between two causal directed acyclic graph (DAG) models given i.i.d. samples from each model is considered. This is of interest for example in genomics, where changes in the structure or edge weights of the underlying causal graphs reflect alterations in the gene regulatory networks. We provide the first provably consistent method for directly estimating the differences in a pair of causal DAGs without separately learning two possibly large and dense DAG models and computing their difference. Our two-step algorithm first uses invariance tests between regression coefficients of the two data sets to estimate the skeleton of the difference graph and then orients some of the edges using invariance tests between regression residual variances. We demonstrate the properties of our method through a simulation study and apply it to the analysis of gene expression data from ovarian cancer and during T-cell activation.

E1267: **Identifiability of total causal effects from observational data**
Presenter: **Emilija Perkovic**, ETH Zurich, Switzerland
Co-authors: Markus Kalisch, Marloes Maathuis
One of the most commonly used methods for estimating total causal effects from graphs learned from observational data is covariate adjustment. Previously, we developed a graphical criterion that is sound and complete for covariate adjustment in graphs learned from observational data. However, not all total causal effects are identifiable from observational data. Furthermore, not all total causal effects that are identifiable from observational data are identifiable through covariate adjustment. We outline how to improve the identifiability of total causal effects with covariate adjustment through the addition of background knowledge and discuss some preliminary results on the gap between covariate adjustment and identifiability of total causal effects from observational data.

E0811: **Graphical criteria for efficient covariate adjustment in causal linear structural equation models**
Presenter: **Leonard Henckel**, ETH Zurich, Switzerland
Co-authors: Emilija Perkovic, Marloes Maathuis
When the underlying causal graph is known, there exists a sound and complete graphical criterion for when a covariate set allows for unbiased causal effect estimation. However, typically a large number of sets fulfils this criterion. Restricting ourselves to the causal linear structural equation model setting, we introduce graphical criteria that allow one to identify in many cases which of two valid adjustment sets provides the smaller asymptotic variance. Even though this result only induces a partial ordering, it can be used to identify a set that always provides the optimal asymptotic variance. Alternatively, when this set cannot be used, the graphical criterion can also be used for strictly beneficial pruning.

E1183: **Detection of instrumental variables in causal models: An algorithmic framework**
Presenter: **Maciej Liskiewicz**, University of Luebeck, Germany
Co-authors: Benito van der Zander
Instrumental variables (IVs) are widely used tool to identify causal effects from observational data. For this purpose IVs have to be exogenous, i.e., causally unrelated to all variables in the model except the explanatory variable. Since these IVs do not exist in many model instances, the approach has been generalized to conditional IVs that only require exogeneity conditioned on a set of covariates. Another generalization is to use instrumental sets that allow us to identify causal effects if no single instrument exists. This has led to a wider choice of potential IVs. However, a significant barrier to the applications of this method is of algorithmic nature: So far, it was not clear whether such generalized IVs can be found efficiently. We address two issues with generalized IVs. First, we discuss new natural concepts of IVs, which interpolate between the existing notions. Second, we provide effective algorithms for detection of such IVs and show NP-hardness for the most generalized levels. Together this implies a complete and constructive solution to causal effect identification using IVs in linear causal models.

EO186 Room C1 MODEL SPECIFICATION TESTS Chair: Maria Dolores Jimenez-Gamero

E0339: **New consistent goodness-of-fit tests based on V-empirical Laplace transforms**
Presenter: **Bojana Milosevic**, University of Belgrade, Serbia
Co-authors: Marija Cuparic, Marko Obradovic
New goodness-of-fit tests are proposed which employ the equidistribution characterizations. Based on V-empirical Laplace transforms of equidistributed statistics, test statistics of L^p-type are constructed. Their asymptotic properties are derived. To assess their quality, the approximate Bahadur efficiency is used. For small sample sizes, a simulated power study is performed. The tests are shown to be very efficient and powerful in comparison to some common competitors.

E0730: **New goodness-of-fit tests based on the functionals of U-empirical processes**
Presenter: **Ksenia Volkova**, Saint-Petersburg State University, Russia
New goodness-of-fit tests are proposed. Tests statistics are functional of U-empirical processes. We discuss limiting distributions of new tests and find the logarithmic large deviation asymptotics of test statistics under null-hypothesis. In order to select the most sensitive test we compare them

by the asymptotic method using the concept of asymptotic relative efficiency against some common alternatives. Conditions of the local asymptotic optimality of new statistics are given.

E0733: **Modeling social preferences by using penalized phi-divergence measures**
Presenter: **Virtudes Alba-Fernandez**, University of Jaen, Spain
Co-authors: Maria Dolores Jimenez-Gamero, Francisca Jimenez-Jimenez
A test is proposed for the model selection problem for multinomial data based on penalized phi-divergences. The use of penalized phi-divergences is motivated by the presence of zero frequencies. An application to the selection of models to classify individuals according to their social preferences, say, altruistic, cooperative, individualistic, competitive and aggressive, according to their behaviour in a decomposed game, is used to exemplify the proposed technique.

E1145: **A characteristic function based test for serial independence in vector autoregressive models**
Presenter: **James Allison**, Northwest University, South Africa
Co-authors: Simos Meintanis, Joseph Ngatchou-Wandji
Tests for serial independence of an arbitrary finite order for the innovations in vector autoregressive models are considered. The tests are expressed as L_2–type criteria involving the difference of the joint empirical characteristic function and the product of corresponding marginals. Asymptotic as well as Monte-Carlo results are presented.

E1223: **A model specification test for the variance function in nonparametric regression**
Presenter: **Juan-Carlos Pardo-Fernandez**, Universidade de Vigo, Spain
Co-authors: Maria Dolores Jimenez-Gamero
The problem of testing for the parametric form of the conditional variance is considered in a fully nonparametric regression model. A test statistic based on a weighted L_2-distance between the empirical characteristic functions of residuals constructed under the null hypothesis and under the alternative is proposed and studied theoretically. The null asymptotic distribution of the test statistic is obtained and employed to approximate the critical values. Finite sample properties of the proposed test are numerically investigated in several Monte Carlo experiments. The developed results assume independent data. Their extension to dependent observations is also discussed.

EO514 Room D1 RECENT ADVANCES IN NONPARAMETRIC METHODS Chair: Marinho Bertanha

E1274: **A rational approach to inference on multiple parameters**
Presenter: **Brigham Frandsen**, Brigham Young University, United States
An approach is proposed to inference on multiple parameters that produces valid simultaneous inferences on multiple parameters while maintaining precision on the parameter or parameters of greatest interest. The approach allows inference to reflect differing preferences for precision that the researcher may have across parameters, resulting in hypothesis tests that are more powerful and confidence regions with shorter projections on the parameters that the researcher cares more about, while remaining jointly valid across all parameters. A researcher using the procedure specifies in advance non-negative weights that correspond to the relative preference for precision across parameters. The proposed confidence procedure chooses a confidence region to minimize the weighted sum of the projections on the parameter dimensions. A decision theoretic framework presents axioms for researcher preferences under which the proposed procedure is optimal.

E1275: **Inference for moments of ratios with robustness against large trimming bias and unknown convergence rate**
Presenter: **Yuya Sasaki**, Vanderbilt University, United States
Co-authors: Takuya Ura
Statistical inference for moments of the form $E[B/A]$ is considered. A naive sample mean is unstable with small denominator A. A method of robust inference is developed, and a data-driven practical choice of trimming observations with small A is proposed. Our sense of the robustness is twofold. First, bias correction allows for robustness against large trimming bias. Second, adaptive inference allows for robustness against unknown convergence rate. The proposed method allows for closer-to-optimal trimming, and more informative inference results in practice. This practical advantage is demonstrated for inverse propensity score weighting through simulation studies and real data analysis.

E1285: **External validity in fuzzy regression discontinuity designs**
Presenter: **Marinho Bertanha**, University of Notre Dame, United States
Co-authors: Guido Imbens
Fuzzy regression discontinuity designs identify the local average treatment effect (LATE) for the subpopulation of compliers, and with forcing variable equal to the threshold. We develop methods that assess the external validity of LATE to other compliance groups at the threshold, and allow for identification away from the threshold. Specifically, we focus on the equality of outcome distributions between treated compliers and always-takers, and between untreated compliers and never-takers. These equalities imply continuity of expected outcomes conditional both on the forcing variable and the treatment status. We recommend that researchers plot these conditional expectations and test for discontinuities at the threshold to assess external validity.

E1286: **A correction for regression discontinuity designs with group-specific mismeasurement of the running variable**
Presenter: **Otavio Bartalotti**, Iowa State University, United States
Co-authors: Steven Dieterle, Quentin Brummet
When the running variable in a regression discontinuity (RD) design is measured with error, identification of the local average treatment effect of interest will typically fail. While the form of this measurement error varies across applications, in many cases there is a group structure to the measurement error. We develop a procedure to make use of this group-specific measurement error structure to correct estimates obtained in a regression discontinuity framework using auxiliary data. This procedure extends the prior literature on measurement error on the running variable by leveraging auxiliary information in order to account for more general forms of measurement error. Additionally, we develop adjusted asymptotic variance and standard errors that take in consideration the variability introduced by the nonparametric estimation of nuisance parameters from auxiliary data. Simulations provide evidence that the proposed procedure adequately corrects for measurement error introduced bias and tests using the new adjusted formulas exhibit empirical coverage closer nominal test size than naive alternatives. We provide two empirical illustrations to demonstrate that correcting for measurement error can either reinforce the results of a study or provide a new empirical perspective on the data.

E1297: **Dealing with a technological bias: The difference-in-difference approach**
Presenter: **Dmitry Arkhangelskiy**, CEMFI, Spain
A nonlinear model is constructed for causal inference in the empirical settings where researchers observe individual-level data for a few large clusters over at least two time periods. It allows for identification (sometimes partial) of the counterfactual distribution, in particular, identifying average treatment effects and quantile treatment effects. The model is flexible enough to handle multiple outcome variables, multidimensional heterogeneity, and multiple clusters. It applies to the settings where the new policy is introduced in some of the clusters, and a researcher additionally has information about the pretreatment periods. We argue that in such environments we need to deal with two different sources of bias: selection and

technological. In my model, we employ standard methods of causal inference to address the selection problem and use pretreatment information to eliminate the technological bias. In case of one-dimensional heterogeneity, identification is achieved under natural monotonicity assumptions. The situation is considerably more complicated in the case of multidimensional heterogeneity where we propose three different approaches to identification using results from transportation theory.

EO162 Room E1 RECENT DEVELOPMENTS IN NETWORK DATA ANALYSIS Chair: Krishnakumar Balasubramanian

E0205: **Global spectral clustering of dynamic networks**
Presenter: **David Choi**, Carnegie Mellon, United States
A new method (PisCES) is presented for finding time-varying community structure in dynamic networks. The method implements degree-corrected spectral clustering, with a smoothing term to promote similarity across time periods. We prove that this method converges to the global solution of a nonconvex optimization problem, which can be interpreted as the spectral relaxation of a smoothed K-means clustering objective. We also show that smoothing is applied in a time-varying and data-dependent manner; for example, when a drastic change point exists in the data, smoothing is automatically suppressed at the time of the change point. Finally, we show that the detected time-varying communities can be visualized through the use of sankey plots.

E0498: **Theoretic and computational guarantee for mean-field variational Bayes methods on community detection**
Presenter: **Anderson Ye Zhang**, University of Chicago, United States
Mean field variational inference is widely used in statistics and machine learning to approximate posterior distributions. Despite its popularity, there exist remarkably little fundamental theoretical justifications. The success of variational inference mainly lies in its iterative algorithm, which, to the best of our knowledge, has never been investigated for any high-dimensional or complex model. We will describe the statistics/computation interface of the iterative algorithm of mean field variational inference. We will study it from a frequentist perspective, quantifying it by posterior contraction. For community detection problem, we will show that the iterative algorithm has a linear convergence to the optimal statistical accuracy within log n iterations. The technique can be extended to analyzing Expectation- maximization and Gibbs sampler with similar guarantees obtained, which will be briefly described. The considered community detection problem provides a test case and playground, and it is promising to understand mean field under a general class of latent variable models.

E0500: **Transform-based unsupervised point registration and unseeded low-rank graph matching**
Presenter: **Yuan Zhang**, Ohio State University, United States
Unsupervised estimation of the correspondence between two point sets has long been an attractive topic to CS and EE researchers. We focus on the vanilla form of the problem: matching two point sets that are identical over a linear transformation. We propose a novel method using Laplace transformation to directly match the underlying distributions of the two point sets. Our method provably achieves a decent error rate within polynomial time and does not require continuity conditions many previous methods rely on critically. Our method enables network comparison without strong model assumptions when node correspondence is unknown.

E0954: **Optimal rates for community estimation in the weighted stochastic block model**
Presenter: **Min Xu**, University of Pennsylvania, United States
Co-authors: Po-Ling Loh, Varun Jog
Community identification in a network is an important problem in fields such as social science, neuroscience, and genetics. Over the past decade, stochastic block models (SBMs) have emerged as a popular statistical framework for this problem. However, SBMs have an important limitation in that they are suited only for networks with unweighted edges; in various scientific applications, disregarding the edge weights may result in a loss of valuable information. We study a weighted generalization of the SBM, in which observations are collected in the form of a weighted adjacency matrix and the weight of each edge is generated independently of an unknown probability density determined by the community membership of its endpoints. We characterize the optimal rate of misclustering error of the weighted SBM in terms of the Renyi divergence of order 1/2 between the weight distributions of within-community and between-community edges, substantially generalizing existing results for unweighted SBMs. Furthermore, we present a principled, computationally tractable algorithm based on discretization that achieves the optimal error rate without assuming knowledge of the weight densities.

E1170: **On estimation and inference in latent structure random graphs**
Presenter: **Minh Tang**, Johns Hopkins University, United States
Co-authors: Avanti Athreya, Youngser Park, Carey Priebe
A latent structure model (LSM) random graph is defined as a random dot product graph (RDPG) in which the latent position distribution incorporates both probabilistic and geometric constraints, delineated by a family of underlying distributions on some fixed Euclidean space, and a structural support submanifold from which the latent positions for the graph are drawn. For a one-dimensional latent structure model with known structural support, we show how spectral estimates of the latent positions of an RDPG can be used for efficient estimation of the parameters of the LSM. We describe how to estimate or learn the structural support in cases where it is unknown, with an illustrative focus on graphs with latent positions along the Hardy-Weinberg curve. Finally, we use the latent structure model formulation to test bilateral homology in the Drosophila connectome.

EO136 Room F1 DIMENSION REDUCTION AND HIGH-DIMENSIONAL SUPERVISED LEARNING Chair: Andreas Artemiou

E0327: **A nonnegative robust linear model for deconvolution of proportions with biological applications**
Presenter: **Hyonho Chun**, Boston University, United States
Estimating mixing rates of a sample mixture is a popular problem in biomedical studies. Recently, it is applied to find immune cell infiltration in tumor samples. The main methodological challenge is tackling the non-Gaussian nature of gene expression data. Although a probabilistic model via multinomial or Poisson distributions would be a solution, such a model often becomes unidentifiable. An alternative is using robust regression because non-Gaussianity is often manifested as too high or too small expression values. We propose a non-negative robust linear model (NRLM) approach that yields robust yet interpretable mixing rate estimates. In the simulation study, NRLM shows a robust performance for finding the relative abundance of specified components when a large amount of noise is present. More importantly, the proposed approach accurately estimates the absolute level of the specified components in the presence of un-specified ones. Finally, it shows a superior performance when applied to deep deconvolution of blood samples.

E0313: **Counting process based dimension reduction methods for censored outcomes**
Presenter: **Ruoqing Zhu**, University of Illinois at Urbana-Champaign, United States
Co-authors: Donglin Zeng, Qiang Sun, Tao Wang
A class of dimension reduction methods for right censored survival data is proposed by using a counting process representation of the failure process. Semiparametric estimating equations are constructed to estimate the dimension reduction subspace for the failure time model. The proposed method addresses two fundamental limitations of existing approaches. First, using the counting process formulation, it does not require

any estimation of the censoring distribution to compensate the bias in estimating the dimension reduction subspace. Second, the nonparametric part in the estimating equations is adaptive to the structural dimension, hence the approach circumvents the curse of dimensionality. Asymptotic normality is established for the obtained estimators. We further propose a computationally efficient approach that simplifies the estimation equation formulations and requires only a singular value decomposition to estimate the dimension reduction subspace. Numerical studies suggest that our new approaches exhibit significantly improved performance for estimating the true dimension reduction subspace. We further conduct a real data analysis on a skin cutaneous melanoma dataset from The Cancer Genome Atlas. The proposed method is implemented in the R package "orthoDr".

E0346: **The central mean envelope for dimension reduction**
Presenter: **Chung Eun Lee**, University of Tennessee, Knoxville, United States
Co-authors: Xin Zhang, Xiaofeng Shao

A new envelope model, called central mean envelope, is introduced which generalizes a previous envelope model in two aspects. One aspect is that the central mean envelope does not impose a linear mean structure which can be viewed as a model-free method. Furthermore, the central mean envelope can have a heteroscedastic error to reduce the dimension. In particular, we seek a minimum subspace that reduces the conditional variance matrix of Y given X and fully captures the conditional mean dependence between Y and X. To estimate the central mean envelope, we use the martingale difference divergence matrix (MDDM) which measures the conditional mean dependence. Moreover, if there exists the heteroscedasticity, we use the slicing method or k-means method to find the subspace that reduces the conditional variance matrix. Theory is also provided regarding the consistency of the projection matrix associated with the central mean envelope. Favorable finite sample performance is demonstrated via simulations in comparison with some existing methods.

E0194: **Covariate-adjusted tensor classification in high-dimensions**
Presenter: **Yuqing Pan**, Florida State University, United States
Co-authors: Qing Mai, Xin Zhang

In contemporary scientific research, it is of great interest to predict a categorical response based on a high-dimensional tensor (i.e. multi-dimensional array) and additional covariates. This mixture of different types of data leads to challenges in statistical analysis. Motivated by applications in science and engineering, we propose a comprehensive and interpretable discriminant analysis model, called CATCH model (in short for Covariate-Adjusted Tensor Classification in High-dimensions), which efficiently integrates the covariates and the tensor to predict the categorical outcome. The CATCH model jointly models the relationships among the covariates, the tensor predictor, and the categorical response. More importantly, it preserves and utilizes the structures of the data for maximum interpretability and optimal prediction. To tackle the new computational and statistical challenges arising from the intimidating tensor dimensions, we propose a penalized approach to select a subset of tensor predictor entries that has direct discriminative effect after adjusting for covariates. We further develop an efficient algorithm that takes advantage of the tensor structure. Theoretical results confirm that our method achieves variable selection consistency and optimal classification error, even when the tensor dimension is much larger than the sample size. The superior performance of our method over existing methods is demonstrated in extensive simulated and real data examples.

E1735: **A general framework for sparse sufficient dimension reduction**
Presenter: **Wei Luo**, Zhejiang University, China

Sparse sufficient dimension reduction incorporates the sparsity assumption in variable selection into sufficient dimension reduction (SDR), and improves the latter in both the interpretability and the estimation accuracy. We construct a general framework to modify the ordinary SDR methods into sparse SDR methods, based on the recent serial work on uniform semiparametric SDR methods. The motivation comes from the observation that the minimum average variance estimator (MAVE) and the semiparametrically efficient estimator for the central mean subspace are asymptotically equivalent when the data are homoscedastic for regression, the justification of which makes independent contribution to the SDR literature. We show that, under certain regularity conditions, the sparse SDR methods based on the proposed framework have the variable selection consistency and the asymptotic normality, and enjoy a weak oracle property. In high-dimensional cases, we justify the asymptotic consistency of the proposed sparse sliced inverse regression. From simulation studies, the proposed sparse SDR methods have superior performance than the comparable existing sparse SDR methods.

EO332 Room G1 SURVIVAL ANALYSIS AND COPULA Chair: Takeshi Emura

E0461: **Nonparametric estimation of the joint distribution of two gap times under various types of censoring and truncation**
Presenter: **Carla Moreira**, University of Vigo - Faculty of Economics, Spain
Co-authors: Jacobo de Una-Alvarez, Ana Cristina Santos

The statistical analysis of consecutive gap times is an issue of much importance in a number of fields, including engineering, economy, epidemiology, and survival analysis. Particular difficulties appear, for example, when information on a cohort is obtained through intermittent visits or successive cross-sections; in such cases, special combinations of left-truncated, right-censored and interval censored data will appear. We describe one of such complicated settings in a three-state progressive model, and we introduce an inverse-probability-weighted type estimator for the joint distribution of two gap times which takes the aforementioned censoring and truncation issues into account. The performance of the proposed estimator is investigated through simulations, considering that the possible dependence structure between the two gap times is ruled by a parametric copula. For illustration purposes, the estimator is applied to data from the EPIPorto adults cohort study.

E1060: **Parametric and semiparametric modeling of doubly truncated lifetimes under time-restricted data collection schemes**
Presenter: **Achim Doerre**, University of Rostock, Germany, Germany

Double truncation occurs when individuals are observed if and only if their variable of interest is within two, possibly random, truncation bounds. In survival analysis, double truncation has been an ongoing topic due to multiple applications in areas including medicine, engineering and economics. The quasi-independence of the truncation variables and the variable of interest, an essential and critical assumption of many methods for truncation, is studied within a multidimensional Poisson process modeling framework. We primarily focus on a type of double truncation which is induced by time-restricted data collection, for which we argue that conventional nonparametric methods often have limited practical capacity. Additional considerations concerning asymptotics and the available parametric and semi-parametric modeling alternatives further motivate the need for a suitable goodness-of-fit assessment. We investigate the latter via Bayesian methods. Finally, we apply the proposed methods to a dataset of German companies whose age at insolvency is recorded conditional on their insolvency event.

E0225: **A time-varying joint frailty-copula model for analyzing recurrent events and a terminal event**
Presenter: **Ming Wang**, Pennsylvania State University, United States

Recurrent events could be stopped by a terminal event, which commonly occurs in biomedical and clinical studies. In this situation, the non-informative censoring assumption could be failed because of potential dependency between these two event processes, leading to invalid inference if analyzing recurrent events alone. The joint frailty model is widely used to jointly model these two processes. Recurrent events and terminal event processes are conditionally independent given the subject-level frailty. Furthermore, the correlation between the terminal event and the recurrent events is constant over time. We are motivated by the Cardiovascular Health Study (CHS). Both the correlation and the latent health status might

change during the follow-up period. We propose a time-varying joint frailty-copula model to relax these two assumptions under the Bayesian framework. The simulation studies show that the performance of our proposal, compared with the joint frailty model, has smaller absolute bias and mean squared error. Finally, we apply our method to analyze the CHS data potential to identify risk factors for myocardial infarction and stroke. We also quantify the correlation between these two types of events and all-cause death.

E1083: **Using nested Archimedean copulas to investigate the correlation structure in udder infection times**
Presenter: **Roel Braekers**, Hasselt University, Belgium
The imposed correlation structure on clustered multivariate time to event data, is in most cases taken as of a simple nature. In the shared frailty model, for example, all pairwise correlations between event times in a cluster are taken the same. When modelling the infection times for the four udder quarters clustered within a cow, more complex correlation structures are needed that will also give more insight into the infection process. We choose a copula approach to study more complex correlation structures in clustered infection times. We are able to model the marginal distributions separately from the association parameters, leaving them unaffected by the imposed association structure between the clustered event times. We use both Archimedean and nested Archimedean copula functions to model the associations. After introducing the different copula models, we compare them using likelihood ratio tests and explore the association structures by conditional probabilities. Afterwards we use simulations to validate the size and power of the different likelihood ratio tests used to discriminate between the copula models. Furthermore, we simulate from different copula families to look at the robustness of the association estimates when the association structure is misspecified.

E0913: **Joint correlation rank screening for semi-competing risks data with ultrahigh-dimensional gene features**
Presenter: **Liming Xiang**, Nanyang Technological University, Singapore
Co-authors: Mengjiao Peng
Ultrahigh dimensional gene features are often collected in modern cancer studies, where the number of gene features is extremely larger than the sample size. We propose a joint screening procedure for survival data subject to semicompeting risks with ultrahigh dimensional covariates. The method employs the ranking of the correlation between covariates and the joint survival distribution of both nonterminal and terminal event times. It is model-free and easy to implement as it only requires Kaplan-Meier estimators for the joint survival function of both event times. Theoretical properties of the proposed method are established. Simulation results show that the proposed screening procedure works very well for different types of marginal models. The practical utility is illustrated through the analysis of data from a breast cancer study.

EO242 Room H1 NEW METHODOLOGIES AND ADVANCES IN SURVIVAL AND RELIABILITY Chair: Juan Eloy Ruiz-Castro

E0396: **A proportional hazards model under bivariate censoring and truncation**
Presenter: **Marialuisa Restaino**, University of Salerno, Italy
Co-authors: Hongsheng Dai
Bivariate survival data have received considerable attention recently. However, most existing research works have focused on bivariate survival analysis when one component is censored or truncated and the other is fully observed. Only recently bivariate survival function estimation when both components are censored and truncated has received considerable attention. In order to evaluate the incidence of covariates on the duration time, the proportional hazards model is used. The focus is on the estimation of the regression coefficients in the Cox Proportional Hazards model, when the components are both censored and truncated. Moreover, we take into account that truncation could affect directly the hazard function. A simulation study and an application on read data are conducted to investigate the performance of the estimators of the unknown parameters.

E0445: **Towards reliable spatial prediction**
Presenter: **Gabrielle Kelly**, University College Dublin, Ireland
Co-authors: Raquel Menezes
Estimation of the variogram and associated parameters in spatial analysis is important for assessing spatial dependence and in predicting values of the measured variable at unsampled locations i.e. kriging. A simulation study is implemented to compare the performance of (i) Gaussian restricted maximum likelihood (reml) estimation, (ii) curve-fitting by ordinary least squares and (iii) nonparametric Shapiro-Botha estimation for estimating the covariance structure of a stationary Gaussian spatial process and a spatial process with t-distributed margins. Processes with Matern covariance functions are considered and the parameters estimated are the nugget, partial sill and practical range. Both parametric and nonparametric bootstrap distributions of the estimators are computed and compared to the true marginal distributions of the estimators. Gaussian reml is the estimator of choice for both Gaussian and t-distributed data and all choices of Matern variogram. However, accurate estimation of the Matern shape parameter is critical to achieving a good fit while this does not affect the Shapiro-Botha estimator. The parametric and nonparametric bootstrap both performed well, the latter being better for the Shapiro-Botha estimates. A numerical example, obtained from environmental monitoring, is included to illustrate the application of the methods and the bootstrap.

E0871: **Joint modelling of bivariate longitudinal data: Application to the recovery of sexual function and urinary continence**
Presenter: **Federica Nicolussi**, University of Milan, Italy
Co-authors: Marcella Mazzoleni, Simone Frassoni, Massimo Monturano, Rino Bellocco, Vincenzo Bagnardi
The following methodological issues occur in the context of the longitudinal study of sexual dysfunction and urinary incontinence after radical prostatectomy: (i) high dropout rate due to the extremely sensitive nature of the investigated outcomes; (ii) correlation between the two outcomes; (iii) non-linearity of the recovery trajectories. To address all these issues, we propose the use of a joint modelling approach, including a bivariate linear mixed model with splines for the two outcomes and a proportional hazards model for the time to dropout. We applied the model to data from consecutive patients underwent robotic-assisted radical prostatectomy at European institute of oncology from May 2015 to July 2016. Pre- and post-surgical sexual and urinary functional conditions were evaluated using the expanded prostate cancer index composite questionnaire. Six hundred forty three patients were included in the analysis. At one year after surgery, only 55% of patients returned the questionnaire. Parameters estimation was based on the maximisation of the likelihood function achieved through the implementation of an EM algorithm. A Gauss Hermite approximation was also used for some of the integrals involved. To assess the effect of nonrandom dropout mechanisms on the parameter estimates, we calculated the index of local sensitivity to non-ignorability.

E0974: **Using Coxian phase-type distributions and survival trees to model length of stay of elderly patients in Italy hospitals**
Presenter: **Adele Marshall**, Queens University Belfast, United Kingdom
Co-authors: Mariangela Zenga
Coxian phase-type distributions are a special type of Markov chain which describe the time which elapses until a certain event occurs as a series of sequential phases. The Coxian phase-type distribution will be used for modelling the length of stay in hospital of elderly patients. The data consists of records for four years (2012-2015 inclusive) for both private and public hospitals in all 21 regions of Italy for all patients aged 65 and over. The optimum number of phases for modelling the elderly patient data was determined to be three. It was found that the optimal three phase Coxian distribution was a better fit to the data than normal, Weibull and log-normal distributions. Survival methods have previously been used to investigate the length of time elderly patients in Italy spend in hospital and had shown that depending on the hospital, the patient length of stay may

differ. This is further investigated by producing survival trees for the current, more recent data, where the covariates are analysed to evaluate which variables significantly influence patient survival.

E0621: **A multi-state k-out-of-n: G system subject to multiple events and loss of units**
Presenter: **Juan Eloy Ruiz-Castro**, University of Granada, Spain
Co-authors: Mohammed Dawabsha
A k-out-of-n: G system is an n-system that works if at least k units are operational. We consider a multi-state k-out-of-n: G system that evolves in discrete time subject to multiple events. Each unit can undergo repairable and non-repairable failures. When a non-repairable failure occurs the unit is removed and if the failure is repairable then the unit goes to the repair facility for corrective repair. When the number of units in the system is less than k then the system is replaced by new and identical one. The system is modelled and several interesting measures are worked out in a well-structured form in transient and stationary regime. Matrix analytic methods are used. The results have been implemented computationally with Matlab.

EO172 Room I1 COMPUTATIONAL STATISTICS IN DISTRIBUTION THEORY Chair: Andriette Bekker

E0303: **The Mobius distribution with spiral motion**
Presenter: **Mohammad Arashi**, Shahrood University of Technology, Iran
Using the bivariate spherically symmetric beta, a family of 3-parameter distributions, namely Mobius distribution proposed long back, which has support the unit disk in two dimensions. Hence, a model for controlling orientation, the degree of concentration, and off-centeredness. The model applied to a bivariate dataset consisting ozone concentration and wind direction. However, for some purposes say, the ozone concentration divided by 120. We elaborate the use of spiral motion for Mobius distribution. It is numerically shown that by carefully selecting the parameter of the spiral, the existing model in the literature can be improved, in the sense of having smaller AIC for model selection purposes.

E0507: **The likelihood ratio test of independence for random size samples**
Presenter: **Carlos Coelho**, NOVA University of Lisbon, Portugal
Co-authors: Filipe Marques
The focus is on the distribution of the likelihood ratio test statistic used to test the independence of two sets of variables when samples have random sizes, originated from different underlying distributions, namely Poisson, Binomial and Negative Binomial. Situations are identified when the distribution of the statistic has closed finite form representations, while for the other situations sharp approximations are developed, being shown how sharp upper bounds on the approximation error may be obtained. Using a real data set, the test is performed assuming, for the derivation of the distribution of the likelihood ratio statistic, a random sample size and the results are compared with the ones obtained using the traditional approach.

E0527: **A new matrix variate gamma distribution with applications**
Presenter: **Anis Iranmanesh**, Mashhad Branch, Islamic Azad University, Iran
A generalized matrix variate gamma distribution, which includes a trace function in the kernel of the density, is introduced. Some important statistical properties including Laplace transform, distributions of some functions are derived. A new matrix t-type family of distributions is also generated by the proposed matrix variate gamma distribution. Some other applications are also addressed and studied in the bivariate form.

E0661: **Generalized Rayleigh-exponential-Weibull distribution and progressively type-I interval-censored data**
Presenter: **Din Chen**, University of North Carolina, United States
Co-authors: Yuhlong Lio, Tanita Cronje
A new family of 3-parameter Generalized Rayleigh-Exponential-Weibull(GREW) distributions is proposed to unify the family of 2-parameter Generalized Rayleigh (GR) distribution, the family of 2-parameter Generalized Exponential (GE) distribution and the family of 2-parameter Weibull (W) distribution. This GREW distribution includes a variety of commonly used statistical distributions, such as, Rayleigh distribution, GR distribution, exponential distribution, GE distribution, Weibull distribution, etc., and therefore can provide better model fitting in real applications. We illustrate the application of this new GREW distribution to fit a typical progressive type-1 interval-censored data from 112 patients with plasma cell myeloma treated at the National Cancer Institute and show the superior model fitting using the value of fitted log-likelihood function and the Kolmogorov-Smironov goodness-of-fit statistic. A series of Monte-Carlo simulation studies are further designed to evaluate the performance of parameter estimation for this new distribution family.

E1062: **A wrapped skew generalised normal family**
Presenter: **Theodor Loots**, University of Pretoria, South Africa
Co-authors: Andriette Bekker
A scale mixture of the uniform and generalised gamma distribution results in a very flexible parametric family of distributions. This family includes a skew-generalised normal distribution as a special case. The resultant family is wrapped in the usual fashion, yielding a new wrapped parametric family of distributions. Simulations are carried out using the corresponding stochastic representation, while parameter estimation is performed using the method of trigonometric moments, as well as that of maximum likelihood. Finally, this new family is fitted to natural phenomena such as wind directional data.

EO404 Room L1 SOFT CLUSTERING Chair: Maria Brigida Ferraro

E0532: **Computational efficiency for fuzzy clustering**
Presenter: **Yuichi Mori**, Okayama University of Science, Japan
Co-authors: Takatsugu Yoshioka, Masahiro Kuroda
One of the most widely used soft clustering algorithms is the Fuzzy c-means clustering (FCM), which searches a reasonable form of clustering in which each data point belongs to multiple clusters. FCM therefore requires high computational cost due to the iterative computation to estimate two parameters (membership and cluster centroid matrices) alternately. Because this computational algorithm is a kind of alternating least squares (ALS) algorithm, so a general procedure to accelerate ALS type of iteration using the vector epsilon algorithm can be applied to FCM computation to obtain the computational results faster than the original computation. The performance/efficiency of the vector epsilon accelerated FCM algorithm is evaluated in simulations under several conditions e.g., data size, the number of original clusters, the number of estimated clusters, accelerated parameter (membership or centroid) and so on. These numerical experiments demonstrate that the vector epsilon accelerated FCM accelerates the computation twice or more as fast as the original one.

E0653: **The fclust R package for fuzzy clustering: A new version**
Presenter: **Alessio Serafini**, Sapienza Universita di Roma, Italy
Co-authors: Maria Brigida Ferraro, Paolo Giordani

Fuzzy clustering is a technique extensively used in several domains of research to discover fuzzy partitions of data sets. The fclust package is a useful toolbox for fuzzy clustering in R programming language. Differently from the other fuzzy clustering packages available on CRAN, fclust implements not only the most known fuzzy clustering algorithm, the fuzzy k-means (fkm), but also many extensions of fkm, cluster validity indices and visualization tools. The new version allows us to use fuzzy relational clustering algorithms for partitioning mixed-type data, a new version of Gustafson-Kessel algorithm to avoid singularity in the covariance matrix, fuzzy versions of some cluster comparison methods, and few minor changes as automatically selection of number of clusters.

E1172: **An infinite mixture model for clustering of multiplex data**
Presenter: **Silvia DAngelo**, University of Rome La Sapienza, Italy
Co-authors: Michael Fop, Marco Alfo

Social network analysis is a well-known and growing branch of statistics. Network structures, either single or multivariate, may arise in various contexts and have been investigated in a broad variety of fields. A popular approach to model this type of data is by means of latent variables, which are assumed to influence the observed structure. In particular, latent space models allow us to describe the observed structure by means of an unobserved latent space; units close in the latent space are assumed to be more likely to connect. In many network data, units have the tendency to cluster into communities and this feature has been largely investigated in the context of single networks. A clustering framework for multivariate network data is proposed based on infinite mixtures of Gaussian distributions. We make use of a single latent space and estimate the model parameters within a hierarchical Bayesian framework. A real data application will be presented.

E1484: **Semi parametric mixtures of generalised linear models**
Presenter: **Sollie Millard**, University of Pretoria, South Africa
Co-authors: Frans Kanfer, Mohammad Arashi

Mixtures of generalised linear models and an extension to a semi-parametric mixture setting are considered. The link function is replaced by a non-parametric estimate thereof. This approach allows for more flexibility since the non-parametric link function gives access to a larger subset of distributions in the exponential family, whilst retaining much of the structure of a generalised linear model. The performance of the proposed procedure is evaluated through a simulation study considering various input settings. An industry case study using a semi-parametric alternative to mixtures of logistic regressions is also presented.

E1489: **Applications of scale mixtures in the exponential family**
Presenter: **Frans Kanfer**, University of Pretoria, South Africa
Co-authors: Sollie Millard, Mohammad Arashi

Robust mixtures of scale mixtures of distributions from the exponential family are considered, specifically, mixtures of scale mixtures of Gaussian distributions. Estimating parameters using the EM algorithm is discussed. Selecting the mixing distribution a gamma distribution results in a mixture of t distributions. Regression based on mixtures of scale mixtures of Gaussian distributions is considered. An industry case study, using mixture regression with t distribution errors is discussed.

EO292 Room M1 NEW DEVELOPMENTS ON ROBUSTNESS AND FUNCTIONAL DATA ANALYSIS Chair: Juan Romo

E0649: **Robustness on big functional data depths**
Presenter: **Alicia Nieto-Reyes**, Universidad de Cantabria, Spain
Co-authors: John Aston

Functional depth ranks the data in a functional data set. We compare through simulations the robustness of several computationally efficient functional depths that can be applied in the big data setting. In turns, we see that the notion of depth that satisfies the properties of statistical functional depth results in a better performance under the presence of outliers.

E0767: **Notions of robustness and stability in machine learning**
Presenter: **Andreas Christmann**, University of Bayreuth, Germany

There exist many different notions of robustness and stability in robust statistics and in machine learning. The goal is to compare these notions and discuss advantages and disadvantages.

E1498: **A clustering procedure for multivariate functional data based on a Mahalanobis type distance**
Presenter: **Andrea Martino**, Politecnico di Milano, Italy
Co-authors: Andrea Ghiglietti, Francesca Ieva, Anna Maria Paganoni

Clustering functional data can be a difficult task, because of the dimensionality of the space the data belongs to. To address the difficulties arising from the study of functional data, several approaches have been proposed along the years. A standard procedure consists in reducing the infinite dimensional problem to a finite one, approximating the data with elements from a finite dimensional space. Since this approach may lead to losing some important information about the data, we propose a novel clustering technique for samples of multivariate functions. The method consists in a k-means algorithm in which the distance between the curves is measured using a metric that generalizes the Mahalanobis distance in Hilbert spaces. This procedure is able to consider the correlation and the variability along all the components of the functional data. The proposed algorithm is tested in a simulation study and compared with the k-means based on other distances commonly used for clustering multivariate functional data. Finally, the method is applied to two case studies, concerning growth curves and ECG signals.

E1525: **Selection of unsupervised classification methods for functional data**
Presenter: **Lucas Fernandez Piana**, Universidad de Buenos Aires and CONICET, Argentina
Co-authors: Sergi Gonzalez, Ana Justel, Julio Rodriguez-Puerta, Marcela Svarc

In cluster analysis there is neither an accepted common criterion to evaluate the performance of different procedures, nor a lower bound that indicates the difficulty of the problem. In this context, it is important to develop criteria to compare different clustering procedures on the same data set. The aim is to propose criteria for comparing different partitions in the context of functional data. Inspired by the classification problem, where the methods are easily comparable by the misclassification rate, we build a confusion matrix based on local and global depths. The method is applied to clusters of trajectories and back-trajectories arriving to the Byers Peninsula, located at the western coast of the Livingston Island (South Shetland Islands, Antarctica). Ten years of 5-day back trajectories each six hours, computed with the Hybrid Single-Particle Lagrangian Integrated Trajectory model (HYSPLIT) are analyzed.

E1715: **High dimensional tensor regression for neuroimaging data**
Presenter: **Montserrat Fuentes**, Virginia Commonwealth University, United States

Imaging data with thousands of spatially-correlated data points are common in many fields. In Neurosciences, magnetic resonance imaging (MRI) is essential for studying brain structure and activity. Modeling spatial dependence of MRI data at different scales is key. It could allow for accurate testing for significance in neural activity. The high dimensionality presents modeling and computational challenges. Methods that account for

spatial correlation often require cumbersome matrix evaluations which are prohibitive for large data. Thus, current methods typically reduce dimensionality by modeling covariance among regions of interest coarser or larger spatial units rather than among voxels. However, ignoring spatial dependence at different scales could drastically reduce our ability to detect activation patterns in the brain, and hence produce misleading results. We introduce a novel Bayesian Tensor approach, treating the brain image as response and having a vector of predictors. Parameter estimates using a generalized sparsity principle are provided. A fully Bayesian approach to characterize different sources of uncertainty is employed. We demonstrate posterior consistency and develop a computationally efficient algorithm. The effectiveness is illustrated through simulations and the analysis of the effects of cocaine addiction on the brain structure. The aim is to identify the effects of demographic information and cocaine addiction on the functioning of the brain.

EO430 Room N1 MODELLING EXTREMES WITH COVARIATES Chair: Juan Juan Cai

E0342: **Improving precipitation forecast using extreme quantile regression**
Presenter: **Jasper Velthoen**, Delft University of Technology, Netherlands
Co-authors: Juan Juan Cai, Geurt Jongbloed, Maurice Schmeits

Aiming to predict extreme precipitation forecast quantiles, a nonparametric regression model that features a constant extreme value index is proposed. Using local linear quantile regression and an extrapolation technique from extreme value theory, we develop an estimator for conditional quantiles corresponding to extreme high probability levels. We establish uniform consistency and asymptotic normality of the estimators. In a simulation study, we examine the performance of our estimator on finite samples in comparison with existing methods. On a precipitation data set in the Netherlands, our estimators have more predictive power compared to the upper member of ensemble forecasts provided by a numerical weather predication model.

E0646: **Regression type models for extremal dependence**
Presenter: **Linda Mhalla**, HEC Montreal, Canada
Co-authors: Miguel de Carvalho, Valerie Chavez-Demoulin

A vector generalized additive modelling framework is discussed for taking into account the effect of covariates on angular density functions in a multivariate extreme value context. The proposed methods are tailored for settings where the dependence between extreme values may change with time and/or other covariates. We will devise a penalized maximum log-likelihood estimator, discuss details of the estimation procedure, and its consistency and asymptotic normality. The empirical analysis reveals relevant dynamics of the dependence between extreme air temperatures in two alpine resorts during the winter season.

E1309: **Efficient adaptive covariate modelling for extremes**
Presenter: **Philip Jonathan**, Lancaster University / Shell Research Limited, United Kingdom
Co-authors: David Randell, Elena Zanini, Emma Ross, Matthew Jones

The characteristics of extreme values are usually dependent on covariates. For example, the severity of the ocean environment in a storm typically depends on a number of covariates, including the direction of storm propagation and the season of storm occurrence. Reliable practical application of extreme value methods therefore needs to accommodate the effects of covariates in a statistically- and computationally- efficient manner, and to quantify uncertainties in inferences carefully. Penalised spline representations for the variation of extreme value parameters with multi-dimensional covariates have been demonstrated to be useful and flexible; inference however becomes computationally unwieldy as the dimensionality of the covariate space and the complexity of covariate effects increase. For this reason, we explore alternative representations for covariate effects in extreme value models which yield more efficient, scalable inference without loss of (too much) flexibility. Specifically, for 2-dimensional covariates, we consider (a) Bayesian adaptive regression spline (BARS) parameterisations, and (b) Bayesian piecewise constant representations (motivated by previous work and Voronoi partitions of the covariate domain) of covariate effects in non-stationary extreme value models, and compare the performance of these models with those based on (c) Bayesian penalised B-spline representations.

E1414: **Bias-corrected estimation for conditional Pareto-type distributions with random right censoring**
Presenter: **Yuri Goegebeur**, University of Southern Denmark, Denmark
Co-authors: Armelle Guillou , Jing Qin

Bias-reduced estimation of the extreme value index is considered in conditional Pareto-type models with random covariates when the response variable is subject to random right censoring. The bias-correction is obtained by fitting the extended Pareto distribution locally to the relative excesses over a high threshold using the maximum likelihood method. Consistency and asymptotic normality of the estimators are established under suitable assumptions. The finite sample behaviour is illustrated with a simulation experiment and the method is applied to some real data sets.

E0935: **Latent group structures with heterogeneous distributions: Identification and estimation**
Presenter: **Wendun Wang**, Erasmus University Rotterdam, Netherlands
Co-authors: Xuan Leng, Heng Chen

Panel data are often characterized by cross-sectional heterogeneity, and a flexible yet parsimonious way of modeling heterogeneity is to cluster units into groups. A group pattern of heterogeneity may exist not only in the mean but also in the other characteristics of the distribution. To identify latent groups and recover the heterogeneous distribution, we propose a clustering method based on composite quantile regressions. We show that combining the strength across multiple panel quantile regression models improves the precision of the group membership estimates if the group structure is common across quantiles. Asymptotic theories for the proposed estimators are established, while their finite-sample performance is demonstrated by simulations. We finally apply the proposed methods to analyze the cross-country output effect of infrastructure capital.

EO128 Room O1 CHOOSING BANDWIDTHS AND TUNING PARAMETERS Chair: Lola Martinez-Miranda

E0511: **Data-based selection of the tuning parameter appearing in certain families of goodness-of-fit tests**
Presenter: **Carlos Tenreiro**, University of Coimbra, Portugal

The situation, common in the current literature, is that of a whole family of location-scale/scale invariant test statistics indexed by a set Λ of real numbers, is available to test the goodness of fit of F, the underlying distribution function of the real-valued iid random variables, to a location-scale/scale family of distribution functions. The power properties of the tests associated with the different statistics usually depend on $\lambda \in \Lambda$, called the "tuning parameter", which is the reason that its choice is crucial to obtain a performing test procedure. We address the data-dependent choice of λ in the set Λ, assumed to be finite, as well as the calibration of the associated goodness-of-fit test procedure. Examples of existing and new tuning parameter selectors are discussed, and the methodology presented, of combining different test statistics in a single test procedure, is applied to well-known families of test statistics for normality and exponentiality.

E0907: **Taking advantage of the optimal kernel in nonparametric density estimation**
Presenter: **Maria Isabel Borrajo**, Lancaster University, Spain
Co-authors: Jose E Chacon, Alberto Rodriguez-Casal
Density estimation has been extensively studied, particularly in the context of nonparametric statistics. During the last decades many advances have been made: the introduction of the histogram, the kernel density estimator with all its variations and bandwidth selection methods, splines base methodology and so on. We propose a new density estimator based on a previous theory, which only requires the kernel to be a L_2-function. In this way, we let the kernel vary in shape and scale, and at the same time, we avoid the problem of bandwidth selection. The proposal is to use the empirical characteristic function by selecting a cut-off point to remove the extra noise that commonly appears in the tails; we define a new data-driven procedure to select this truncation point in the frequency domain and then apply Fourier transforms to obtain the target, i.e., the density estimation. The good performance of our proposal is illustrated in an extensive simulation study, in which we have also included the existing competitors.

E0934: **A non-model-based approach to bandwidth selection for kernel intensity estimators**
Presenter: **Marie-Colette Van Lieshout**, CWI/UT, Netherlands
Co-authors: Ottmar Cronie
A new bandwidth selection method is discussed for kernel estimators of spatial point process intensity functions. The method is based on an optimality criterion motivated by the Campbell formula applied to the reciprocal intensity function. The new method is fully nonparametric, does not require knowledge of higher-order moments, and is not restricted to a specific class of point process. Our approach is computationally straightforward and does not require numerical approximation of integrals.

E0485: **Bandwidth selection for prediction in regression**
Presenter: **Ines Barbeito**, University of A Coruna, Spain
Co-authors: Stefan Sperlich, Ricardo Cao
The smoothed bootstrap method has been used in the context of prediction, in which the response variable of the target population remains unknown. Specifically, this bootstrap procedure is used for the purpose of bandwidth selection in regression estimation. The aim is to establish a new bootstrap bandwidth selector based on the exact expression of the bootstrap version of the mean average squared error of some approximation of the kernel regression estimator. This is very useful since Monte Carlo approximation is avoided for the implementation of the bootstrap selector. Furthermore, the distribution of the target population no longer needs to be estimated.

E0909: **Multiplicative local linear hazard estimation and best one-sided cross-validation**
Presenter: **Lola Martinez-Miranda**, Universidad de Granada, Spain
Co-authors: Maria Luz Gamiz, Jens Perch Nielsen
Detailed mathematical statistical theory is developed for a new class of cross-validation techniques of local linear kernel hazards and their multiplicative bias corrections. The new class of cross-validation combines principles of local information and recent advances in indirect cross-validation. A few applications of cross-validating multiplicative kernel hazard estimation do exist in the literature. However, detailed mathematical statistical theory and small sample performance are introduced and further upgraded to our new class of best one-sided cross-validation. Best one-sided cross-validation turns out to have excellent performance in its practical illustrations, in its small sample performance and in its mathematical statistical theoretical performance.

EO368 Room Q1 GOING ROBUST: NEW DEVELOPMENTS AND APPLICATIONS Chair: Vanda Lourenco

E0772: **Detecting outliers on microarrays with mixtures of normal and heavy-tailed distributions**
Presenter: **Alexandra Posekany**, Danube University Krems, Austria
A common assumption in statistical modelling is a normal error distribution, however many data sets in fields like biology or economics do not follow these, independent of sample sizes. To robustify inference we employ a Bayesian hierarchical mixture model which simultaneously performs bio-informatical inference and detects outliers on a gene and array level. This is obtained by mixing normal mixture components with Student's t distributed ones to identify the over-dispersed part of data which may originate from biological processes or laboratory work. In our application, we present several microarray data, which generally show over-dispersed noise behaviour. In recent years, microarrays were less regarded in molecular biological research, but have been introduced to clinical practice which makes the detection of outliers indicating problems in the medical and bio-informatical analyses even more relevant. In addition to provide a better inference of differential expression, the goal is to identify noisy genes in a gene and array level. Thus, we wish to identify whether single arrays are responsible for this behaviour to provide a quality control for clinical practice.

E0970: **A generalized spatial sign covariance matrix**
Presenter: **Jakob Raymaekers**, KULeuven, Belgium
Co-authors: Peter Rousseeuw
The well-known spatial sign covariance matrix (SSCM) carries out a radial transform which moves all data points to a sphere, followed by computing the classical covariance matrix of the transformed data. We study more general radial functions. It is shown that the eigenvectors of the generalized SSCM are still consistent. The breakdown value of the resulting scatter matrix is derived and the influence function is calculated. A simulation study indicates that the best results are obtained when the inner half of the data points are not transformed and points lying far away are moved toward the center.

E1187: **Robust multivariate methods based on the weighted likelihood**
Presenter: **Luca Greco**, University of Sannio - Benevento, Italy
Co-authors: Claudio Agostinelli
Standard data reduction techniques, such as principal component analysis, discriminant analysis, cluster analysis, exhibit lack of robustness with respect to the occurrence of outliers, anomalous values that can completely break down classical procedures, hence leading to unreliable conclusions. This unpleasant behavior stems from the fact that they rely on the sample mean vector and sample covariance matrix. Then, robust data reduction methods can be defined by supplying robust estimates of multivariate location and scatter. Furthermore, formal rules for the purpose of outlier detection can be obtained. The interest focuses on those techniques driven by the employ of weighted likelihood multivariate estimates. Weighted likelihood estimation is characterized by the evaluation of unit specific data dependent weights lying in the interval $[0, 1]$, aiming at downweighting the effect of anomalous observations. The weights depend on the so called Pearson residuals, aimed at comparing the data, summarized by a non-parametric density estimate, and the model. In a multivariate setting Pearson residuals are obtained from the univariate distribution of Mahalanobis distances. Then, the effect of large residuals is bounded by a suitable residual adjustment function in the estimation process.

E1185: **Robustness in Bayesian inference**
Presenter: **Laura Ventura**, University of Padova, Italy
Co-authors: Erlis Ruli, Nicola Sartori
The aim is to review the properties and applications of the so-called robust posterior distributions, i.e. posterior distributions derived from the combination of a robust pseudo-likelihood function or an unbiased estimating function with suitable prior information. Examples of pseudo-likelihoods are the composite, the empirical and the quasi-likelihoods, while unbiased estimating functions include as special instances M-estimating functions and proper scoring rules. From a theoretical point of view we illustrate how to perform robust Bayesian inference from pseudo-likelihoods and estimating equations. From a practical point of view, we show the simple but effective application of robust posterior distributions in challenging examples.

E0173: **Robust clustering based on determinants-and-shape constraints**
Presenter: **Luis Angel Garcia-Escudero**, Universidad de Valladolid, Spain
Co-authors: Agustin Mayo-Iscar, Marco Riani, Andrea Cerioli
Model-based clustering mostly relies on the maximization of classification and mixture likelihoods. Trimming principles can be added to these maximum likelihood maximization in order to robustify them. Trimmed adaptions of traditional (classification) EM can be applied with this aim. However, non-interesting or "spurious" clusters, made of few almost collinear observations, can be easily detected when no proper constraints on the components scatter matrices are considered. Therefore, only trimming is not enough and appropriate constraints are required in order to achieve better robustness performance. Establishing an upper bound on the scatter matrices determinant ratio seems to be a sensible idea for constraining scatter matrices by applying an affine equivariant type of constraints. However, degeneracy issues are not fully solved. On the other hand, we will see how some extra mild constraints on the shape matrices elements can be useful in this framework. A computationally feasible algorithm will be presented and the proposed methodology is illustrated through simulations and real data examples.

EO260 Room O2 CSDA JOURNAL: TIME SERIES AND NONPARAMETRIC METHODS Chair: Stephen Pollock

E1071: **A frequency domain bootstrap for general stationary processes**
Presenter: **Jens-Peter Kreiss**, Technische Universitaet Braunschweig, Germany
Co-authors: Marco Meyer, Efstathios Paparoditis
Existing frequency domain methods for bootstrapping time series have a limited range. Essentially, existing frequency domain bootstrap procedures cover the case of linear time series with independent innovations, and some even require the time series to be Gaussian. We propose a new frequency domain bootstrap method which is consistent for a much wider range of stationary processes and can be applied to a large class of periodogram-based statistics. It introduces a new concept of convolved periodograms of smaller samples which uses pseudo periodograms of subsamples generated in a way that correctly imitates the weak dependence structure of the periodogram. We show consistency for this procedure for a general class of stationary time series, ranging clearly beyond linear processes, and for general spectral means and ratio statistics. Furthermore, we show how existing bootstrap methods can be corrected using this new approach. The finite sample performance of the new bootstrap procedure is illustrated via simulations.

E1103: **Change points detection and identification for high dimensional data**
Presenter: **Ping-Shou Zhong**, University of Illinois at Chicago, United States
High-dimensional time series are often collected when a large number of features are measured repeatedly over time. An important challenge of such data is the existence of both spatial and temporal dependence, as well as the high dimensionality. We will introduce our proposed nonparametric methods developed for detecting and identifying change points for high dimensional data. Our methods do not impose explicit restriction on the growth rate between data dimension and sample size, and allow very general spatial and temporal dependence. The proposed change points estimation procedure is shown to be consistent in estimating the locations of multiple change points. The rate of convergence depends on data dimension and signal-to-noise ratio. Computation and applications of the proposed methods to time-course microarray data and functional MRI data will be discussed.

E1112: **Density estimators that can be plugged in**
Presenter: **Eric Beutner**, Maastricht University, Netherlands
Co-authors: Henryk Zaehle
Density estimators that can be plugged in are considered. The notion plug-in property of density estimators has been introduced in 2003. It means that statistical functionals like the expected value or the second moment can be estimated by replacing the density by its estimate in the respective formula. We focus on kernel based methods and show that appropriate kernels have the plug-in property for large classes of functionals. It will be demonstrated that the results continue to hold for many stationary time series models.

E0764: **Bias correction for local linear regression estimation using asymmetric kernels via the skewing method**
Presenter: **Masayuki Hirukawa**, Ryukoku University, Japan
The aim is to extend the skewing method that has been originally proposed as a bias correction device for local linear regression estimation using standard symmetric kernels to the cases of asymmetric kernels. The method is defined as a convex combination of two or three local linear estimators. It is demonstrated that as with symmetric kernels, the skewed regression estimator using asymmetric kernels with properly chosen weights can accelerate the bias convergence under sufficient smoothness of the unknown regression curve while not inflating the variance in order of magnitude. Orders of magnitude in the bias and variance convergences are the same as those for a local cubic estimator. A remarkable difference between the cases of symmetric and asymmetric kernels can be found in the form of weights. While the weights are constant regardless of the position of the design point for symmetric kernels, they vary with the design point for asymmetric kernels. Finite-sample performance of the skewed regression estimator is examined via Monte Carlo simulations in comparison with a local cubic smoother, and an empirical application is considered.

E0654: **Generalised additive dependency inflated models including aggregated covariate**
Presenter: **Young Kyung Lee**, Kangwon National University, Korea, South
Co-authors: Enno Mammen, Jens Perch Nielsen, Byeong Park
Suppose that X, Y and U are observed and that the conditional mean of U given X and Y can be expressed via an additive dependency of X, $\lambda(X)Y$ and $X+Y$ for some unspecified function λ. This structured regression model can be transferred to a hazard model or a density model when applied on some appropriate grid, and has important forecasting applications via structured density models including age-period-cohort relationships. In case the conditional mean of U approximates a density, the regression model can be used to analyse the age-period-cohort model, also when exposure data are not available. In case the conditional mean of U approximates a marker dependent hazard, the regression model introduces new relevant age-period-cohort time scale interdependencies in understanding longevity. A direct use of the regression relationship is the estimation of the severity of outstanding liabilities in non-life insurance companies. The technical approach taken is to use B-splines to capture the underlying 1-dimensional unspecified functions. It is shown via finite sample simulation studies and an application for forecasting future asbestos related

deaths in the UK that the B-spline approach works well in practice. Special consideration has been given to ensure identifiability of all models considered.

EO036 Room P2 BAYESIAN SEMI- AND NONPARAMETRIC MODELLING I Chair: Li Ma

E0404: **Bayesian repulsive Gaussian mixture model with its applications to EHR**
Presenter: **Yanxun Xu**, Johns Hopkins University, United States
Co-authors: Fangzheng Xie

A general class of Bayesian repulsive Gaussian mixture models that encourage well-separated clusters is developed. It aims at reducing potentially redundant components produced by independent priors for locations (such as the Dirichlet process). The asymptotic results for the posterior distribution of the proposed models are derived, including posterior consistency and posterior contraction rate in the context of nonparametric density estimation. More importantly, we show that compared to the independent prior on the component centers, the repulsive prior introduces additional shrinkage effect on the tail probability of the posterior number of components, which serves as a measurement of the model complexity. In addition, an efficient and easy-to-implement blocked-collapsed Gibbs sampler is developed based on the exchangeable partition distribution and the corresponding urn model. We evaluate the performance and demonstrate the advantages of the proposed model through extensive simulation studies and real data analysis.

E0440: **A Bayesian nonparametric approach for causal inference with semi-competing risks**
Presenter: **Michael Daniels**, University of Florida, United States

A Bayesian nonparametric (BNP) model is developed in order to assess the treatment effect in semi-competing risks, where a nonterminal event may be censored by a terminal event, but not vice versa. Semi-competing risks are common in brain cancer trials with death being censored by cerebellar progression. We propose a flexible BNP approach to model the joint distribution of progression and death events, thereby effectively inferring the marginal distributions of progression time and death time, characterizing within-subject dependence structure, predicting the progression and death times given a patient's covariate, and quantifying uncertainties of all estimates. More importantly, we define a causal effect of treatment, which can be estimated from the data and has a nice causal interpretation. We perform extensive simulation studies to evaluate the proposed BNP model. The simulations show that the proposed model can accurately estimate the treatment effect in semi-competing risks setup. We also implement the proposed BNP model on data from a brain cancer Phase II trial.

E0517: **Dirichlet processes and copulas**
Presenter: **Clara Grazian**, University of Oxford, United Kingdom
Co-authors: Gianluca Mastrantonio, Enrico Bibbona

The Dirichlet process is a stochastic process defined on a space of distribution functions and which depends on a scaling parameter and a base distribution. It has an explicit representation called stick breaking representation, which is almost surely discrete, i.e. depends on some weighted atoms generated from the base distribution. The Dirichlet process may be extended in a hierarchical version, so that several processes share the same set of atoms with process dependent weights, which has a stick breaking representation. The original construction of the hierarchical Dirichlet process considers weights that are independent for different processes. We propose a way to introduce a dependence in the marginal distribution of the vectors of weights, by imposing a Gaussian copula whose correlation matrix has a given dependence structure (for instance, implying spatial dependence). We also prove that the dependence structure imposed on the (independent) components of the stick breaking representation is automatically transferred to the vectors of weights and that the order in which the components are taken does not matter. This representation of the hierarchical Dirichlet process may be used to produce a nonparametric clustering, where the weights of the Dirichlet process represents the probabilities to be allocated to each cluster and the dependence is among the weights relative to the same cluster.

E1207: **Augmented conditional sampler for nonparametric mixture models**
Presenter: **Bernardo Nipoti**, Trinity College Dublin, Ireland
Co-authors: Antonio Canale, Riccardo Corradin

Nonparametric mixture models based on the Pitman-Yor (PY) process are a flexible tool for density estimation and clustering. Two main classes of algorithms, namely marginal and conditionals, have been considered in literature. We propose a new algorithm, named Augmented Conditional Sampler (ACS), which, although technically conditional, is closely reminiscent of the Polya urn marginal scheme and features the same degree of interpretability. Unlike its most popular conditional competitors, the ACS does not rely on the stick-breaking representation of the underlying PY process and turns out to be more robust to the choice of the parameters characterising the distribution of the underlying PY process. The performance of the ACS is investigated and compared with popular competitors, by means of an extensive simulation study. Finally, the proposed sampler is used as the building block of a new algorithm for carrying out posterior inference based on a class of dependent nonparametric priors.

E0980: **On Bayesian uncertainty quantification in sparse high-dimensional models**
Presenter: **Botond Szabo**, Leiden University, Netherlands

The reliability of Bayesian uncertainty quantification in sparse high-dimensional models is investigated for various choices of the prior distribution, including the horseshoe and spike-and-slab-prior. We show that under necessary and mild assumptions one can achieve good frequentist coverage for the credible sets (subject to a possible blow up factor of the credible set).

EO040 Room Q2 NOVEL BAYESIAN APPLICATIONS AND METHODS Chair: Christopher Hans

E0781: **Astrophysical deconvolution when the convolution function is imprecise**
Presenter: **David van Dyk**, Imperial College London, United Kingdom

Estimating parameters that quantify the physical environments of solar and stellar atmospheres requires a detailed understanding of complex instrumentation and/or atomic physics, both of which can sometimes be modeled as forward convolutions. The weighting functions for these convolutions are themselves the product of precursor statistical analyses and may involve another set of parameters of scientific interest. In practice the errors arising from the precursor analyses are ignored in secondary analyses. We present a Bayesian framework for coherently accounting for these uncertainties in the secondary analysis. This framework allows us to estimate both the physical parameters of interest and the convolution functions. In principle, multiple data sets sharing common convolution functions can be combined for more precise inference. Unfortunately, however, this may allow biases stemming from misspecification in some of the analyses to spread to others. We consider how comparing the individual analyses can diagnose such bias and how the results of the combined secondary analyses can be fed back to improve estimation of their parameters of a precursor analysis.

E1082: **Causal inference from complex observational data**
Presenter: **Alexander Volfovsky**, Duke University, United States

A classical problem in causal inference is that of matching treatment units to control units in an observational dataset. This problem is distinct from simple estimation of treatment effects as it provides additional practical interpretability of the underlying causal mechanisms that is not available

without matching. Some of the main challenges in developing matching methods arise from the tension among (i) inclusion of as many covariates as possible in defining the matched groups, (ii) having matched groups with enough treated and control units for a valid estimate of average treatment effect in each group, (iii) computing the matched pairs efficiently for large datasets, and (iv) dealing with complicating factors such as non-independence among units. We propose the Fast Large-scale Almost Matching Exactly (FLAME) framework to tackle these problems for categorical covariates. At its core this framework proposes an optimization objective for match quality that captures covariates that are integral for making causal statements while encouraging as many matches as possible. We demonstrate that this framework is able to construct good matched groups on relevant covariates and further extend the methodology to incorporate continuous and other complex covariates.

E0382: **Parametric and non-parametric Bayesian approaches to spatial 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 empirically investigate urban design principles of the past half century. We will focus on one particular direction: spatial-temporal modeling of the change in crime over the past decade in the city of Philadelphia. We will show that Bayesian parametric spatial models can improve the accuracy of crime predictions (compared to simpler methods) by inducing both global and local shrinkage between proximal neighborhoods. However, there is a need for even more sophisticated crime models that take into account the geography of the city and find larger regions that share similar trends in crime over time, which motivates nonparametric approaches that can clusters of neighborhoods. Conventional Bayesian nonparametric clustering priors, such as the Dirichlet process, do not naturally handle spatial data. Thus, we will explore different strategies for introducing spatial cohesion into our nonparametric spatial clustering models.

E1016: **Formalizing the use of expert judgement in uncertainty quantification of computer models**
Presenter: **Leanna House**, Virginia Tech, United States

Deterministic computer models or simulators are used regularly to assist researchers in understanding the behavior of complex physical systems when real-world observations are limited. However, simulators are often imperfect representations of physical systems and may introduce layers of uncertainty into model-based inferences that are hard to quantify. To formalize the use of expert judgement in assessing simulator uncertainty, a method, called reification, has been previously proposed that decomposes the discrepancy between simulator predictions and reality by an improved, hypothetical computer model known as a reified simulator. One criticism of reification is that validation is, at best, challenging; only expert critiques can validate the subjective judgements used to specify a reified simulator. We develop a procedure to quantify the advantages of reification for fast, modular simulators. The procedure is explained and implemented within the context of a rainfall-runoff. We show that reification leads to informed judgements of simulator uncertainty.

EP002 Room Ground Level Hall POSTER SESSION I Chair: Elena Fernandez Iglesias

E1390: **Multivariate functional subspace methods for classifying high-dimensional longitudinal data**
Presenter: **Tatsuya Fukuda**, Chuo University, Japan
Co-authors: Toshihiro Misumi, Hidetoshi Matsui, Sadanori Konishi

Classification of high-dimensional longitudinal data with multiple classes plays an important role in various fields of science, such as medical research, meteorology and ecology, and those data are often measured at different time points for individual. Existing approaches for classifying multivariate observations are mainly restricted to data measured at same time points. We propose a novel multi-class classification procedure for high-dimensional longitudinal data based on CLAss-Featuring Information Compression (CLAFIC) method with the help of multivariate functional principal component analysis. We call this method multivariate functional subspace method. The multivariate functional subspace method can be used to classify unlabeled data according to the distance between the data and a subspace for each class, obtained by a multivariate functional principal component analysis. In modeling process of longitudinal observations to functional data, we use a smoothing technique via regularized basis expansions with Bernstein polynomials. We examine the performance of this method through the analysis of a real data set and a simulation study.

E1413: **Preliminary test estimation in system regression models using asymmetric loss functions**
Presenter: **Judy Kleyn**, University of Pretoria, South Africa
Co-authors: Sollie Millard, Mohammad Arashi

The system regression model is considered. We study the performance of the feasible preliminary test estimator both analytically and computationally, under the assumption that constraints may hold on the vector parameter space. This performance is analysed through a Monte Carlo simulation study under bounded and or asymmetric loss functions. An application of the so-called Cobb Douglas production function in economic modelling together with the results from the simulation study shows that the bounded linear exponential loss function outperforms the linear exponential loss function by comparing risk values.

E1427: **Robust Bayesian estimation using the gamma divergence**
Presenter: **Tomoyuki Nakagawa**, Tokyo University of Science, Japan
Co-authors: Shintaro Hashimoto

In the Bayesian analysis, it is well known that ordinary Bayesian estimator is not robust against outliers. The robust Bayesian estimation against outliers is proposed by using the density power divergence. They characterized the robustness in terms of the influence function. However, it is known that the estimator using the density power divergence does not work well the estimation for the scale parameter, and are unstable when the contamination ratio is not small. These facts were discussed previously in a frequentist viewpoint. However, it was shown that the estimator using the gamma divergence can make a stable estimate even when the contamination ratio is not small. We propose the robust Bayesian estimation using the gamma divergence. Furthermore, the selection of priors is also an important problem in the robust Bayesian estimation. We also propose the two type objective priors for the robust Bayesian estimation.

E1444: **An R package for Cramer-von Mises goodness-of-fit tests in regression models**
Presenter: **Sandie Ferrigno**, University Nancy Lorraine/INRIA Nancy, France
Co-authors: Marie-Jose Martinez, Romain Azais

Let $Y = m(X) + \sigma(X)\varepsilon$ be a regression model, where $m(\cdot)$ is the regression function, $\sigma^2(\cdot)$ the variance function and ε the random error term. Methods to assess how well a model fits a set of observations fall under the banner of goodness-of-fit tests. Many tests have been developed to assess the different assumptions for this kind of model. Most of them are "directional" in that they detect departures from mainly a given assumption of the model. Other tests are "global" in that they assess whether a model fits a data set on all its assumptions. We focus on the task of choosing the structural part $m(\cdot)$. It gets most attention because it contains easily interpretable information about the relationship between X and Y. To valid the form of the regression function, we consider three nonparametric tests based on a generalization of the Cramér-von Mises statistic. The first two are directional tests, while the third is a global test. To perform these goodness-of-fit tests based on a generalization of the Cramér-von Mises statistic, we have developed a R package providing an easy-to-use tool for many users. The use of the package is illustrated using simulated to compare the three implemented tests.

E1447: **On model selection via penalized likelihood for square contingency tables**
Presenter: **Kouji Tahata**, Tokyo University of Science, Japan
Co-authors: Ukyo Matsushima
The issues of symmetry and asymmetry arise naturally for the analysis of square contingency table with ordinal categories. Various types of symmetries have been proposed. When we focus on a problem of model selection, we can use some information criteria, and also restrict to hierarchical log-linear model to use the difference of likelihood ratio chi-square statistics. Recently, for the problem of model selection, penalized likelihood approaches are used in many situations, i.e., the least absolute shrinkage and selection operator. We consider methods of model selection by using the penalized likelihood for the class of certain asymmetry models. The class includes symmetry, quasi-symmetry, conditional symmetry, linear diagonals-parameter symmetry, and ordinal quasi-symmetry models. For some examples, the proposed method may be useful to select a symmetry model in the class.

E1486: **Bayesian sparse factor models with overlapping blocks**
Presenter: **Ilsang Ohn**, Seoul National University, Korea, South
Co-authors: Yongdai Kim
Sparse factor models have proven useful for describing dependency in high-dimensional data. Whereas sparsity of the factor loading matrix improves both interpretability and predictive performance, it may yield a covariance matrix having too many zero correlations, which is not desirable in many situations including genetic data where the variables are expected to be highly correlated. The aim is to propose a Bayesian sparse factor model with the corresponding covariance having overlapping block structure. The proposed factor model is able to capture strong dependency between random variables with the relatively small number of parameters. We introduce a novel prior distribution on the factor loading matrix, which provides lots of flexibility and enables scalable posterior computation. We show on a number of datasets that our model outperforms other competitors.

E1555: **The TTS R library for predicting the mechanical behavior of viscoelastic materials**
Presenter: **Mario Francisco-Fernandez**, Universidade da Coruna, Spain
Co-authors: Antonio Meneses, Salvador Naya, Javier Tarrio-Saavedra, Ricardo Cao
The TTS library, implemented in R software, has been performed to estimate the mechanical properties of amorphous materials (mainly polymers) subjected to thermal and mechanical loads. It allows us to study the reliability of amorphous polymers in the framework of material science. It is the first free software tool available to perform this task, apart from commercial software. It provides predictions of mechanical properties of amorphous viscoelastic materials, e.g. modulus and strain, at short and long observation times using the time-temperature superposition (TTS) principle and statistical analysis. The TTS package provides tools to estimate master curves that characterize materials from a thermal-mechanical point of view. The master curve is the trend of some viscoelastic property depending on time/frequency at the required temperature. It is calculated by shifting vertically and horizontally the curves obtained at other temperatures. It provides mechanical property estimates at wider ranges of time/frequencies than the experimental ones. The TTS library proposes, implements and explains step by step a new method for obtaining the shift factors in a TTS analysis. This method is based on horizontal shifting of the first derivative function of viscoelastic property curves, and it is implemented by the application of B-spline regression and bootstrap.

E1557: **Strategies for evaluation of the selected partition in cluster analysis**
Presenter: **Osvaldo Dias Lopes da Silva**, Universidade dos Acores, Portugal
Co-authors: Aurea Sandra Toledo de Sousa
In cluster analysis there are several problems, among which, we highlight the absence of knowledge about the real structure of the data, due to the fact that we cannot recognize artificial structures imposed by the algorithms used. The aim is to present a methodological approach in order to evaluate the adequacy of the selected partition, as the most significant, based on stopping criteria rules. The comparison of a partition obtained from real data with several partitions, with the same number of clusters, obtained from randomly generated matrices in the reference hypothesis of absence of structure, allows us to test if the partition under evaluation has a structure which is relevant, that stands out from the inevitable influence of the algorithms used. The application of this methodology is exemplified with a real dataset and the main conclusions are presented and discussed.

E1573: **Joint modeling for longitudinal data and binary outcome via h-likelihood**
Presenter: **Toshihiro Misumi**, Yokohama City University, Japan
Joint modeling techniques of longitudinal covariate and binary outcome have attracted considerable attention in medical research areas. Joint models provide a powerful tool to explore how strongly associated is a longitudinal trajectory of biomarker with an event of interest. The strategy for estimating joint models is to define a joint likelihood based on two sub-models with shared random effects, i.e. linear random effects models for the longitudinal sub-model, and logistic models with random effects for the binary sub-model. A numerical integration, however, is required in the estimation algorithm for the joint likelihood, and a computational problem arises as the assumed sub-models become more complex. In order to overcome the issue, we propose a joint modeling procedure by using a h-likelihood to avoid the numerical integration in the estimation algorithm. The estimating procedure is expected to reduce the computational cost. We conduct some Monte Carlo simulation studies to examine the effectiveness of our proposed modeling procedure, and then apply the method to the analysis of the real data.

E1618: **A fast automatically calibrated resampling method for evaluating multinomial model fit**
Presenter: **Ioulia Papageorgiou**, Athens University of Economics and Business, Greece
Many tests of goodness-of-fit can be reduced to testing a hypothesis about the parameters from a multinomial distribution. However, the classic multinomial goodness-of-fit test based on Pearson's χ^2 is only asymptotic and may be biased in small samples or when the contingency table is sparse. A general goodness-of-fit approach has been recently proposed in the literature using calibrated simulation. Comparative data are generated based on the maximum-likelihood estimates from the observed data to assess systematic discrepancies between observed and simulated data from the fitted model. In common with the posterior predictive p-value in Bayesian statistics, the p-value of the approach is generally non-uniform under the null model, and calibration is required to check its significance. We introduce a simple variant of the calibrated simulation approach and provide theory-based modifications that result in uniform p-values in multinomial goodness-of-fit testing, thereby removing the need for calibration. We illustrate and evaluate the method in a variety of contexts including item response theory modelling, latent class analysis and capture-recapture data modelling. The new method is shown to have nominal type I error rates whilst being computationally much less intensive than alternative resampling methods.

E1632: **Outlier detection using auxiliary variable dependent monitoring in scanner data**
Presenter: **Youngrae Kim**, Seoul National University, Korea, Korea, South
Recently, statistical agencies of major countries analyze the scanner data, which include the information about transaction, such as price and quantity. It can be important to detect sudden change or anomalies in transactions. We turn this problem into outlier detection problem in price change rates. Because the existing methods use only the transaction price information, there is a difficulty that can not deal with the situation with the abnormal sales amount. We develop the procedure to detect the anomaly of the transaction by considering both the transaction price and quantity information. We estimate variance of price change rate using quantity information with kernel estimation method, to create abnormal point

detection criteria. And we can show the criterion on the control chart. We divided the two periods, estimation period and monitoring period, and propose appropriate length of estimation period through a simulation study. We numerically compare the performance of our method to existing methods which did not use volume information.

E1646: **A control chart methodology for functional data**
Presenter: **Javier Tarrio-Saavedra**, Universidade da Coruna, Spain
Co-authors: Miguel Flores, Salvador Naya, Ruben Fernandez Casal, Veronica Bolon, Carlos Eiras
Anomaly detection in the industry and the control of the quality of products and services have usually been developed by the application of univariate and multivariate control charts. However, the problem of ongoing monitoring of data (due to the advances in sensoring in the framework of the Industry 4.0) requires more sophisticated tools that can be applied to autocorrelated critical to quality variables. Many of these new data, generally curves, can be studied as functional data. This is the case of energy consumption, temperatures and relative humidity among other variables, measured in buildings. These new complex data require of innovative solutions by researchers in statistical quality control based on the application of functional data analysis techniques (FDA. A methodology to build process control charts for functional data is proposed. The control consists of two phases: Phase I of process calibration and Phase II of process monitoring. For Phase I, a control chart for functional data based on functional data depth and outlier detection is developed. In Phase II, another control chart based on rank control charts and functional depth is also proposed to monitor the process. A comprehensive simulation study and real data application have been performed. Variable selection methods are also studied in advance to FDA control charts.

E1650: **A new nonparametric estimator of the regression function**
Presenter: **Issam El Hattab**, ENCG-Casablanca, Hassan 2 University, Morocco
A new kernel-type estimator of the regression function is considered. The proposed methodology is based on expressing the regression function in terms of a copula density function. There are basically no restrictions on the choice of the kernel function in our setup, apart from satisfying some mild conditions. The selection of the bandwidth, however, is more problematic. Under some regularity conditions, we establish the asymptotic properties for the proposed estimator, namely uniform-in-bandwidth consistency with exact rate. Some numerical studies using Monte Carlo simulations and an empirical application are given to examine the finite-sample performance of our methods.

E1651: **Trends in extremes of storm environments**
Presenter: **Jonathan Koh**, EPFL, Switzerland
Co-authors: Erwan Koch, Anthony Davison
Severe thunderstorms can have devastating impacts on society. Concurrently high values of Convective Potential Energy (CAPE) and Storm Relative Helicity (SRH) have been found to be favourable to severe weather. Hence, it is highly relevant to model the probabilistic behaviour of their extreme values. We consider a large area of the Continental US over the period 1979-2015. Using extreme-value theory and an appropriate multiple testing procedure, we show that there is a time trend in the location parameter of the generalized extreme-value distribution for SRH maxima in April and May, for CAPE maxima in April, and for the maxima of the product variable SRH $\times\sqrt{\text{CAPE}}$ in April and May. Moreover, we show that the El Nino-Southern Oscillation appears to be a good covariate for the location parameter of the generalized extreme-value distribution for SRH maxima in February. These meteorological results and extreme-value analyses are consistent with what is expected under global warming; as such, they are of significant interest to the severe thunderstorms research community.

E1686: **Probabilistic clustering methods in data analysis of macro-datasets**
Presenter: **Aurea Sandra Toledo de Sousa**, Universidade dos Azores , Portugal
Co-authors: Helena Bacelar-Nicolau, Osvaldo Dias Lopes da Silva, Leonor Bacelar-Nicolau
The extraction of useful knowledge from huge data sets stored in large databases is an important task. One possible solution for analysing high-dimensional micro-data sets is the prior identification of classes (usually of individuals) in such databases, whose description is then made through macro-data matrices (also called symbolic data matrices). From a proximity matrix containing similarities or dissimilarities between the pairs of elements to be classified, either classic or probabilistic aggregation criteria can subsequently be applied. We use hierarchical clustering methods based on the weighted generalized affinity coefficient, and on probabilistic aggregation criteria, which apply the transformation by the probabilistic distribution function of appropriated sample statistics. The most relevant clustering structures, obtained with the hierarchical clustering analysis of two datasets taken from the literature of complex data analysis, are described. Their results are compared to those obtained with other clustering algorithms. The results show that the clustering probabilistic approach performs well over both datasets, which also happen with similar previous studies on either simulated or real data sets.

E1720: **Structure learning of continuous time Bayesian networks via penalized likelihood methods**
Presenter: **Maryia Shpak**, Uniwersytet Marii Curie-Sklodowskiej w Lublinie, Poland
Co-authors: Blazej Miasojedow
A Continuous Time Bayesian Network (CTBN) is a time homogeneous Markov process, which is decomposed into processes whose transition intensities depend on the other processes in the network. The dependence structure between the intensities is encoded by a directed graph. CTBNs are widely used for modeling different phenomena from biology, chemistry, social sciences among many others. The problem of learning the structure of the CTBN is a challenging task. We present the solution to this problem based on the penalized maximum likelihood method. Our method can be applied to both fully and partially observed data. In the case of partially observed data we propose the efficient MCMC algorithm to solve the underlying optimization problem. Both theoretical and numerical results will be presented.

E1470: **Overcoming inequalities in intergroup competitions: Heterogeneous preferences and cultural values**
Presenter: **Francisca Jimenez-Jimenez**, Universidad de Jaen, Spain
Previous research provides strong evidence that competition between groups enhances intragroup cooperation but, also, that intergroup inequalities might discourage cooperative behaviour. We design an experiment to control the degree of intergroup and intragroup (in)equality in a context which has shown to be successful in promoting cooperation: an all-can-win intergroup competition. By forming groups according to individuals' propensities to cooperate, we investigate the role and persistence of players' types in a competitive context. We are also interested in analysing the relationship between cooperative/competitive behaviour and cultural values concerned on (in)equality, as vertical and horizontal individualism and collectivism. We find that the positive effect of all-can-win competition on cooperation is robust to intergroup inequalities. With competition, significant types-based differences vanish at group-level and individual-level. Surprisingly, the groups that are ranked 1st more frequently are those formed by the top cooperators and by the bottom cooperators. Lastly, we find that the role of cultural values to explain cooperation depends on contextual conditions rather than being structural personality traits.

E1467: **Simulation of links between temperature extremes and atmospheric circulation in regional climate models**
Presenter: **Jan Kysely**, Institute of Atmospheric Physics AS CR, Czech Republic
Co-authors: Eva Plavcova
Atmospheric circulation is an important driver of temperature anomalies and extreme temperature events in mid-latitudes. Since climate models

suffer from biases in the simulation of temperature extremes, it is important to study whether those biases are associated with large-scale atmospheric flow. We evaluate links of summer heat waves/winter cold spells to circulation types in an ensemble of 19 EURO-CORDEX regional climate model (RCM) simulations in central Europe and compare them against observations over 1980-2005. We find that the RCMs reproduce in general the observed circulation significantly conducive to temperature extremes, with zonal flow reducing probability of extremes in both seasons, while advection of warm/cold air from the south-easterly/north-easterly quadrant playing dominant role in developing heat waves/cold spells. Because of these links, simulation of temperature extremes in the RCMs is strongly affected by biases in atmospheric circulation. For almost all examined simulations, the persistence of circulation supertypes (i.e. types with common flow characteristics grouped together) is significantly overestimated, which may contribute to enhanced tendency to group days with large temperature anomalies into sequences. This bias is manifested in development of too-long heat waves/cold spells in the RCMs, and points to limited credibility of possible future scenarios of these temperature extremes based on the RCM simulations.

E1755: **Group iterative hard-thresholding and generalized linear models in genetics**
Presenter: **Benjamin Chu**, University of California, Los Angeles, United States
Co-authors: Kevin Keys, Kenneth Lange, Janet Sinsheimer, Hua Zhou

SNP-by-SNP (single nucleotide polymorphism) association testing is currently the de-facto statistical analysis employed for Genome Wide Association Studies (GWAS). This analysis approach ignores joint effects of SNPs. Iterative hard-thresholding (IHT) is one of the most scalable algorithms that performs multivariate model selection without shrinking effect sizes and circumvents the use of p-values. We implemented IHT in Julia to analyze GWAS data as a module under the open sourced statistical genetics ecosystem: OpenMendel. We modified the hard-thresholding operator to enforce sparsity on a group-level as well as within-group level. This accommodates for linkage disequilibrium because only the top predictors in each group are selected. Then we extend the framework of IHT to generalized linear models (GLM) so we can model non-continuous, non-normal response data. Our implementation enjoys built-in parallelism. We applied IHT on real and simulated datasets to demonstrate model quality, algorithm robustness, and scalability. For geneticists, our method offers multivariate model selection and maintains comparable speed/memory usage to some of the fastest algorithms available today. For theorists, we investigate properties of the group hard-thresholding operator, and derive best step sizes for IHT in the GLM setting.

E1756: **Penalty method for variance component selection**
Presenter: **Juhyun Kim**, University of California, Los Angeles, United States
Co-authors: Hua Zhou

Variance components models, also known as mixed effects model, are central themes in statistics. When there is a large number of variance components, one may wish to select a subset of those that are associated with response. Existing methods are limited to finding random components at individual level or within one variance component. We propose selection of variance components based on a penalized log-likelihood with adaptive penalty. This is achieved via a majorization-minimization (MM) algorithm, which is well known for being simple, numerically stable, and easy to implement. Performance of the proposed methodology is shown empirically through simulation studies and real data analysis. In theory, we establish a non-asymptotic error bound for the output from the algorithm and characterize the region in which the MM iterates converge to a global optimum of the population likelihood. This result provides a theoretical guideline in terminating MM iterations.

E1762: **Statistical approaches of the extreme rainfall events variability using hourly data, NW Spain**
Presenter: **Elena Fernandez Iglesias**, University of Oviedo, Spain
Co-authors: Gil Gonzalez-Rodriguez, Jorge Marquinez

Flash floods represent a significant natural hazard in the small mountainous catchments in the NW of Spain. More than 2500 torrents are affected by floods related to short-term storm rainfall. Several studies on rainfall variability have been carried out in Spain. Most of them are focused on the Mediterranean area by using daily data. Little or marginal information is available with hourly data, especially in the Cantabrian area. The aim is to analyze the behavior of these kinds of extreme rainfall events during the last five decades in the Spanish Cantabrian area by using selected hourly precipitation data from six meteorological gauges. By considering large quantiles cutoff, different scenarios of extreme events are handled. The increasing trend of annual series related with the number, duration and total precipitation of these extreme events is statistically analyzed by applying bootstrap tests of isotonicity.

CI017 Room A0 BAYESIAN MACROECONOMETRICS Chair: Toshiaki Watanabe

C0175: **A Bayesian approach for inference on probabilistic surveys**
Presenter: **Marco Del Negro**, Federal Reserve Bank of New York, United States

A non-parametric Bayesian approach is proposed to the estimation of forecast densities in probabilistic surveys. We use it to study the evolution of the subjective forecast distribution for the U.S. Survey of Professional Forecasters over the past forty years, focusing especially on second moments. We show that the variance of aggregate forecast distribution fell substantially from the eighties to the nineties (the "conquest"), and fell again after the Fed announced its long term inflation goal. We also show that disagreement (heterogeneity in the mean forecasts) plays a minor role, but that heterogeneity in uncertainty is very large. The "conquest" amounted to convincing high-uncertainty forecasters that inflation is under control. We also find that only a fringe of forecasters place any significant probability of the possibility of a return to the seventies. The likelihood of deflation in the aftermath of the Great Recession was significant (almost ten percent for the average forecaster) but declined to one percent or less for most forecasters thereafter.

C0176: **Economic predictions with big data: The illusion of sparsity done**
Presenter: **Giorgio Primiceri**, Northwestern University, United States

The aim is to compare sparse and dense representations of predictive models in macroeconomics, microeconomics and finance. To deal with a large number of possible predictors, we specify a prior that allows for both variable selection and shrinkage. The posterior distribution does not typically concentrate on a single sparse or dense model, but on a wide set of models. A clearer pattern of sparsity can only emerge when models of very low dimension are strongly favored a priori.

C0180: **Non-Markovian regime switching models**
Presenter: **Chang-Jin Kim**, University of Washington, United States
Co-authors: Jaeho Kim

The non-Markovian regime switching model is revisited. This model employs an autoregressive continuous latent variable in order to specify the dynamics of the latent regime-indicator variable. We show that, in spite of the non-Markovian nature of the regime indicator variable, the Markovian property of this continuous latent variable allows us to easily estimate the model within the Bayesian framework without any approximations. In particular, we show that the conventional Gibbs sampling is enough in generating the regime indicator variable as well as the continuous latent variable conditional on all the parameters of the model and data. For an application to business cycle modeling of postwar US real GDP, a modified version of a Markovian switching model is slightly preferred to a non-Markovian switching model by the Bayesian model selection criterion. For an application to volatility modeling of the weekly stock return, a non-Markovian switching model with endogenous switching or the leverage

effect is strongly preferred to Markovian switching models.

CO578 Room P1 ADVANCES IN SPATIAL ECONOMETRICS Chair: Federico Martellosio

C0805: **Misspecification testing in spatial autoregressive models**
Presenter: **Francesca Rossi**, University of Verona, Italy
Co-authors: Jungyoon Lee, Peter CB Phillips
Spatial autoregressive (SAR) and related models offer flexible yet parsimonious ways to model spatial or network interaction. However, SAR specifications typically rely on a particular parametric functional form and an exogenous choice of the so-called spatial weight matrix with only limited guidance from theory in making these specifications. Moreover, the choice of a SAR model over other alternatives, such as Spatial Durbin (SD) or Spatial Lagged X (SLX) models, is often arbitrary, raising issues of potential specification error. To address such issues, we develop an omnibus specification test within the SAR framework that can detect general forms of misspecification including that of the spatial weight matrix, functional form and the model itself. The approach extends the framework of conditional moment testing to the general spatial setting. We derive the asymptotic distribution of our test statistic under the null hypothesis of correct SAR specification, show consistency of the test, and provide local power properties. A Monte Carlo study is conducted to study finite sample performance of the test.

C1079: **Consistent specification testing under network dependence**
Presenter: **Abhimanyu Gupta**, University of Essex, United Kingdom
Co-authors: Xi Qu
A series-based nonparametric specification test for a regression function are presented when data are dependent across a network. The framework permits network dependence to be parametric, parametric with increasing dimension, semiparametric or any combination thereof, thus covering a vast variety of settings. These include spatial error models of varying types and levels of complexity. Furthermore, we also cover models in which network dependence arises directly in outcome variables, possibly with dependence complexity increasing with sample size. Despite being applicable so generally, our test statistic is easy to compute and asymptotically standard normal. To prove the latter property, we present a central limit theorem for quadratic forms in linear processes in an increasing dimension setting that may be of independent interest. Finite sample performance is studied in a simulation study and empirical examples illustrate the test with real-world data.

C0406: **How to avoid the zero-power-trap in testing for correlation**
Presenter: **David Preinerstorfer**, Universita libre de Bruxelles, Belgium
In testing for autocorrelation of the errors in regression models the power function of, e.g., "point-optimal-invariant" or "locally-best-invariant" tests can get very low as the correlation gets strong (as sample size stays fixed). We discuss and revisit theoretical results concerning the occurrence of this "zero-power-trap" phenomenon, and suggest ways to avoid it.

C0693: **Matrix spatial specification models**
Presenter: **Samantha Leorato**, University of Rome Tor Vergata, Italy
Co-authors: Andrea Martinelli
Spatial linear regression models with spatial dependence in the errors and in the dependent variable are studied. The spatial dependence is modeled by arbitrary matrix functions, V and M respectively, indexed by a scalar parameter and, eventually, by two (possibly distinct) weight matrices, D and W. This family of models encompasses the main models used in the spatial econometric literature, such as SARAR and MESS models. We define the quasi maximum likelihood estimator and study its asymptotic properties under non-Gaussian errors. By doing so, we provide some insights into the difference between specifications, with emphasis on advantages and shortcomings as well as on interpretation of the parameters and correspondences between models.

C1133: **Adjusted maximum likelihood inference for spatial fixed effects models**
Presenter: **Federico Martellosio**, University of Surrey, United Kingdom
In a likelihood framework, an often successful way to deal with incidental parameters is to "adjust" the profile score of the parameter of interest. We consider the adjusted quasi-maximum likelihood estimator (QMLE) of the spatial parameter in a spatial panel model with individual and/or time fixed effects. The adjusted QMLE coincides with the QMLE if covariates are not present in the model. When covariates are present, the adjusted QMLE can be more accurate than the QMLE. Saddlepoint confidence intervals for the spatial autoregressive parameter based on the adjusted QMLE are proposed. In simulation, they perform very well against other higher-order methods.

CO168 Room A2 ADVANCES IN TIME SERIES AND FINANCIAL ECONOMETRICS Chair: Roxana Halbleib

C0233: **U.S. aggregate output measurement: A common trend approach**
Presenter: **Gabriele Fiorentini**, University of Florence, Italy
Co-authors: Tincho Almuzara, Enrique Sentana
Signal-extraction techniques are applied to US GDP and GDI to produce an improved aggregate output measure, emphasising the presence of a common trend in levels whose absence would imply an empirically implausible diverging statistical discrepancy. We also study the consequences of ignoring this common trend, which can introduce important biases in maximum likelihood estimators and considerably reduce the Kalman smoother precision. Our theoretical and Monte Carlo results characterise the severity of misspecification as inversely proportional to an R^2 measure of common trend observability. This R^2 is high in the data (1952Q1-2015Q4). Therefore, we conclude misspecification is small but not negligible.

C0457: **Estimation of multivariate factor stochastic volatility models**
Presenter: **Christian Muecher**, University of Konstanz, Germany
Co-authors: Roxana Halbleib, Giorgio Calzolari
A frequentist procedure is introduced to estimate multivariate factor stochastic volatility models. The estimation is done in two steps. First, the factor loadings and idiosyncratic variances are estimated by approximating the dynamic factor model with a static one. Second, we apply the Efficient Method of Moments with GARCH(1,1) as an auxiliary model to estimate the stochastic volatility parameters governing the dynamic latent factors and idiosyncratic noises. Based on various simulations, we show that our procedure outperforms existing approaches in terms of accuracy and efficiency and it has clear computational advantages over the existing Bayesian methods.

C0786: **Modelling temporal dependence of realized variances with vines**
Presenter: **Yarema Okhrin**, Universitaet Augsburg, Germany
Models for realized volatility are proposed which take the specific form of temporal dependence into account. Current popular methods use the idea of mixed frequencies for forecasting of the realized volatility, but neglect the potential non-linear and non-monotonic temporal dependence. The proposed approaches utilize vine copulas to mimic different memory properties. As a linear benchmark, we choose HAR and MIDAS models, both of which account for the persistence of volatility. Both benchmarks can be seen as special cases of the suggested modelling approach. The models

are evaluated within an extensive empirical study both in- and out-of-sample. The forecasting ability of competing models is compared statistically and it shows that vine methods are significantly superior over the linear approach in modeling time dependencies of realized volatilities.

C0862: **Indirect inference estimation of misspecified DSGE asset pricing models using nonlinear vector autoregressions**
Presenter: **Julie Schnaitmann**, Universitat Konstanz, Germany
Co-authors: Joachim Grammig, Dalia Elshiaty

A two-step indirect inference strategy is proposed in order to analyze a class of DSGE asset pricing models. These models, which combine a one-sector stochastic growth model with external habit preferences and capital adjustment costs, have so far only been calibrated but not made subject to econometric analysis. The proposed strategy draws on and extends the idea of estimating misspecified DSGE models by sequential partial indirect inference (SPII). Drawing on the SPII philosophy, we use binding functions that facilitate the consistent estimation of some structural model parameters of interest, while treating others as nuisance parameters. We acknowledge that these parameters do not necessarily capture economic reality, but they are necessary to generate model-implied data. Moreover, we separate the estimation of the parameters dependent on the parts of the asset pricing model that they influence. Specifically, they are classified into technology parameters, which are estimated separately in the first step. The estimation of the investor preference parameters is performed in the second step, using the first-step estimates as input. The auxiliary parameters are delivered by the class of recently developed nonlinear vector autoregressive (NVAR) models, which allow for more flexible impulse-response functions than a standard VAR model.

C0706: **Determination of vector error correction models in high-dimensions**
Presenter: **Melanie Schienle**, Karlsruhe Institute of Technology, Germany
Co-authors: Chong Liang

A shrinkage type methodology is provided which allows for simultaneous model selection and estimation of vector error correction models (VECM) when the dimension is large and can increase with sample size. Model determination is treated as a joint selection problem of cointegrating rank and autoregressive lags under respective practically valid sparsity assumptions. We show consistency of the selection mechanism by the resulting lasso-VECM estimator under very general assumptions on dimension, rank and error terms. Moreover, with computational complexity of a linear programming problem only, the procedure remains computationally tractable in high dimensions. We demonstrate the effectiveness of the proposed approach by a simulation study and an empirical application to recent CDS data after the financial crisis.

CO122 Room B2 TOPICS IN FINANCIAL ECONOMETRICS Chair: Leopold Soegner

C0392: **Risk reduction and portfolio optimization using clustering methods**
Presenter: **Anna-Katharina Thoes**, TU Kaiserslautern, Germany
Co-authors: Joern Sass

Diversification is one of the main pillars of investment strategies. The prominent $1/N$-portfolio, which puts equal weight on each asset, is apart from its simplicity a method which is hard to outperform in realistic settings. But depending on the number of considered assets this method can lead to very large portfolios. We investigate how the number of assets can be reduced and which advantages and disadvantages arise. The idea is to reduce the number of chosen assets by clustering. Using clustering techniques the possible assets are separated into non-overlapping clusters and the assets within a cluster are ordered by their Sharpe ratio. Then the best asset of each portfolio is chosen to be a member of the new portfolio with equal weights, the cluster portfolio. We show that this portfolio inherits the advantages of the $1/N$-portfolio and can even outperform it empirically. To this end different clustering methods and performance measures are used to compare the portfolios on simulated and real data. To explain the observations on real data, we finally derive corresponding results in different model settings.

C0421: **Optimal granularity for portfolio choice**
Presenter: **Katarina Lucivjanska**, Pavol Jozef Safarik University in Kosice, Slovakia
Co-authors: Alex Weissensteiner, Nicole Branger

Many optimization-based portfolio rules fail to beat the simple $1/N$ rule out-of-sample because of parameter uncertainty. We suggest a grouping strategy in which we first form groups of equally weighted stocks and then optimize over the resulting groups only. This strategy aims at balancing the trade-off between the benefits from optimization and the losses from estimation risk. We rely on Monte-Carlo simulations to illustrate the performance of the strategy, and we derive the optimal group size for a simplified setup. Furthermore, we show that estimation risk also has an impact via the criterion by which the assets are sorted into groups (like the expected excess returns or betas), but does not negate the grouping approach. Out of sample back-tests confirm the validity of our grouping strategy empirically.

C0606: **Estimators of the boundary in inverse first exit problems**
Presenter: **Klaus Poetzelberger**, WU Vienna, Austria

First-passage problems for the Brownian motion (W_t) or general diffusion processes, have important applications. Given a boundary $b(t)$, the distribution of the first-exit time τ^b has to be computed, in most cases numerically. Inverse boundary crossing probabilities assume that the distribution of τ^b is given and the boundary b has to be found. The boundary and the density of τ^b satisfy a Volterra integral equation. We propose and analyze estimators of b. The first class of estimators are solutions of stochastic versions of the Volterra equation. The second class of estimators uses the idea of approximating the boundary $b(t)$ by a piecewise boundary $\tilde{b}_m(t)$. Define $W^m = (W_{t_1}, \ldots, W_{t_m})$. The density of $\tau^{\tilde{b}_m}$ given $W^m = w^m$ is available in closed form. The EM estimator iterates the estimation and maximization steps. The Bayesian estimator additionally chooses a prior on b and then uses Gibbs sampling to iterate the generation of $b|(W^m, \tau_1, \ldots, \tau_n)$ and $W^m|(b, \tau_1, \ldots, \tau_n)$. Typical inverse problems are sequential testing in statistics or the estimation of a ruin boundary, for instance in credit risk modelling. A company defaults, if a process (V_t), called the value of the firm, crosses a boundary $b(t)$. (V_t) cannot be observed. It is correlated with (S_t), which includes published relevant information on (V_t).

C0846: **Sensitivity of boundary crossing probabilities of the Brownian motion**
Presenter: **Sercan Gur**, Vienna University of Economics and Business, Austria
Co-authors: Klaus Poetzelberger

The aim is to analyze the sensitivity of boundary crossing probabilities of the Brownian motion to perturbations of the boundary. The first and second order sensitivities, i.e. the directional derivatives of the probability are derived. Except in cases where boundary crossing probabilities for the Brownian bridge are given in closed form, the sensitivities have to be computed numerically. We propose an efficient Monte Carlo procedure.

C1146: **Price efficiency in markets with stochastic latency**
Presenter: **Stefan Voigt**, WU (Vienna University of Economics and Business), Austria
Co-authors: Nikolaus Hautsch, Christoph Scheuch

The purpose is to analyze how stochastic latency in the transaction settlement process, as introduced by distributed ledger systems, affects price efficiency. The time-consuming settlement process exposes arbitrageurs to price risk and imposes limits to arbitrage. We derive boundaries to price efficiency imposed by stochastic latency and show that larger price differences are consistent with higher expected latency, higher uncertainty in latency, or higher volatility. We parametrize stochastic latency in the Bitcoin network and estimate boundaries for high-frequency orderbook data

from several exchanges. Stochastic latency explains about 74% of observed price differences adjusted for transaction costs.

CO320 Room C2 ADVANCES IN CREDIT RISK MODELLING Chair: Raffaella Calabrese

C0366: **Identifying hidden patterns in credit risk survival data using generalised additive models**
Presenter: **Jonathan Crook**, University of Edinburgh, United Kingdom
Co-authors: Viani Djeundje
Survival models for credit card default have several advantages over cross sectional credit scoring models. They provide more information to analysts, they can be used to model provisions for IFRS9 and the inclusion of macroeconomic variables means they can be used for stress testing. However, in such models the hazard of default is typically expressed as a simple linear combination or covariates. We investigate the predictive accuracy of survival models if the hazards are expressed as GAMs. Specifically, we parameterise hazard models in terms of penalised splines. We estimate the parameters using frequentist and Bayesian methods applied to a large portfolio of credit card accounts. We compare the predictive accuracy of models without splines with those with splines, applied in turn to application, behavioural, to macroeconomic variables, and to models with all of these types of variables. We find that GAM specifications have higher predictive accuracy than linear models. The results suggest that some applications for accounts may be less attractive to a lender when GAMs are used rather than linear models. Similarly, expected profit may be more accurately estimated with a GAM model than a simple linear model.

C0529: **Behavioural attitudes and financial performance: New ideas for segmenting bank customers**
Presenter: **Caterina Liberati**, University of Milano-Bicocca, Italy
Co-authors: Galina Andreeva
Credit scoring models are generally trained using customers credit history and demographics. Recent works in interdisciplinary studies have showed that alternative source of information, such as psychological traits or behavioural attitudes, can aid to improve default prediction. We would like to verify if an explorative segmentation based on financial knowledge, preferences and personality traits is able to detect different customer typologies also in terms of financial performance. This could be the case when the bank management has to deal with new clients or with those without a long banking history. The segmentation has been derived employing hierarchical clustering on factorial scores coming from non-linear Principal Component Analysis (PCA). The kernel-PCA allowed data to be mapped indirectly in a very-high-dimensional space F where is simple to construct a hyperplane that divides the points into arbitrary clusters. The choice of the kernel functions and its parameters provided different kernel factors which produced, in turn, alternative clusters solutions. All the partitions have been ranked using alternative criteria that measure aspects of clusters validity.

C0665: **Access to finance and growth of innovative SMEs after Brexit**
Presenter: **Marta DeglInnocenti**, University of Southampton, United Kingdom
Co-authors: Raffaella Calabrese, Si Zhou
A new perspective is offered on the link between the literature on firms' investment decisions during period of uncertainty and the literature on access to finance for innovative Small and Medium Enterprises (SMEs). In particular, it provides nuanced evidence about innovative SMEs expectations of growth and their ability to access debt finance after Brexit. By using a unique survey, we find that innovative SMEs expect to be more financially constrained after Brexit. Furthermore, the results show that innovative SMEs have not only changed their strategies by cutting their employment, but also expect lower growth. Finally, we also find that there is a spatial bias in expectations; innovative firms outside London expect to be more financially constrained and to grow less than those located in London.

C0804: **A Bayesian spatial sample selection model with an application to credit constraints for small businesses**
Presenter: **Michaela Kesina**, ETH Zurich, Switzerland
Co-authors: Raffaella Calabrese
A spatial autoregressive probit model is developed that corrects for sample selection bias. We consider a two-step model with a spatial lag of a latent variable in both the selection and outcome equation. We suggest to jointly estimate both steps using a Bayesian Markov chain Monte Carlo (MCMC) simulation approach. We explore the finite sample properties of the estimator using Monte Carlo simulations. We apply the proposed model to data on UK Small and Medium-size Enterprises (SMEs) to estimate how neighbourhood effects can help to explain the external financing decisions of UK SMEs and their access to credit.

C1106: **Network-based PARX models to measure contagion in credit default counts**
Presenter: **Paolo Giudici**, University of Pavia, Italy
Co-authors: Arianna Agosto
PARX models (Poisson Autoregression with Exogenous Covariates) are based on the assumption that a count time series, conditional on its past, follows a Poisson distribution with a time-varying autoregressive intensity whose formulation includes a set of exogenous covariates. The model was applied previously to the time series of monthly defaults among Moody's rated industrial companies in the period 1982-2012. The weakness of the approach, when applied to credit risk modelling, is the way it defines contagion: the autoregressive component may be influenced by misspecification or latent effects. The aim is to overcome this problem by explicitly defining a contagion component, based on a network model formulation, considering several time series corresponding to different economic sector and/or geographical areas.

CO148 Room D2 ENERGY ECONOMICS Chair: Francesco Ravazzolo

C0714: **A copula-based modelling of peak district heating demand and outdoor temperature**
Presenter: **Andrea Menapace**, Free University of Bozen-Bolzano, Italy
Co-authors: F Marta L Di Lascio, Maurizio Righetti
The complex dependence relationship between peak district heating demand and outdoor temperature is examined by using the copula function. The aim is to derive a copula-based conditional probability distribution of heat demand given weather conditions and provide a probability law sensitive to extreme events. We analyse data concerning the district heating system of the city of Bozen-Bolzano. Daily maxima heating demand and the corresponding outdoor temperature have been observed from January 2014 to November 2017. We perform a three-step analysis. First, the serial correlation of the two time series of interest, taken separately, is analysed through a seasonal autoregressive integrated moving average model. Second, the dependence relationship between the two series of uncorrelated residuals is investigated using copula models. Third, the conditional probability distribution of peak heat demand given outdoor temperature is derived. The selected copula model indicates that the two investigated phenomena tend to co-move closely together for the whole observed period and, when taking into account the conditional behaviour of interest, in the case of extreme climatic events, we find a high probability of thermal energy demand reaching its peak. From this finding we derive useful implications for the management and production of thermal energy.

C1002: **Measuring the spillover effects of commodities prices and international markets on the Russian stock exchange**
Presenter: **Robert Hornegold**, Heriot-Watt University, United Kingdom
Co-authors: Marco Lorusso, Michele Costola
The connectedness between Russian economy sectors (including the government sector), commodities prices and international markets is investigated by using a previous approach. The analysis is based on variance decompositions from high-dimensional vector autoregressions to characterize connectedness in returns vs. volatility spillovers for the sample period 2005-2018. We focus on specific events related to shocks in the Russian economy, commodities market and international markets. The estimated results show that both returns and volatility of the Russian stock market index (MOEX) are greatly influenced by periods of persistent high and low oil prices. Moreover, we find that Oil & Gas is the Russian sector that mostly affects MOEX. The contribution of this industry is particularly strong during two periods: the plunge of oil prices started in 2015 and the international sanctions imposed by US and EU to the Russian economy in 2013. Interestingly, the Oil & Gas industry has important spillover effects on both returns and volatility of Russian government bonds. As far as other commodities are concerned, coal makes the most significant contribution to MOEX returns, whilst natural gas makes the greatest contribution to MOEX volatility for the sample period. Finally, our findings indicate that EU and US stock markets have an important influence on MOEX returns during the global financial crisis of 2007-2009.

C1174: **The RES-induced switching effect across fossil fuels: An analysis of day-ahead and balancing prices**
Presenter: **Angelica Gianfreda**, Free University of Bozen-Bolzano, Italy
Co-authors: Lucia Parisio, Matteo Pelagatti
The empirical literature on electricity markets has highlighted a strong cointegrating relationship governing the dynamics of electricity and fuel prices. More recently, the massive introduction of RES in electricity generation, fostered by generous supporting schemes, has influenced the shape and position of the supply function and consequently the equilibrium prices. We believe that the new competitive scenario may have influenced the fuel-electricity nexus with a different impact in day-ahead and balancing markets. Taking into account the Northern Italian zone characterized by a high solar PV and hydro penetration, we provide empirical evidence of the evolving fuels-electricity nexus across two samples characterized by low (2006-08) and high (2013-15) RES levels. We conduct the analysis taking into account both day-ahead and, for the first time, balancing market sessions. Results indicate that fuel prices are much less relevant in determining the dynamics of electricity prices in recent years characterized by high RES penetration. On the contrary, taking into account flexible thermal sources, we show that in the second sample balancing and fuel prices (especially gas) are in a long run equilibrium.

C1188: **Value at risk and extreme value theory for oil prices: a long memory multivariate GARCH approach**
Presenter: **Malvina Marchese**, Cass Business School, United Kingdom
Co-authors: Malvina Marchese, Francesca Di Iorio
The purpose is to examine the performance of several short and long memory multivariate volatility models for the crude oil spot and futures returns of three major oil markets, namely Brent WTI and Dubai, to estimate value and risk and the expected shortfalls for long and short positions. The results show that long memory multivariate GARCH models with asymmetries significantly over perform their short memory counterparts from a risk management perspective. Optimal portfolio weights and optimal hedge ratios are calculated with the fractionally integrated DCC model in order to suggest a crude oil hedge strategy. The empirical findings suggest holding crude oil futures to spot for Brent and Dubai and oil spot to futures for WTI. Finally, we show that the fractionally integrated DCC model is most effective in term of reducing the variance of the portfolio.

C0947: **Global oil market uncertainty and oil prices**
Presenter: **Bing Xu**, Heriot-Watt University, United Kingdom
New measures of global oil-market uncertainty are proposed and they are related to the real price of oil. Supply and demand are clearly important, a factor which has been overlooked in this debate is the impact of uncertainty on the oil market. So far, the literature has primarily relied on macroeconomic or oil price uncertainty proxies such as the implied or realized volatility of stock market returns or oil prices, the cross-sectional dispersion of forecasts. The potential issues are they are based on structure of specific theoretical models or relied on single or small number of observable indicators. To the best of our knowledge, we are the first one to use factor augmented vector autoregression to construct time-varying global oil market uncertainty in a data-rich environment. Our estimates display significant independent variations from popular uncertainty proxies. We also find there is a negative role for oil market uncertainty as a determinant of oil prices.

CO150 Room E2 TOPICS IN MATHEMATICAL FINANCE AND MACHINE LEARNING Chair: Jan Vecer

C0671: **Dynamic scoring: Probabilistic model selection based on utility maximization**
Presenter: **Jan Vecer**, Charles University, Czech Republic
A novel and general approach of model selection is proposed for probability estimates that can also be applied in the time evolving setting. The basic idea is that any discrepancy between two different probability estimates opens a possibility to compare them by setting a trade on a hypothetical betting market that trades probabilities. The mechanism of this market and the behavior of agents in this market is described. An agent maximizing some utility function determines the optimal bet size for given odds. This procedure produces supply and demand functions, the size of the bet as a function of a trading probability. These functions are analytical for the choice of logarithmic and exponential utility functions. Having two probability estimates and the corresponding supply and demand functions, the trade matching these two estimates happens at the resulting equilibrium at the intersection of the supply and demand functions. It is shown that an agent using true probabilities will realize a profit when trading against any other set of probabilities.

C0836: **Detecting arbitrage in the spot foreign exchange market**
Presenter: **Stephen Taylor**, New Jersey Institute of Technology, United States
A theoretical and computational framework is proposed for the detection and identification of arbitrage opportunities among spot currency exchange rates. We obtain sufficient conditions for excluding the triangular arbitrage opportunities in a market with or without market frictions, i.e. transaction costs. Then, we propose an efficient computational approach not only to detect triangular arbitrage opportunities in real time but also to identify the combinations of currencies associated with the arbitrage. Finally, we discuss a graph theoretic formulation of the maximum arbitrage detection problem and present associated techniques used to identify the arbitrage magnitude. In numerical studies, we utilize empirical data of foreign currency exchange rates to substantiate our theoretical findings and demonstrate the efficiency of the proposed computational approach.

C0867: **Mean-field approximation of large banking network with defaults**
Presenter: **Tomoyuki Ichiba**, University of California Santa Barbara, United States
The dynamics of the cash reserves of an interconnected banking system are considered. Whenever a bank defaults (by letting its reserve reach a given threshold), each bank is impacted negatively, via an instantaneous jump on its reserve level. We also take into account the arrival of new financial institutions in the system. The underlying dynamics of such system is written in the spirit of the spiking neural network models as studied previously. We study the mean field limit of such system and focus in particular on its stationary distribution and on learning about the network structure. We also discuss the effects of network structure in the mean-field system.

C0988: **Local time, running supremum and variable annuity guarantee benefits with ratchet option**
Presenter: **Runhuan Feng**, University of Illinois at Urbana-Champaign, United States
Co-authors: Chongda Liu
Guaranteed lifetime withdrawal benefit (GLWB) with the step-up feature (a.k.a ratchet option) is among the most popular riders for variable annuity contracts. Most of the literature on the valuation problem are based on numerical solutions. A unique perspective is offered to formulate the valuation problem through the application of the local time and running supremum. Consequently, we obtain analytical solutions to risk-neutral value of the GLWB rider with roll-up feature in the waiting time and the step-up feature throughout lifetime, which lead to highly efficient algorithms for pricing and dynamic hedging of GLWB riders.

C1236: **From ratings to credit losses, in good and bad times: Correlation modeling of credit risk and macroeconomic variables**
Presenter: **Libor Pospisil**, Moody's Analytics, United States
In order to perform calculations set out in the new international accounting standards, IFRS9 and CECL, financial institutions need models that would project losses on credit portfolios under hypothetical macroeconomic scenarios. While we can relate this kind of calculation to the well-known topic of stress testing, the IFRS9 and CECL applications often have a limitation: financial institutions may only have agency ratings as the credit risk measures of their portfolios. Agency ratings are typically considered Through-the-Cycle (TTC) risk measures, not reflecting the current economic environment. Projecting losses if only rating is known therefore must include two components: accounting for how the current economic environment affects credit risk as of now, leading to Point-in-Time (PIT) view of credit, and quantifying effects of the hypothetical scenarios that describe a possible future path of the economy. We present several time series and correlation models, which allow us to project credit losses when only rating of a portfolio is known. We then relate these models to the theory of TTC and PIT credit measures, as well as to the standard stress testing methods for credit portfolios.

CO330 Room F2 ECONOMETRIC ANALYSIS OF COMMODITIES AND COMMODITY FUTURES Chair: Claudia Wellenreuther

C0235: **A model of grains prices with application to the impact of biofuels**
Presenter: **Christopher Gilbert**, Johns Hopkins University, Italy
An econometric model of world grains and oilseed (wheat, corn and soybeans) markets is reported. The model is used to address the relative importance of demand and supply shocks and the extent of their price impact. We ask whether it makes sense to think of grains as a single composite commodity or whether idiosyncratic crop-specific factors remain important. We apply the model to analyze the impact of the growth of the use of grains as biofuel feedstocks.

C0394: **A Bayesian commodity style-integration framework**
Presenter: **Nan Zhao**, Cass Business School, City, University of London, United Kingdom
Co-authors: Ana-Maria Fuertes
The literature abounds with multiple long-short investment styles designed to capture the commodity futures risk premia of which momentum, hedging pressure and term structure are the most well-known. Combining investment styles is strongly motivated a priori because by relying on a composite signal that exploits K lowly-correlated signals, the integrated-style portfolio is likely to capture a larger premium consistently over time. A key decision that an investor pursuing a style-integrated portfolio faces is how to dynamically choose the weights to allocate to each style. A Bayesian style-integration (BI) approach is developed which allows the investor explicitly to account for parameter and model uncertainty. Focusing on the allocation problem of a commodity futures investor that seeks exposure to the well-documented hedging pressure, term structure and momentum styles, the Bayesian integration is confronted with the widely-used nave equal-weight integration (EWI) and the optimized integration approach that obtains the style weights at each portfolio formation time by utility maximization. We find that the proposed BI approach yields a superior Sharpe ratio and certainty equivalent return than the EWI and OI approaches. The findings are robust to transaction costs, variants of the sophisticated OI integration, longer ranking windows, and different economic period analysis.

C0577: **Oil jump risk**
Presenter: **Craig Pirrong**, University of Houston, United States
Co-authors: Nima Ebrahimi
The aim is to evaluate (a) the predictive power of oil price jump risk premia, and (b) whether these risk premia are priced in the cross-section of stock returns. We find that upside and downside jump risk premia extracted from crude oil futures prices have considerable power to predict economic indicators including GDP growth, consumption growth, and investment. The upside jump premium also has some power to predict equity index returns and oil futures returns. Further, prior to 2011, the upside jump premium is a driver of the cross-sections of equity returns. After controlling for the oil jump risk premium, the oil variance risk premium (which has been found to be priced in some previous research) is no longer explains the cross-section of stock returns.

C0848: **Crude oil price movements and institutional traders**
Presenter: **Celso Brunetti**, Bocconi University and Federal Reserve Board, United States
Co-authors: Jeffrey Harris
The role of hedge fund, swap dealer and arbitrageur activity in the crude oil market is analyzed. Using confidential position data on institutional investors, we first analyze the linkages between trader positions and fundamentals. We find that these institutional positions reflect fundamental economic factors. Subsequently, we adopt a Markov regime-switching model with time varying probabilities and find institutional positions contribute incrementally to the probability of regime changes displaying the synchronization patterns modeled previously. Conditioning on hedge fund activity and arbitrageur activity significantly improves our probability estimates, demonstrating that institutional positions can be useful in determining whether price trends resembling bubble patterns will continue or reverse.

C1243: **Price discovery in commodity markets: On the contribution of speculators**
Presenter: **Martin Stefan**, University of Muenster, Germany
Co-authors: Martin T Bohl, Pierre Siklos, Claudia Wellenreuther
Previous literature on price discovery in commodity markets is mainly focused on whether the spot or the futures market dominates the price discovery process. Little attention, however, has been paid to the question of how the price discovery process is affected by futures speculation. Using different measures for speculation and hedging and a new price discovery metric, the present study analyzes this relationship for various agricultural commodities. The results indicate that speculative activity generally reduces the level of noise in the futures market, while increasing the relative contribution of the market to the price discovery process.

CO098 Room G2 SMALL-SAMPLE ASYMPTOTICS Chair: Benjamin Holcblat

C0271: **An extended empirical saddlepoint approximation for intractable likelihoods**
Presenter: **Matteo Fasiolo**, University of Bristol, United Kingdom
The use of simulation-based inferential approaches is widespread in computational biology and ecology. We will focus on one such approach: Synthetic Likelihood (SL). This method reduces the observed and simulated data into a set of features or summary statistics, and quantifies the discrepancy between them through a synthetic likelihood function. While requiring less tuning than some alternative approaches (such as approximate Bayesian computation), SL has the drawback of relying on the summary statistics being approximately normally distributed. We will describe how this shortcoming can be addressed by adopting a more flexible density estimator: the Extended Empirical Saddlepoint Approximation (ESA). This new density estimator is able to capture large departures from normality, while being scalable to high dimensions. This can lead to more accurate parameter estimates, relative to the Gaussian alternative.

C0437: **Limiting saddlepoint relative errors under purely Tauberian conditions**
Presenter: **Andrew Wood**, The University of Nottingham, United Kingdom
Co-authors: Ronald W Butler
Most theoretical results on the relative errors of saddlepoint approximations in the extreme tails have involved placing at least some conditions on the density/mass function. As a result, checking the validity of such conditions is problematic when density/mass functions are intractable, as is typically the case in important practical applications involving convolved, compound, and first-passage distributions as well as for MGFs that are regularly varying. We present a novel condition which ensures the existence of a positive finite limiting relative error for saddlepoint density functions. This condition, which is rather weak, is expressed entirely in terms of the moment generating function MGF, hence the description purely Tauberian. The focus will be mainly on the case in which there is a positive gamma distributional limit. We show how to check the new condition in important classes of models in this setting.

C0214: **On solutions to estimating equations and the empirical saddlepoint approximation of their distribution**
Presenter: **Fallaw Sowell**, Carnegie Mellon University, United States
Co-authors: Benjamin Holcblat
Many statistics correspond to a solution to estimating equations, so that the latter are often used to conduct inference. We study solutions to estimating equations, and the empirical saddlepoint (ESP) approximation of their distribution. When estimating equations are nonlinear, they may have multiple solutions. Under general assumptions, we prove that, for any solution, there exists an arbitrary close measurable solution, and that the distribution of the solutions corresponds to the intensity measure of a point random field. If the set of solutions has no accumulation point, we establish the global consistency and asymptotic normality of the ESP approximation. Monte-Carlo simulations and an empirical application illustrate the performance of the ESP approximation.

C0256: **Small sample proportional hazards inference, with application to mortgage prepayment**
Presenter: **John Kolassa**, Rutgers, the State University of New Jersey, United States
Proportional hazards regression techniques are often used to model event time data subject to censoring. Small samples involving discrete covariates with strong effects can lead to infinite maximum partial likelihood estimates. A methodology is presented for eliminating nuisance parameters estimated at infinity using approximate conditional inference. Conventional higher-order likelihood inference may then be applied to remaining parameter components. Techniques will be applied to models for mortgage prepayment.

C0399: **Large and moderate deviations for option pricing**
Presenter: **Antoine Jacquier**, Imperial College London, United Kingdom
Large and moderate deviations results for option prices and implied volatilities are proved for a large class of stochastic volatility models, in particular regarding their small-time and large-time behaviours. We show how to use these asymptotic estimates in order to construct efficient importance sampling estimates.

CO564 Room H2 SPECULATIVE BUBBLES Chair: Robinson Kruse-Becher

C0280: **Date-stamping multiple bubble regimes**
Presenter: **Emily Whitehouse**, Newcastle University, United Kingdom
Co-authors: Dave Harvey, Steve Leybourne
Identifying the start and end dates of explosive bubble regimes has become a prominent issue in the econometric literature. Recent research has demonstrated the advantage of a model-based minimum sum of squared residuals estimator, combined with Bayesian Information Criterion (BIC) model selection, over recursive unit root testing methods in providing accurate date estimates for a single bubble. However, in the context of multiple bubbles, a large number of models are possible, making such a model-based method unattractive to practitioners. We propose a two-step procedure for dating multiple bubbles. First, recursive unit root tests are used to identify a 'date window' in which we believe a bubble starts and ends. Second, a model-based BIC approach is used to estimate the regime change points within each window. Monte Carlo simulations highlight the effectiveness of our procedure. In addition, our method allows us to distinguish between different types of bubble behaviour (such as whether or not each bubble crashes before reverting back to a unit root process) and date these crash regimes. The advantages of our procedure over existing methods of bubble dating are shown through empirical application to financial and macroeconomic time series.

C0298: **Risk measures under explosiveness**
Presenter: **Christoph Wegener**, IPAG Business School, France
Co-authors: Dominique Guegan, Robinson Kruse-Becher, Hans-Joerg von Mettenheim
Financial asset bubbles can be characterized by periods of expansion and collapse. Expansions are often modeled as explosive processes for the asset price. Ignoring such explosiveness leads to misspecified Value-at-Risk (VaR) and related measures, e.g. Expected Shortfall. We use the definition of mildly explosive autoregression to analyze how the VaR is affected during an explosive regime and several forms of collapses. We find that the unadjusted down-side VaR is overestimated in explosive periods and also misspecified during the collapse. The form of the misspecification strongly depends on several factors: (i) horizon of the VaR forecast, (ii) duration and strength of the explosive regime (as measured by the length of the explosive subsample and the explosive root), and (iii) the nature of the collapse. These insights are demonstrated by means of a parametric model for which we establish a number of theoretical findings. The size of the effects (in terms of capital requirements) are quantified by means of an extensive Monte Carlo simulation study. We propose a correction term to be added to the VaR which accounts for the unexpected loss due to a burst. In our empirical applications, we demonstrate the merits and limits of the suggested VaR adjustments.

C0623: **Mild explosivity in recent crude oil prices**
Presenter: **Roderick McCrorie**, University of St Andrews, United Kingdom
Co-authors: Isabel Figuerola-Ferretti, Ioannis Paraskevopoulos
An analysis of oil prices during and in the aftermath of the Global Financial Crisis is provided. The mildly explosive/multiple bubbles testing strategy is used to assess whether there were any price departures from an underlying stochastic trend and whether any such departures can be explained by fundamentals or other proxy variables. The test dates two significant time periods in both Brent and WTI nominal and real front-month futures prices: a mildly explosive episode during the 2007-08 spike, prior to the peak of the Global Financial Crisis; and a shorter, negative such episode during the recent price decline, whose commencement is dated around a key OPEC meeting in November 2014. Evidence using other commodity prices points to explanatory factors beyond commodity markets. A demand-side fundamental is found to be decisive in the episode in mid-2008; excess speculation is not. U.S. shale oil production, although contributing to the post-June 2014 price decline, is not decisive. We find no evidence the CBOE Volatility Index (VIX) decisively affected oil price levels during the sample period. The results are compared and contrasted with those obtained previously via a forecasting approach based on a structural vector autoregressive model without financial variables. The results offer evidence based on formal statistical testing to resolve a number of recent controversies in the oil price literature.

C0491: **Exuberance: Sentiments driven buoyancy**
Presenter: **Anurag Banerjee**, Durham University, United Kingdom
Co-authors: Guillaume Chevillon
The expected growth in consumption is based on consumer sentiment, which can produce temporary bubbles in the asset markets. We present a model based on by consumption CAPM, where prices are driven by expectations in consumption growth. Our econometric model is flexible random coefficient model that allows multiple bubbles impact of sentiments on asset price dynamics: "buoyancy" driven by local optimism. We estimate models with candidate "sentiments" variables for forecasting asset returns. We also propose a test under the null hypothesis that consumer "sentiments" do not matter and the asset prices are simple driven by a random walk.

C1007: **Monitoring asset price bubbles in real-time in the presence of nonstationary volatility**
Presenter: **Matei Demetrescu**, University of Kiel, Germany
A real-time statistical monitoring of speculative bubbles is proposed which take into account the inherent feature of nonstationarity volatility in the innovation process of the time series of interest. Nonstationary volatility is a common but often overlooked occurrence in financial and macroeconomic time series data, such that abstracting from this process results in overdetection of bubbles. We compute critical values from a boundary value function that changes at each point in time in such a way that, asymptotically, volatility variations don't matter. We showcase proposed testing strategy in detecting real-time speculative bubbles in the valuation ratios of the S&P 500 during the dot-com bubble as well as in bitcoin prices, yielding bubble detection rates that are robust to structural breaks.

CO605 Room I2 SPECIFICATION TESTING IN FINANCIAL ECONOMETRICS Chair: Dominik Wied

C0187: **Weighted CUSUM tests for the stability of the correlation structure of multivariate volatility models**
Presenter: **Marco Barassi**, University of Birmingham, United Kingdom
Co-authors: Brendan McCabe, Yuqian Zhao
A weighted approximation of semi-parametric CUSUM statistics is proposed for the stability of correlation structures of multivariate volatility models. These weighted CUSUM tests are constructed so as to enhance their power in either ends of a sample. Since for CUSUM-type change-point tests the usual Sup-statistics are calculated over a trimmed sample, this renders impossible to detect change points in the trimmed proportion of the sample as well as in the vicinity of it. As such, these tests tend to lose the power when breaks occur close to either ends of the sample ($T-t^* \to \infty$ and $t^* \to \infty$, T is the sample size and $1 \leq t^* \leq \infty$ is the break location). Our tests overcome this type of issues using two types of weight functions, namely $q_1(u,\alpha) = (u\cdot(1-u))^{\alpha}, \quad 0 \leq u \leq 1$ and $q_2(u,\alpha) = (u\cdot(1-u)\cdot\log\log\frac{1}{u\cdot(1-u)})^{\alpha}, \quad 0 \leq u \leq 1$, where $0 < \alpha < 1/2$ is a self-selected parameter. We derive the limiting distribution of the proposed tests under the null and assess their performance by means of Monte Carlo methods. The results obtained suggest that weighted CUSUM tests with weight function $[u\cdot(1-u)]^{\alpha}$ $(0 < \alpha < 1/2)$ exhibit better performance. As an application, we use both standard semi-parametric CUSUM and weighted CUSUM tests to detect harmful events in the U.S. equity market in the years 2014, 2015 and 2016.

C0186: **A monitoring procedure for detecting structural breaks in factor copula models**
Presenter: **Florian Stark**, University of Cologne, Germany
Co-authors: Hans Manner, Dominik Wied
A new monitoring procedure based on moving sums (MOSUM) is proposed for detecting single or multiple structural breaks in factor copula models. The test compares parameter estimates from a rolling window to those from a historical data set and analyzes the behavior under the null hypothesis of no parameter change. The case of multiple breaks is also treated. In the model, the joint copula is given by the copula of random variables which arise from a factor model. This is particularly useful for analyzing data with high dimensions. Parameters are estimated with the simulated method of moments (SMM). We analyze the behavior of the monitoring procedure in Monte Carlo simulations and a real data application. We consider an online procedure for predicting the day-ahead value-at-risk based on the suggested monitoring procedure.

C0253: **A non-parametric cusum-type test for testing relevant change in copulas**
Presenter: **Tim Kutzker**, University of Cologne, Germany
Co-authors: Florian Stark, Dominik Wied
A new non-parametric test is proposed for detecting relevant breaks in copula functions. The hypothesis is of the form $H_0 : ||C_1(u) - C_2(u)|| \leq \Delta$ versus $H_1 : ||C_1(u) - C_2(u)|| > \Delta$, where Δ is an adjustable size to allow for difference in the copulas C_1 and C_2. The test is based on suitable differences between empirical copulas and it is shown that the test statistic is asymptotically normally distributed when the marginals are known. Critical values are determined by bootstrap approximations. In the case of known marginals, it is sufficient to bootstrap the variance, while we bootstrap the whole testing process in the case of unknown marginals. We analyze the behavior of the test in Monte Carlo simulations and a real data application.

C0571: **Detecting relevant changes in the mean of non-stationary processes: A mass excess approach**
Presenter: **Weichi Wu**, Ruhr University Bochum, Germany
Co-authors: Holger Dette
The focus in on the problem of testing if a sequence of means $(\mu_t)_{t=1,\ldots,n}$ of a non-stationary time series $(X_t)_{t=1,\ldots,n}$ is stable in the sense that the difference of the means μ_1 and μ_t between the initial time $t = 1$ and any other time is smaller than a given threshold, that is $|\mu_1 - \mu_t| \leq c$ for all $t = 1,\ldots,n$. A test for hypotheses of this type is developed using a bias corrected monotone rearranged local estimator and asymptotic normality of the corresponding test statistic is established. As the asymptotic variance depends on the location of the critical roots of the equation $|\mu_1 - \mu_t| = c$, a new bootstrap procedure is proposed to obtain critical values and its consistency is established. As a consequence, we are able to quantitatively

describe relevant deviations of a non-stationary sequence from its initial value. The results are illustrated by means of a simulation study and by analyzing data examples.

C1152: **Monitoring cointegration in a system of homogeneous cointegrating regressions**
Presenter: **Etienne Theising**, University of Cologne, Germany
Co-authors: Martin Wagner, Dominik Wied
The aim is to extend a previous procedure for detecting a structural change in a system of homogeneous cointegrating regressions. Therefore, we construct test statistics based on the residuals of fully modified OLS estimators to account for error serial correlation and regressor endogeneity and to obtain nuisance parameter free limiting distributions. In particular, we consider the pooled FM-OLS estimator for homogeneous panel cointegration. We, however, focus on the finite N case and consider only large T asymptotics. We examine four different detectors along with " self-normalized " versions of those and additionally compare them to their equivalents based on residuals of the classical FM-OLS estimator applied to each cointegrating regression separately to assess the relative performance in terms of small size distortion under the null of no structural change, and power and detection time under various alternatives.

CO064 Room M2 MACROECONOMIC POLICIES AND MACROECONOMETRICS Chair: Etsuro Shioji

C0478: **Infrastructure investment news and business cycles in Japan: Evidence from the VAR with an external instrument**
Presenter: **Etsuro Shioji**, Hitotsubashi, Japan
The Trump Administration's push for a massive infrastructure spending has renewed interest in the macroeconomic effects of public investment. Past efforts to estimate these effects have been marred by the "fiscal foresight problem": as most fiscal actions are announced well before their implementation, the traditional identification scheme which relies on the timing and the amount of actual spending could be misleading. The difficulty by utilizing the VAR with an external instrument is overcome. My instrument is an indicator of the private sector's expectations about future policies; it utilizes cross sectional variations across individual construction firms' stock returns, on the days that important news about policies arrive. Results indicate that an anticipated public investment shock has a significantly positive effect on GDP of Japan.

C0483: **The new area-wide model II: An updated version of the ECB's micro-founded model for forecasting and policy analysis**
Presenter: **Gunter Coenen**, European Central Bank, Germany
A detailed exposition of an updated version of the ECB's New Area-Wide Model (NAWM) of the euro area is provided. Besides several changes to its original specification reflecting the practical uses of the model in the policy process over the past few years, the updated version of the NAWM includes a rich financial intermediary sector with the threefold aim of (i) accounting for a genuine role of financial frictions and the propagation of financial disturbances originating in the financial sector, (ii) capturing the prominent role of bank lending rates in the transmission of monetary policy in the euro area, and (iii) providing a structural framework useable for assessing the macroeconomic impact of the ECB's large-scale asset purchases conducted in recent years.

C0681: **Empirical analysis on the effects of Japanese fiscal policy under the effective lower bound**
Presenter: **Hiroshi Morita**, Hosei University, Japan
For the Japanese economy, we examine whether the effectiveness of fiscal policy is amplified under the effective lower bound of nominal interest rates using a non-linear time-varying parameter vector autoregression model. The model adopts the Tobit specification on the interest rate equation in order to isolate the role of monetary policy stance on the effectiveness of government spending shock by comparing the two values of fiscal multiplier computed based on the actual and implied interest rate, respectively. We find that government spending shock greatly reduces the real interest rate under the zero interest rate policy, so that the effective lower bound of short-term interest contributes to enhance the effect of fiscal policy on output.

C0737: **Identifying factor-augmented vector autoregression models via changes in shock variances**
Presenter: **Yohei Yamamoto**, Hitotsubashi University, Japan
The aim is to propose a new method for the structural identification of factor-augmented vector autoregression models based on changes in the unconditional shock variances that occur on a historical date. The proposed method can incorporate both observed and unobserved factors in the structural vector autoregression system and it allows the contemporaneous matrix to be fully unrestricted. We derive the asymptotic distribution of the impulse response estimator and consider a bootstrap inference method. Monte Carlo experiments show that the proposed method is robust to the misspecification of the contemporaneous matrix unlike the existing methods. Both the asymptotic and bootstrap methods obtain a satisfactory coverage rate when the shock of an observed factor is studied, although the latter is more accurate when the shock of an unobserved factor is considered. An empirical illustration based on data used previously provides similar point estimates and somewhat wider confidence intervals.

C0852: **Frequency-wise causality analysis in infinite order vector autoregressive processes**
Presenter: **Mototsugu Shintani**, University of Tokyo, Japan
Co-authors: Ryo Kinoshita, Kosuke Oya
The asymptotic properties of a frequency-domain causality measure are derived by using the vector autoregressive model of infinite order and proposes a test for causality at a particular frequency. Simulation results confirm that our procedure works well with sample size typically available in practice. We illustrate the usefulness of our method via an application to financial data.

CO176 Room N2 TIME SERIES ECONOMETRICS I Chair: Antonio Montanes

C0627: **A new approach to dating the reference cycle**
Presenter: **Lola Gadea**, University of Valencia, Spain
A new approach is proposed to the analysis of the reference cycle turning points. Each individual pair of specific peaks and troughs from a set of disaggregated coincident economic indicators is viewed as a realization of a mixture of an unspecified number of separate bivariate Gaussian distributions whose different means are the reference turning points. These dates break the sample into separate reference cycles, whose phase shifts are modeled by a hidden Markov chain. The transition probability matrix is constrained so that the specification is equivalent to a multiple change-point model. Bayesian estimation of finite Markov mixture modeling techniques is suggested to estimate the model. Several Monte Carlo experiments are used to show the accuracy of the model to date reference cycles that suffer from short phases, uncertain turning points, small samples and asymmetric cycles. In the empirical section, we show the high performance of our approach to identifying the US reference cycle, with little difference from the timing of the turning point dates established by the NBER.

C1074: **Panel data cointegration analysis with structural instabilities**
Presenter: **Josep Lluis Carrion-i-Silvestre**, Universitat de Barcelona, Spain
Co-authors: Anindya Banerjee
Spurious regression analysis in panel data when time series are cross-section dependent is analyzed. The set-up is general enough to include multiple structural breaks that can affect the deterministic component and the common factor component. We show that consistent estimation of

the long-run average parameter is possible once cross-section dependence is controlled using cross-section averages in the spirit of the common correlated effects approach. This result is used to design individual and panel cointegration test statistics that accommodate the presence of structural breaks that can induce parameter instabilities on the deterministic component, the cointegration vector and the common factor loadings.

C1054: **The dynamics of external reaction functions: The role of financial globalization and risk aversion**
Presenter: **Juan Sapena**, Catholic University of Valencia, Spain
Co-authors: Mariam Camarero, Cecilio Tamarit

The aim is to reexamine the issue of the external sustainability in a dynamic framework. The emergence of large and persistent external imbalances is one of the most crucial events in international finance. We depart from standard empirical tests of intertemporal sustainability conditions, by following the approach initiated by Henning Bohn, that proposes the estimation of a linear reaction function for the trade balance to a measure of external debt position, such as net foreign liabilities. To estimate the fiscal reaction function, we extend Bohn's approach into a time-varying external reaction function for a group of 17 countries panel, paying special attention to Eurozone members, but also including some relevant OECD countries as well for the period 1970-2016. The main advantage of our empirical approach is that it captures the dynamics of the external reaction function, in search of main sources of heterogeneity among EMU countries. In particular, we explore the incidence of financial globalization and global risk aversion on the dynamics of the model for the selected countries along the period considered. For the estimation of the model, we employ a State-Space framework for modeling panel time series, which extends the simple framework generally employed at the literature, into a panel-data time-varying parameters framework, combining both fixed (either common and country-specific) and varying components.

C1129: **Is there just a climate change?**
Presenter: **Antonio Montanes**, University of Zaragoza, Spain

The evolution of temperature across worldwide is analyzed. The use of a previous methodology leads us to reject the convergence null hypothesis. Rather, we can find the presence of several convergence clubs. Thus, we can conclude the existence of different climate changes instead of the presence of a single one.

C1537: **Multiple long-run equilibria: Threshold cointegration**
Presenter: **Jesus Gonzalo Munoz**, Universidad Carlos III de Madrid, Spain
Co-authors: Jun Yi Peng

The extension of linear cointegration to a non-linear framework has always considered a single long-run equilibrium: non-linearity comes from the adjustment toward it. Non-linear cointegration with multiple long-run equilibria is analyzed via threshold cointegration. It develops the testing (non-linear cointegration), inference (threshold cointegration) and representation theorem (quasi ECM). Several applications are shown where cointegration with different equilibria is not rejected while standard linear cointegration is.

EO326 Room Aula 4 MODEL SELECTION Chair: Keith Knight

E0195: **Debiasing the debaised lasso with bootstrap**
Presenter: **Sai Li**, University of Pennsylvania, United States
It is proven that under proper conditions, bootstrap can further correct the bias of the debiased lasso estimator for statistical inference of low-dimensional parameters in high-dimensional linear regression. We prove that the required sample size for inference with bootstrapped debiased lasso, which involves the number of small coefficients, can be of smaller order than the existing ones for the debiased lasso. Therefore, our results reveal the benefits of having strong signals in high-dimensional inference. Our theory is supported by results of simulation experiments, which compare coverage probabilities and lengths of confidence intervals with and without bootstrap, with and without debiasing.

E0226: **On the model selection properties and uniqueness of the lasso**
Presenter: **Ulrike Schneider**, Vienna University of Technology, Austria
Co-authors: Karl Ewald
The model selection properties of the lasso estimator are investigated in finite samples with no conditions on the regressor matrix X. We show that which covariates the lasso estimator may potentially choose in high dimensions (where the number of explanatory variables p exceeds sample size n) depends only on X and the given penalization weights. This set of potential covariates can be determined through a geometric condition on X and may be small enough (less than or equal to n in cardinality), so that the lasso estimator acts as a low-dimensional procedure also in high dimensions. Related to the geometric conditions in our considerations, we also provide a necessary and sufficient condition for uniqueness of the lasso solutions.

E1077: **Sparse portfolio selection via the sorted L_1 norm**
Presenter: **Malgorzata Bogdan**, University of Wroclaw, Poland
Co-authors: Philipp Johannes Kremer, Sandra Paterlini, Sangkyun Lee
A financial portfolio optimization framework is introduced that allows us to automatically select the relevant assets and estimate their weights by relying on a sorted L_1-Norm penalization, henceforth SLOPE. Our approach is able to group constituents with similar influence on the overall portfolio risk. We show that depending on the choice of the penalty sequence, SLOPE can span the entire set of optimal portfolios on the risk-diversification frontier, from minimum variance to the equally weighted. To solve the optimization problem, we develop a new efficient algorithm, based on the alternating direction method of multipliers. Our empirical analysis shows that SLOPE yields optimal portfolios with good out-of-sample risk and return performance properties, by reducing the overall turnover through more stable asset weight estimates. Moreover, using the automatic grouping property of SLOPE, new portfolio strategies, such as sparse equally weighted SLOPE-EW portfolio, can be developed to exploit the data-driven detected similarities across assets.

E1408: **Model selection with missing data**
Presenter: **Sylvain Sardy**, University of Geneva, Switzerland
Model selection in high-dimensional linear models is considered when some entries of the regression matrix are missing. The goal is to be the least affected by the missing values so as to achieve high true positive rate and low false discovery rate in the search for the true underlying covariates.

EO240 Room Aula 5 RECENT ADVANCES IN COMPLEX DATA ANALYSIS Chair: Alejandro Murua

E1461: **Bayesian lasso time-course data clustering**
Presenter: **Alejandro Murua**, University of Montreal, Canada
Co-authors: Folly Adjogou, Wolfgang Raffelsberger
A flexible model is developed for the analysis and clustering of time-course or longitudinal data. The model combines functional analysis and model-based clustering. The functional framework is used to model time-course data. Principal functional components are described by score coefficients which embed the curves in a much lower-dimensional space. Model based clustering is performed on the score space, thus avoiding the curse of dimensionality in the curves' space. The model is embedded into a Bayesian framework. We first develop an approximation of the marginal log-likelihood MLL that allows us to perform a MLL based model selection. We then developed a Bayesian version of the lasso and elastic-net penalty in order to render the model selection step more efficient. The number of clusters as well as the dimension of the score space are determined via this Bayesian lasso penalty model. We show some applications to the analysis of gene-expression data.

E1564: **Approximate confidence distribution computing: An effective likelihood-free method with statistical guarantees**
Presenter: **Wentao Li**, Newcastle University, United Kingdom
Co-authors: Suzanne Thornton, Min-ge Xie
Approximate Bayesian computing has grown increasingly popular for intractable likelihood model. However, complications arise in the theoretical justification for Bayesian inference conducted from this method with a non-sufficient summary statistic. We seek to re-frame approximate Bayesian computing within a frequentist context and justify its performance by standards set on the frequency coverage rate. In doing so, we develop a new computational technique called approximate confidence distribution computing, yielding theoretical support for the use of non-sufficient summary statistics in likelihood-free methods. Furthermore, we demonstrate that approximate confidence distribution computing extends the scope of approximate Bayesian computing to include data-dependent priors without damaging the inferential integrity. This data-dependent prior can be viewed as an initial 'distribution estimate' of the target parameter which is updated with the results of the approximate confidence distribution computing method. A general strategy for constructing an appropriate data-dependent prior is also discussed and is shown to often increase the computing speed while maintaining statistical guarantees. We supplement the theory with simulation studies illustrating the benefits of the confidence distribution method, namely the potential for broader applications than the Bayesian method and the increased computing speed compared to approximate Bayesian computing.

E1574: **A Cramer moderate deviation theorem for general self-normalized sums**
Presenter: **Jiasheng Shi**, The Chinese University of Hong Kong, Hong Kong
Co-authors: Qi-Man Shao, Lan Gao
Asymptotic theory for self-normalized sums has been well studied in the past two decades. The aim is to focus on a general self-normalized sums $\frac{\sum_{i=1}^n X_i}{\sqrt{\sum_{i=1}^n Y_i^2}}$, where (X_i, Y_i), for i from 1 to n, are independent random vectors. The Cramer type moderate deviation theorem is obtained under optimal moment condition. Applications to self-normalized dependent random variables, as in the case of longitudinal data and beta mixing time series will also be discussed.

E1576: **On higher order moment and cumulant estimation**
Presenter: **Chunxue Li**, The Chinese University of Hong Kong, Hong Kong
Co-authors: Chun Yip Yau, Kun Chen, Lok Hang Chan, Chung Wang Wong
Moments and cumulants are fundamental in statistical analysis. Particularly, many models designed for longitudinal data require estimation for higher order moments. A natural and popular approach to moment and cumulant estimation is based on sample average. However, these sample average estimators may perform poorly. We derive uniformly minimum variance unbiased estimators for raw moments, centered moments, and cumulants of any order. Explicit formulas of the estimators are obtained for a number of common distributions. Extensive simulation studies demonstrate that the proposed estimators can perform much better than the corresponding sample average estimators.

EO418 Room Aula B LARGE-SCALE AND COMPLEX DATA ANALYSIS Chair: George Michailidis

E0825: **Anomaly detection in static networks using egonets**
Presenter: **Srijan Sengupta**, Virginia Tech, United States
Network data has rapidly emerged as an important and active area of statistical methodology. We consider the problem of anomaly detection in networks. Given a large background network, we want to detect whether there is a small anomalous subgraph present in the network, and if such a subgraph is present, we want to identify nodes constitute the subgraph. We propose an inferential tool based on egonets to answer this question. The proposed method is simple and easily extends to several network models, while being computationally efficient and naturally amenable to parallel computing. Using synthetic networks, we demonstrate that the egonet method works well under a wide variety of network models. We obtain interesting results by applying the method on several well-studied benchmark datasets.

E1169: **Geometric inference in admixture models**
Presenter: **Yuekai Sun**, University of Michigan, United States
A class of admixture models suitable for a variety of data types is considered. Fast and provably accurate inference algorithms are developed by accounting for the model's convex geometry and low dimensional simplicial structure. Thanks to the strong connection to the Voronoi tessellation and properties of the Dirichlet distribution, the proposed inference algorithm is shown to achieve consistency and strong error bound guarantees on a range of model settings and data distributions. The effectiveness of our model and the learning algorithm is demonstrated by simulations and by analyses of text and financial data.

E1317: **A Bayesian framework for joint estimation of multiple networks**
Presenter: **George Michailidis**, University of Florida, United States
A novel Bayesian approach is developed for joint estimation of multiple graphical models from a dictionary of possible/suggested ones. This problem arises in many applications, such as understanding co-expression networks from high-dimensional omics data obtained from different biological conditions or disease subtypes. We pursue a pseudo-likelihood based approach which provides robustness and computational efficiency. We establish strong posterior consistency and illustrate the efficacy of the proposed approach on both synthetic and real data.

E1191: **Statistics of stochastic gradient descent: Stability, efficiency, and inference**
Presenter: **Panagiotis Toulis**, University of Chicago, United States
Stochastic gradient descent (SGD) is remarkably multi-faceted: for machine learners it is a powerful optimization method, but for statisticians it is a method for iterative estimation. While several important results are known for optimization properties of SGD, surprisingly little is known about its statistical properties. We will review recent results on doing statistics with SGD, which include analytic formulas for the asymptotic covariance matrix of SGD-based estimators and a numerically stable variant of SGD with implicit updates. Together these results open up the possibility of doing principled statistical analysis with SGD, including classical inference and hypothesis testing. Specifically about inference, we present current work showing that with appropriate selection of the learning rate the asymptotic covariance matrix of SGD is isotropic and parameter-free. As such, some SGD-based estimators can be easily transformed into pivotal quantities, which substantially simplify inference. This is a unique and remarkable property of SGD, even compared to popular estimation methods favored by statisticians, such as maximum likelihood, highlighting the untapped potential of SGD for fast and principled estimation with large data sets.

EO026 Room Aula Magna ADVANCES IN HIGH-DIMENSIONAL AND FUNCTIONAL TIME SERIES ANALYSIS Chair: Dominik Liebl

E0854: **Dynamic function-on-scalars regression**
Presenter: **Daniel Kowal**, Rice University, United States
A modeling framework is developed for dynamic function-on-scalars regression, in which a time series of functional data is regressed on a time series of scalar predictors. The regression coefficient function for each predictor is allowed to be dynamic, which is essential for applications where the association between predictors and a (functional) response is time-varying. For greater modeling flexibility, we design a nonparametric reduced-rank functional data model with an unknown functional basis expansion, which is both data-adaptive and, unlike most existing methods, modeled as unknown for appropriate uncertainty quantification. Within a Bayesian framework, we introduce shrinkage priors that simultaneously (i) regularize time-varying regression coefficient functions to be locally static, (ii) effectively remove unimportant predictor variables from the model, and (iii) reduce sensitivity to the selected rank of the model. A simulation analysis confirms the importance of these shrinkage priors, with substantial improvements over existing alternatives. We develop a novel projection-based Gibbs sampling algorithm, which offers unrivaled computational scalability for fully Bayesian functional regression. We apply the proposed methodology (i) to characterize the effects of demographic predictors on age-specific fertility rates in South and Southeast Asia, and (ii) to analyze the time-varying impact of macroeconomic variables on the U.S. yield curve.

E0932: **Spectral clustering of functional time series**
Presenter: **Anne van Delft**, Ruhr University Bochum, Germany
Co-authors: Holger Dette
Due to the surge of data storage techniques, the need for the development of appropriate techniques to identify patterns and to extract knowledge from the resulting enormous data sets, which can be viewed as collections of dependent functional data, is of increasing interest in many scientific areas. We develop such a technique and introduce a spectral clustering algorithm for time series of functional data. First, we propose a measure to test equality of the spectral density operators of a collection of functional time series. The functional time series are neither supposed independent nor stationary. The measure is based on the aggregation of Hilbert-Schmidt differences of the individual time-varying spectral density operators. Under fairly general conditions, the asymptotic properties of the corresponding estimator are derived and asymptotic normality is established. The introduced statistic lends itself naturally to quantify (dis)-similarity between functional time series, which we subsequently exploit in order to build a spectral clustering algorithm. Our algorithm is the first of its kind in the analysis of nonstationary (functional) time series and enables to discover particular patterns by grouping together 'similar' series into clusters, thereby reducing the complexity of the analysis considerably. The algorithm is simple to implement and computationally feasible.

E1015: **High-dimensional curve estimation in time-varying models**
Presenter: **Stefan Richter**, Heidelberg University, Germany
Co-authors: Jonas Krampe, Jens-Peter Kreiss, Efstathios Paparoditis
Curve estimation for locally stationary processes is considered. We allow the parameter curves describing the non-stationarity to be high-dimensional. Under usual sparsity assumptions we derive concentration inequalities for a likelihood-based lasso estimator. Furthermore, we propose a desparsified lasso approach which allows for a Gaussian limit distribution and hypotheses testing via a bootstrap procedure. The finite-sample behavior of the estimation procedure is analyzed in simulations with tvVAR and tvARCH processes.

E1400: **Change point analysis with functional time series**
Presenter: **Gregory Rice**, University of Waterloo, Canada
Co-authors: Alexander Aue
Methods are considered for detecting and dating changes in both the level and variability of a sequence of curves or functional data objects. Regarding level shifts, we propose a new detection and dating procedure that is "fully functional", in the sense that it does not rely on dimension reduction techniques. To test for changes in variability, we consider methods based on measuring the fluctuations of eigenvalues of the empirical covariance operator. A thorough asymptotic theory is developed for each procedure that highlights their relative strengths and weaknesses when compared to existing methods. An application to annual temperature curves illustrates the practical relevance of the proposed methods.

EO629 Room Sala Convegni STATISTICS IN SPORTS: SOME RECENT DEVELOPMENTS Chair: Xin Liu

E0241: **Positional value in soccer: Expected league points added above replacement**
Presenter: **Xin Liu**, University of Pittsburgh, United States
Co-authors: Konstantinos Pelechrinis, Wayne Winston
Soccer is undeniably the most popular sport world-wide, but at the same time it is one of the least quantified. While there are many reasons for this, one of the main is the difficulty to explicitly quantify the contribution of every player on the field to his team chances of winning. For example, successful advanced metrics such as the (adjusted) +/- that allows for division of credit among a basketball team's players (and ultimately to obtain a view of the wins contributed by a player), fail to work in soccer due to severe co-linearities (i.e., the same players being on the field for the majority of the time). We take a first step towards developing metrics that can estimate the contribution of a soccer player to his team's winning chances. In particular, using data from (i) approximately 20,000 games from 11 European leagues for 8 seasons, as well as, (ii) player ratings from FIFA, we estimate through a Skellam regression model the importance of every line in winning a soccer game. We consequently translate the model to expected league points added (per game) above a replacement player method. This model can be used as a guide for contracts' monetary value decisions. For example, using market value data for approximately 10,000 players we further identify that currently the market clearly under-values defensive line players relative to goalkeepers. Finally, we discuss how this model can be significantly enhanced using optical tracking data.

E0304: **Dynamic modeling of player movement in American football**
Presenter: **Karl Pazdernik**, Pacific Northwest National Laboratory, United States
American football is a game of inches. The offense attempts to gain separation and yardage, while the defense works to collectively and continuously limit these distances. To truly understand the complex spatiotemporal patterns of offensive separation, identification of defensive coverage is necessary. The aim is to outline a novel methodology used to estimate the probability that each defender is tracking each offensive player at predetermined intervals of time within a play using a hidden Markov model. Within each defensive assignment, coverage types also exist. A defender may be in attack mode, in more of a surveillance motion, or may struggle to maintain proper coverage, trailing their assignment. We use a secondary group of hidden states to differentiate between these three behavioral patterns. From these estimated probabilities, unique summary statistics are possible. We can now quantify previously unmeasured statistics such as the amount of attention an offense player receives, the amount of separation a route runner can obtain, a defender's instincts regarding to their ability to diagnose a play, a defender's ability to recover when beaten, and the degree to which a defense attacks the ball carrier. For illustration, both simulated data and data from NFL games obtained through the All-22 game film are used.

E0668: **Martingale testing of football betting odds**
Presenter: **Mark Richard**, Frankfurt School of Finance and Management, Germany
Co-authors: Jan Vecer
A novel approach for martingale testing is presented. The new approach is applied in the situation of football betting odds. The betting odds are the conditional expectation of the ultimate outcome and as such, they should follow a martingale evolution. We propose to use regression analysis to check for the martingale property and apply it on Betfair data from the English Premier League 2016/17 sampled with one-minute frequency during the in-play phase of the game.

E1704: **Towards a comprehensive data-driven evaluation of soccer players performance**
Presenter: **Luca Pappalardo**, ISTI-CNR, Italy
The problem of evaluating the performance of soccer players is attracting the interest of many companies, websites, and the scientific community, thanks to the availability of massive data capturing many events generated during a game (e.g., tackles, passes, shots, etc.). Existing approaches are mainly mono-dimensional, in the sense that they propose a single metric capable of capturing just a single aspect of soccer performance. Recently, the PlayeRank algorithm has been proposed as a data-driven framework offering a principled multi-dimensional and role-aware evaluation of the performance of soccer players. We will show how to validate such a framework through an extensive experimental analysis advised by soccer experts, based on a massive dataset of millions of events pertaining five seasons of the prominent eleven soccer leagues in the world. Finally, we show how the PlayeRank framework can be used to characterize the typical performance of players and to predict their future performance on the basis of their performance history.

EO460 Room A1 STATISTICAL METHODS IN RADIATION RESEARCH Chair: Manuel Higueras

E0284: **Quasi-Poisson regression models for radiation dose estimation from biomarkers**
Presenter: **Jochen Einbeck**, Durham University, United Kingdom
Poisson regression models have a long tradition in the construction of dose-response calibration curves from count-data valued biomarkers. For instance, the current 'gold-standard' in radiation biodosimetry, based on dicentric chromosomes, usually adheres well to the equidispersion assumption of the Poisson distribution. However, there do exist several alternative modern biomarkers which allow for considerably quicker and cheaper analysis than the dicentric one, but often these come at the price of considerable overdispersion, for instance due to inter-individual variation. This holds particularly for the gamma-H2AX assay, a protein-based biomarker which makes use of counts of fluorescent dots produced by the H2AX histone following radiation-induced double-strand breaks. We illustrate the quasi-Poisson model in the context of the gamma-H2AX assay, and show how it can be used to quantify the uncertainty of doses estimated through this biomarker. Finally, the possibility of applying this procedure onto certain cytogenetic biomarkers which feature considerable overdispersion, such as micronuclei, is briefly discussed.

E0381: **Towards a new R package for fitting Poisson linear excess relative risk models**
Presenter: **Manuel Higueras**, Basque Center for Applied Mathematics, Spain

The excess relative risk (ERR) represents the elevated rate of disease (e.g., cancer) per unit of exposure (e.g., ionizing radiation). Relative risk models are typically applied in radiation epidemiology follow-up studies, for instance those which analyze the risk of leukemia or brain tumor in pediatric patients who have been examined with computed tomography scans. It is proposed the fitting of general Poisson linear relative risk models in R as an alternative of EPICURE's AMFIT module, which is the gold standard in radiation epidemiology practice. This fitting is performed by means of the maxLik R package and an efficient implementation of the gradient of the log-likelihood with respect the parameters. These Poisson linear ERR models require data in person-years format with time-depending and cumulative variables, and they can be built in R by means of the pyears function at the survival package, as an alternative of EPICURE's DATAB module. The aim is to join these two tools, person-years table builder and Poisson linear ERR model fitter, in an R package to give a free license and open source alternative to EPICURE, a very specialized proprietary software.

E0766: **A Bayesian hierarchical approach to account for shared and unshared exposure uncertainty in radiation epidemiology**
Presenter: **Sophie Ancelet**, Institut de Radioprotection et de Surete Nucleaire (IRSN), France
Co-authors: Sabine Hoffman, Chantal Guihenneuc

Measurement error is ubiquitous and represents an important source of uncertainty in radiation epidemiology. When not or poorly accounted for, it can lead to biased risk estimates and to a distortion of the exposure-response relationship. One of the main reasons why measurement error is rarely accounted for is that classical methods lack the flexibility to account for complex patterns of exposure uncertainty. In occupational cohort studies, for instance, the type and magnitude of error can change over time depending on the methods of exposure assessment. Moreover, methods of group-level exposure estimation may give rise to errors which are shared between workers belonging to the same group or shared within workers. First, a simulation study is conducted to compare the effects of shared and unshared errors on risk estimation and shape of the exposure-response curve in proportional hazards models. Then, as a flexible framework to deal with complex error structures, several Bayesian hierarchical models are proposed to obtain corrected risk estimates on the association between exposure to radon and lung cancer mortality in the French cohort of uranium miners. The importance of making a careful characterization of shared and unshared errors to account for exposure uncertainty in risk estimates is highlighted.

E0287: **How to detect zero inflation in biological dosimetry data: An exact test for the Poisson distribution**
Presenter: **Amanda Fernandez-Fontelo**, Universitat Autonoma de Barcelona, Spain
Co-authors: Pedro Puig, Elizabeth Ainsbury, Manuel Higueras

The goal of biological dosimetry is to estimate the absorbed ionising radiation dose by an overexposed individual using chromosomal damage. When radiation occurs, damage in DNA is randomly distributed between cells which may be unrepaired to form aberrations (dicentrics). The radiation dose received by an individual is estimated building dose-response-calibration curves through the number of scored dicentrics when human body cells are exposed to radiation. Being exposed to whole (WBI) or partial body irradiation (PBI) determines how to proceed to estimate the absorbed dose. The Manual of the IAEA proposes the u-test to decide whether the irradiation is WBI or PBI. However, this test can sometimes be inappropriate. We propose to use an exact goodness-of-fit test for the Poisson distribution based on the occupational problems. This test allows us to analyze the zero-inflation/deflation of the data, being able to detect PBI. It can be seen as a complement of the u-test, being useful when the dispersion is insignificant, but the number of zeros is anomalous. Data coming from accidents are analyzed. An R Shinny app is presented which performs this and other zero-inflation tests under the Poisson assumption.

EO290 Room Aula A EMPIRICAL PROCESSES AND APPLICATIONS Chair: Henryk Zaehle

E0379: **Some results about the strong approximations of the empirical processes**
Presenter: **Salim Bouzebda**, Universite de Technologie de Compiegne, France

Some results about the strong approximations of the multivariate copulas processes and the multivariate empirical process are presented. We consider an extension of the p-fold multivariate empirical processes. We finally discuss an application about the change point problems.

E0616: **Inference for local distributions at high sampling frequencies: A bootstrap approach**
Presenter: **Ulrich Hounyo**, University at Albany, SUNY, United States

Inference for the local innovations of Ito semimartingales is studied. Specifically, we construct a resampling procedure for the empirical CDF of high-frequency innovations that have been standardized using a nonparametric estimate of its stochastic scale (volatility) and truncated to rid the effect of large jumps. Our locally dependent wild bootstrap (LDWB) accommodate issues related to the stochastic scale and jumps as well as account for a special block-wise dependence structure induced by sampling errors. We show that the LDWB replicates first and second-order limit theory from the usual empirical process and the stochastic scale estimate, respectively, as well as an asymptotic bias. Moreover, we design the LDWB sufficiently general to establish asymptotic equivalence between it and a nonparametric local block bootstrap, also introduced, up to second-order distribution theory. Finally, we introduce LDWB-aided Kolmogorov-Smirnov tests for local Gaussianity as well as local von-Mises statistics, with and without bootstrap inference, and establish their asymptotic validity using the second-order distribution theory. The finite sample performance of CLT and LDWB-aided local Gaussianity tests are assessed in a simulation study as well as two empirical applications. Whereas the CLT test is oversized, even in large samples, the size of the LDWB tests are accurate, even in small samples.

E0850: **Testing the equality of a large number of means of functional data**
Presenter: **Maria Dolores Jimenez-Gamero**, Universidad de Sevilla, Spain

Given k independent samples of functional data, the problem of testing for the equality of their mean functions is considered. In contrast to the classical setting where k is fixed and the sample size from each population increases without bound, k is assumed to be large and the size of each sample is either bounded or small in comparison to k. The asymptotic distribution of the considered test statistic is studied under the null hypothesis of equality of the k mean functions as well as under alternatives.

E0586: **Switching to the new norm: From heuristics to formal tests using integrable empirical processes**
Presenter: **Tetsuya Kaji**, University of Chicago, United States

A frequent concern in empirical research is to ensure that a handful of outlying observations have not driven the key empirical findings. We develop new theory of integrable empirical processes and apply it to construct a formal test of outlier robustness. The key is to observe that statistics related to outlier robustness analyses are represented as L-statistics-integrals of empirical quantile functions with respect to sample selection measures and to consider these elements in appropriate normed spaces. In particular, we characterize weak convergence of empirical distribution functions and sample selection functions in the space of bounded integrable functions, establish the delta method for empirical quantile functions as integrable functions, and then derive the delta method for L-statistics. We also prove the validity of nonparametric bootstrap. An empirical application shows the utility of the proposed test.

EO250 Room Aula C MACHINE LEARNING AND ROBUSTNESS Chair: Andreas Christmann

E0431: **Stability and generalization of stochastic gradient descent for pairwise learning**
Presenter: **Yiming Ying**, State University of New York at Albany, United States
Pairwise learning refers to a learning task which involves a loss function depending on pairs of instances. Most notable examples of pairwise learning include bipartite ranking, metric learning, AUC maximization and minimum error entropy (MEE) principle. We establish the stability results for SGD algorithms for pairwise learning in both convex, strongly-convex and non-convex settings. As a consequence, we derive their generalization error bounds. Finally, we describe our stability results by illustrating some specific examples of pairwise learning such as AUC maximization, metric learning and MEE. The motivation comes from a previous recent work and the results we obtain complement it in the setting of pointwise learning.

E0945: **Statistical learning for modal regression**
Presenter: **Jun Fan**, Hong Kong Baptist University, Hong Kong
The goal of supervised learning is to characterize a conditional distribution that can be used for prediction. Modal regression seeks the conditional mode of a response variable given a set of covariates, providing an alternative to mean regression and quantile regression in the presence of heavy-tailed noise, asymmetric noise or outliers. We study the modal regression estimator involving two types of kernels arising from kernel density estimation and reproducing kernel Hilbert spaces. Consistency results and learning rates are presented. Numerical results will be given to show the efficiency of the proposed method. Moreover, we will also discuss its connections to outlier detection and high-dimensional robust estimation.

E0987: **Efficient kernel-based learning by localization**
Presenter: **Ingo Steinwart**, University of Stuttgart, Germany
Despite the recent successes of (deep) neural networks kernel-based learning (KBL) methods remain one of the most successful learning methods for unstructured small to medium sized classification and regression problems. However, when it comes to large scale applications, their computational requirements, which grow super-linearly in the number of training samples, renders their application infeasible. To address this issue, several approaches that e.g. train KBL on many small chunks of the given large data set separately have been proposed in the literature. We consider such a decomposition strategy, called localized KBL, which is based upon a spatial partition of the input space. For this localized KBL, we derive a general oracle inequality describing its learning performance. Then we apply this oracle inequality to least squares regression using Gaussian kernels and deduce local learning rates that are essentially minimax optimal under some standard smoothness assumptions on the regression function. We further introduce a data-dependent parameter selection method for our localized KBL approach and show that this method achieves the same learning rates as before. Finally, we present some larger scale experiments for our localized KBL showing that it achieves essentially the same test performance as a global KBL for a fraction of the computational requirements.

E1725: **Unconventional regularization for efficient machine learning**
Presenter: **Lorenzo Rosasco**, Unige MIT IIT, Italy
Regularization is classically designed by penalizing or imposing explicit constraints to an empirical objective function. This approach can be derived from different perspectives and has optimal statistical guarantees. However, it postpones computational considerations to a separate analysis. In large scale scenarios, considering independently statistical and numerical aspects often leads to prohibitive computational requirements. It is then natural to ask whether different regularization principles exist or can be derived to encompass at once both statistical and computational aspects. Several ideas in this direction are presented; showing how procedures typically developed to perform efficient computations can often be seen as a form implicit regularization. We will discuss how iterative optimization of an empirical objective leads to regularization, and analyze the effect of acceleration, preconditioning and stochastic approximations. We will further discuss the regularization effect of sketching/subsampling methods by drawing a connection to classical regularization with projection methods common in applied mathematics. We will show how these forms of implicit regularization can obtain optimal statistical guarantees, with dramatically reduced computational properties.

EO607 Room B1 CAUSALITY: MODELING, REASONING, ESTIMATION AND PREDICTION II Chair: Joris Mooij

E0659: **Switching regression models and causal inference in the presence of latent variables**
Presenter: **Rune Christiansen**, University of Copenhagen, Denmark
Co-authors: Jonas Peters
Given a response Y and a vector $X = (X^1, \ldots, X^d)$ of predictors, the problem of inferring direct causes of Y among X is investigated. Models for Y that use its causal covariates as predictors enjoy the property of being invariant across different environments or interventional settings. Given data from such environments, this property has been exploited for causal discovery: one collects the models that show predictive stability across all environments and outputs the set of predictors that are necessary to obtain stability. If some of the direct causes are latent, however, there may not exist invariant models for Y based on variables from X, and the above reasoning breaks down. We therefore extend the principle of invariant prediction by proposing a relaxed version of the invariance assumption. This property can be used for causal discovery in the presence of latent variables if the latter's influence on Y can be restricted. More specifically, we allow for latent variables with a low-range discrete influence on the target Y. This assumption gives rise to switching regression models, where each value of the hidden variable corresponds to a different regression coefficient. We provide sufficient conditions for the existence, consistency and asymptotic normality of the MLE in switching regression models, and construct a test for the equality of such models. These results allow us to prove an asymptotic false discovery control of our causal discovery method.

E0901: **Anchor regression: Heterogeneous data meets causality**
Presenter: **Dominik Rothenhaeusler**, ETH Zurich, Switzerland
Co-authors: Nicolai Meinshausen, Peter Buehlmann, Jonas Peters
Many traditional statistical prediction methods mainly deal with the problem of overfitting to the given data set. On the other hand, there is a vast literature on the estimation of causal parameters for prediction under interventions. However, both types of estimators can perform poorly when used for prediction on heterogeneous data. We discuss the delicate trade-off between predictive performance on the training data and perturbed data. In particular, under a linear structural equation model with exogenous variables, we show that the change in loss under certain perturbations (interventions) can be written as a convex penalty. This motivates anchor regression, a regularization scheme that encourages the estimator to generalize well to perturbed data. Under instrumental variable (IV) assumptions, the procedure naturally provides an interpolation between the solution to ordinary least squares and the IV estimator. The proposed procedure allows statisticians and practitioners to trade-off predictive performance on the distribution of the training data and on distributions which are perturbed versions of what is seen in the training data.

E0961: **On the boundary between qualitative and quantitative measures of causal effects**
Presenter: **Linbo Wang**, University of Toronto, Canada
Co-authors: Yue Wang
Causal relationships among variables are commonly represented via directed acyclic graphs. There are many methods in the literature to quantify

the strength of arrows in a causal acyclic graph. These methods, however, have undesirable properties when the causal system represented by a directed acyclic graph is degenerate. We characterize a degenerate causal system using multiplicity of Markov boundaries, and show that in this case, it is impossible to quantify causal effects in a reasonable fashion. We then propose algorithms to identify such degenerate scenarios from observed data. Performance of our algorithms is investigated through synthetic data analysis.

E1125: **Causal discovery in the presence of missing data**
Presenter: **Cheng Zhang**, Mircrosoft Research, United Kingdom
Co-authors: Ruibo Tu, Kun Zhang, Hedvig Kjellstrom, Paul Ackermann
Missing data are ubiquitous in many domains such as healthcare. Depending on how they are missing, the (conditional) independence relations in the observed data may be different from those for the complete data generated by the underlying causal process and, as a consequence, simply applying existing causal discovery methods to the observed data may lead to wrong conclusions. It is then essential to extend existing causal discovery approaches to find true underlying causal structure from such incomplete data. We aim at solving this problem for data that are missing with different mechanisms, including missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). With missingness mechanisms represented by missingness Graph (m-Graph), we analyze conditions under which addition correction is needed to derive conditional independence/dependence relations in the complete data. Based on our analysis, we propose missing value PC (MVPC), which combines additional corrections with traditional causal discovery algorithm, in particular, PC. Our proposed MVPC is shown in theory to give asymptotically correct results even using data that are MAR and MNAR. Experiment results illustrate that the proposed algorithm can correct the conditional independence for values MCAR, MAR and rather general cases of values MNAR both with synthetic data as well as real-life healthcare application.

EO520 Room C1 QUANTILE REGRESSION METHODS Chair: Mauro Bernardi

E0250: **Quantile regression based seasonal adjustment**
Presenter: **Mohammed Elseidi**, University of Padova, Italy
Co-authors: Massimiliano Caporin
Different types of time series are affected by both deterministic and stochastic seasonal patterns. The usual assumption is that there is a unique seasonal pattern that affects the location and/or the scale of the variable of interest. However, there are cases where we observe, in a given time series, different seasonal patterns affecting the mean and the variance. Furthermore, seasonal patterns might affect higher order moments. Using traditional approaches for seasonal adjustment might not be efficient and does not ensure the adjusted data are free from periodic behaviors in higher order moments. We introduce a seasonal adjustment method based on quantile regression approach that is capable of capturing different forms of seasonal patterns. By describing the seasonal behavior over an approximation of the entire conditional distribution of a variable of interest, we might remove seasonal patterns affecting the mean and/or the variance only, or remove seasonal patterns varying over the distribution of the variable of interest. The results, focusing both on real and simulated data show the flexibility of the approach and the significant improvement over the traditional methods when the periodic behaviors impact on the distribution and/or on higher order moments.

E0545: **Variable selection in quantile varying coefficient models with heteroscedastic error**
Presenter: **Anneleen Verhasselt**, Hasselt University, Belgium
Co-authors: Mohammed Abdulkerim Ibrahim
Quantile regression is a great tool to get a thorough view of the relationship between (the distribution of) a response and covariates. We consider a location-scale quantile varying coefficient model with heteroscedastic error to model longitudinal data. In a longitudinal data setting, it is intuitive to allow the coefficients in the varying coefficient model to vary over time. The functional coefficients are estimated with penalized B-splines. As we allow for heteroscedasticity, the covariates can influence various quantiles of the response differently. Therefore, the problem of variable selection in quantile regression is more challenging. We consider grouped lasso and nonnegative garrote to perform variable selection in the location as well as the scale. When the problem is high-dimensional a two-stage approach, with a first screening stage (independence screening) is used, before applying grouped lasso or nonnegative garrote in the second stage.

E0572: **Sparse nonparametric dynamic graphical models**
Presenter: **Fabrizio Poggioni**, University La Sapienza, Italy
Co-authors: Mauro Bernardi, Lea Petrella
A sparse nonparametric dynamic graphical model is proposed for financial applications in which we employ a semiparametric multiple quantile model with CAViaR specification to describe the marginal distributions and a LASSO-penalized Gaussian copula-VAR model to describe the multivariate distribution of financial returns as a sparse dynamic model. In order to use the multiple quantile models as marginal distributions the estimated quantile functions must be invertible, in this way we can get the marginal CDFs from the estimated multiple quantiles. It is therefore necessary to guarantee the monotonicity of the estimated quantiles and, consequently, the absence of crossing. We contribute to the topic of quantile crossing for semiparametric models by defining a non-crossing parametric space for multiple quantile CAViaR models. Furthermore, we find computationally convenient to include the defined non-crossing necessary conditions as linear constraints to the multiple quantile estimation problem. Finally, we present an empirical application of the proposed methodology.

EO464 Room D1 MODELLING OF HIGH DIMENSIONAL DATA WITH BIOLOGICAL APPLICATIONS Chair: Anne Gegout-Petit

E0543: **Spatio-temporal modelling of the spread of chalara (illness of the ash tree) in France**
Presenter: **Anne Gegout-Petit**, Universite de Lorraine-IECL Inria BIGS, France
Co-authors: Coralie Fritsch, Benoit Marcais, Marie Grosdidier
Chalara is an illness of the tree ash that appears in Europe by Poland in 1992 and was observed for the first time in France in 2008. From 2008 to now, around 500 places of forest are visited every year in the purpose to notice the proportion of infected ashes. The mechanism of the illness is due to a fungus that grows during summer on ash rachis in the previous year's fallen leaves. The spores are produced and are transmitted by wind. According to this knowledge about chalara and in the purpose to a better understanding of the effect of covariates on the spread of the illness, we have built a spatio-temporal mechanistic model of propagation of the illness. It is based on a latent parametric model of spores production accounting on the effect of humidity and temperature and a reaction-diffusion model for the spores diffusion. For the inference, only based on the proportion of infected trees, we have used a Bayesian framework and MCMC simulations.

E0708: **Straightforward finding of differential expressions through intensive randomization in transcriptomic studies**
Presenter: **Dorota Desaulle**, University Paris Descartes, France
Co-authors: Bernard Hainque, Celine Hoffmann, Pascal Bigey, Yves Rozenholc
Transcriptomic data measure proportions of p transcripts to identify differential expressions (DE) from n samples under two or more experimental conditions, through e.g. a multiple testing procedure. Unfortunately, statistical analysis is sensitive to the unknown individual fraction of tissue reacting in the biological experiment. Multiplicative normalization adjusts for this intrinsic sample variability before any differential comparisons.

Apart from inadequate methods based on library size or housekeeping genes, normalizations try to find a proper subset of invariants to estimate the scaling factor. Such strategies fail on simple counter-examples by adjusting expression variabilities across the conditions and/or the DE. In this context of high dimensional data ($n << p$), under the assumption that the majority of analyzed expressions is invariant, we propose a new procedure for finding DE. It is straightforward as it does not rely on a previous good normalization. Instead, the findings are obtained by iterating the following steps: random selection of a small normalization subset regardless its quality, multiple testing for discovery of DE on the normalized expressions controlled by FDR. After the iterations, decreasing sorted rates of detection are compared to upper bound of the probabilities of choosing a (wrong) subset containing at least one DE when the number of true DE varies to obtain the discoveries. Our procedure globally controls the FDR.

E1030: **Multi-task Dirichlet-multinomial regression for detecting global microbiome associations**
Presenter: **Frank Dondelinger**, Lancaster University, United Kingdom

There is evidence that the human gut microbiome influences diseases as disparate as inflammatory bowel disease, cardiovascular disease and schizophrenia. Current statistical techniques for microbiome association studies either rely on a measure of microbiome distance, or on detecting associations with individual bacterial species. A method that extends the latter approach beyond individual species is the multi-task Dirichlet-multinomial model; however, it does not take species relatedness into account. We have improved on that approach in two respects: 1) by incorporating the phylogenetic tree of the microbial species as prior information about their relatedness, and 2) by introducing a hierarchical Bayesian prior that allows us to estimate the global effect of each covariate on the microbiome. We have applied our method to simulated data, and show that it allows for better estimation of global effects compared to a post-processing of the individual effects detected by the previous method. Additionally, we apply the data to two real-world examples in Crohn's disease and in inflammatory bowel disease. We show that our method can reliably detect global associations that are supported by the literature.

E1541: **Bayesian Ising sparse nonparametric model**
Presenter: **Inyoung Kim**, Virginia Tech, United States

A Bayesian variable selection approach is proposed via the graphical model and the Ising model. Our Bayesian variable problem can be considered as a complete graph and described by an Ising model with random interactions. There are several advantages of our approach: it is easy to (1) employ the single-site updating and cluster updating algorithm, both of which are suitable for problems with small sample sizes and a larger number of variables, (2) extend this approach to other regression models, and (3) incorporate graphical prior information. In our approach, the interactions are determined by the linear model coefficients, so we systematically study the performance of different scale normal mixture priors for the model coefficients by adopting the global-local shrinkage strategy. Our results indicate that the best prior for the model coefficients in terms of variable selection should place substantial weight on small, nonzero shrinkage. The methods are illustrated with simulated and pathway genomics data.

EO218 Room E1 PROJECTION PURSUIT Chair: Nicola Loperfido

E0201: **Projection pursuit in high dimensions**
Presenter: **Boaz Nadler**, Weizmann Institute of Science, Israel
Co-authors: Peter Bickel, Gil Kur

Projection pursuit is a classical exploratory technique to detecting interesting low dimensional structure in multivariate data. Motivated by contemporary applications, we study its properties in high dimensional settings. Specifically, we consider projection pursuit on structure-less Gaussian data with identity covariance, as both dimension p and sample size n tend to infinity, with p/n tending to a constant c. Our main results are that: (i) if $c = \infty$, there exist projections whose corresponding empirical cdf can approximate any arbitrary distribution; (ii) if $0 < c < \infty$, not all limiting distributions are possible. Yet, depending on the value of c various non-Gaussian distributions may still be approximated. In contrast, if we restrict to sparse projections, involving only few of the p variables, then asymptotically all empirical cdfs are Gaussian; and (iii) if $c = 0$, then asymptotically all projections are Gaussian. Some of these results extend to mean centered sub-Gaussian data and to projections into k dimensions. Hence, in small n, large p settings, unless sparsity is enforced and regardless of the chosen projection index, projection pursuit, may detect apparent structure that has no statistical significance. Fundamental limitations are revealed on the ability to detect non-Gaussian signals in high dimensional data, in particular via independent component analysis (ICA) and related non-Gaussian component analysis.

E0276: **MultiKurt: An R package for kurtosis-based data analysis**
Presenter: **Cinzia Franceschini**, Alma Mater Studiorum Bologna University, Italy
Co-authors: Nicola Loperfido

Multivariate kurtosis plays an important role in several areas of multivariate analysis: normality testing, projection pursuit, independent component analysis, model-based clustering, portfolio optimization and density approximation. MultiKurt is an R package purported to describe, testing and visualize multivariate kurtosis. In particular, it incorporates state-of-the-art algorithms for computing linear projections which either maximize, minimize or remove kurtosis. The package also computes scalar-valued and matrix-valued measures of multivariate kurtosis.

E0726: **Musings on projection pursuit**
Presenter: **Mu Zhu**, University of Waterloo, Canada

The literature on projection pursuit is perhaps not among the largest in statistics, but its existential outreach may be much wider than we realize. We will reflect on projection pursuit research, show examples of important on-going work that can be viewed as projection pursuit problems, and present some recent collaborative work with psychologists, in which our contribution can be said to have been driven by the very idea of projection pursuit itself.

E0395: **The skew-normal and related distributions as a copula and as a model for skewness persistence**
Presenter: **Chris Adcock**, SOAS - University of London, United Kingdom

An extension of the skew-normal distributions has been recently presented which is a copula. Under this model, the standardized marginal distributions are standard normal. The copula itself depends on the familiar skewing construction based on the normal distribution function. Two topics are considered. First, a number of extensions of the skew-normal copula are presented. Notably these include a case in which the standardized marginal distributions are Student's t, with different degrees of freedom allowed for each margin. In this case, the skewing function is not the distribution function for Student's t, but depends on certain of the special functions. Secondly, several multivariate versions of the skew-normal copula model are presented. These include as particular cases previous distributions. These distributions may be employed to model time varying skewness and skewness persistence, which is known to be a feature of some stock markets.

EO288 Room F1 ROBUST TESTS FOR CHANGE-POINTS IN TIME SERIES Chair: Herold Dehling

E0222: **Nuisance parameter free changepoint detection in non-stationary time series**
Presenter: **Martin Wendler**, University of Greifswald, Germany
Co-authors: Michal Pesta
For short range dependent time series, a new ratio type changepoint test is proposed which does involve any estimation of the long run covariance nor the choice of any tuning parameters. The new test is not affected by heteroskedasticity. The asymptotic distribution is derived with the help of the continuous mapping theorem. We will also present a robustified version of our test statistic and compare it to previous proposals: the ratio test and the self-normalized test. Some simulation results on the finite sample performance will illustrate our findings.

E0802: **Multiple change-point detection under short-range dependence**
Presenter: **Sara Kristin Schmidt**, Ruhr-University Bochum, Germany
Co-authors: Herold Dehling, Roland Fried, Max Wornowizki
A new test statistic is introduced to examine the presence of, potentially multiple, change-points in the variance of an absolutely regular time series. By regarding the test statistic as a U-statistic of a triangular array, a law of large numbers and a central limit theorem are established. Afterwards, asymptotic critical values of the test statistic are obtained and its performance is evaluated and compared to alternatives from the literature via a simulation study.

E0702: **Rank-based change-point analysis for long-range dependent time series**
Presenter: **Annika Betken**, Ruhr-Universitat Bochum, Germany
Co-authors: Martin Wendler
Change-point tests based on rank statistics are considered to test for structural changes in long-range dependent observations. Under the hypothesis of stationary time series as well as under the assumption of a structural change in the data, i.e. under (local) alternatives approaching the null hypothesis of no change, the asymptotic distributions of the corresponding test statistics are derived. For this, we prove a uniform reduction principle for the empirical process in a two-parameter Skorohod space equipped with a weighted supremum norm. Theoretical results are accompanied by simulation studies that are based on an approximation of the distribution of test statistics by subsampling procedures. The finite sample performance of change-point tests is compared for rank statistics resulting from different score functions.

E1330: **Power comparison of nonparametric change-point tests**
Presenter: **Herold Dehling**, Ruhr-University Bochum, Germany
Co-authors: Daniel Vogel, Martin Wendler
Some recent results on the power of nonparametric change point tests are presented. As one example, we will study two change point tests derived from a one sample U-statistic, and compare their power under local as well as under fixed alternatives.

EO196 Room G1 SURVIVAL ANALYSIS FOR CANCER STUDIES Chair: Thomas Scheike

E0270: **The matched cohort study for evaluation of post-treatment events of cancer survivors**
Presenter: **Thomas Scheike**, University of Copenhagen, Denmark
An excess risk regression model for matched cohort data is presented, where the occurrence of some events for individuals with a disease is compared to that of healthy controls that are matched at the onset-of-disease by various factors. By using the matched structure we show how to estimate the excess risk and its dependence on covariates on both proportional or additive form. An individual effect on the background mortality, possibly depending on matching factors, is removed considering differences. The model handles two different timescales, namely attained age and follow-up time. Firstly, we solve estimating equations for the non-parametric and parametric components of the excess risk model, providing large sample properties for the suggested estimators. Then, we report results from a simulation study. Lastly, we describe an application of the method on childhood cancer data, to study the excess risk of cardiovascular events in adults life among childhood cancer survivors.

E0269: **Modelling genetic influence on familial cancer risk: In honor of Fishers 100 year landmark biometrical genetics paper**
Presenter: **Jacob Hjelmborg**, University of Southern Denmark, Denmark
Co-authors: Thomas Scheike, Klaus Holst, Jaakko Kaprio
Modelling dependence among random variables is fundamental in statistics and becomes very much apparent when studying causes of complex traits taking for instance familial relationship into consideration. In 1918 R.A. Fisher published two papers that provided the foundation for studying genetic influence taking sources of variation into account. Variation is key in Fisher's contribution and we touch how the 'second moment' modelling of dependence is applied today, one hundred years after, providing insights to for instance familial risk of cancer disease taking censoring and competing risks into consideration. We further discuss the extension of modelling multiple cancer loci in a twin design setting. R.A. Fisher provided us statistical theory but also contributed with insights and concern in applications - multiple thanks.

E0567: **A screening study for identification of anticancer effects of prescription drugs**
Presenter: **Klaus Kaae Andersen**, Danish Cancer Society Research Center, Denmark
Cancer incidence is increasing on a global scale and cancer mortality is still high for many cancer types. A promising approach to meet the need for improved cancer therapy is repurposing of drugs initially marketed for non-cancer indications. We present preliminary results from an ongoing project where the aim is to search for new potential drug candidates for repurposing, by mounting a systematic screening for prescription drugs that might be associated with improved prognosis among cancer patients. We utilize Danish nationwide registries with continuous long-term data on drug prescriptions, cancer diagnoses, and mortality outcomes. We present an approach to set up the search in an automated way with procedures for ensuring reproducibility of results. Specifically, we use separate Cox models for each drug/cancer combination to estimate harm/benefit of the drug on cancer-specific mortality, while applying propensity scores for confounder adjustment. We adjust for multiple testing of main effects, and in evaluation of effect modification by treatment, relevant effects are filtered by p-values of corresponding main effects. All potential signals are evaluated for meaningful dose-response relationships along with calculation of prognostic indices, C-indices and Kaplan-Meier curves for risk groups. External validation is conducted in independent cohorts.

E1019: **Modelling the hazard of transition into the absorbing state in the illness-death survival model**
Presenter: **Laura Antolini**, Universita Milano Bicocca, Italy
Co-authors: Elena Tassistro, Davide Paolo Bernasconi, Paola Rebora, Maria Grazia Valsecchi
The illness-death survival model is the simplest multi-state model where the transition from the initial state 0 to the absorbing state 2 may involve an intermediate state 1 (e.g. disease relapse). The impact of the transition into 1 on the subsequent transition hazard to 2 enables to gain insights into the disease course. In this setting, the standard approach of analysis is the joint model of the transition hazards from 0 to 2 and from 1 to 2, including time to illness as a time-varying covariate and measuring time from origin even after the transition into state 1. The hazard from 1 to 2 can be also modelled separately only on patients in state 1, measuring time from illness and including time to illness as a fixed covariate. A recently

proposed approach is a joint model where time after the transition into state 1 is measured in both scales and time to illness is included as a time varying covariate. A further possibility is a joint model where time after the transition into state 1 is measured only from illness and time to illness is included as a fixed covariate. Through theoretical reasoning and simulation protocols the use of these models will be discussed when aiming at: a) validating the properties of the illness-death process, b) estimating the impact of time to illness on the hazard from 1 to 2, c) quantifying the impact that the transition into 1 has on the hazard of the absorbing state.

EO132 Room H1 SMALL AREA ESTIMATION Chair: Domingo Morales

E0563: **Estimation of additive parameters under unit-level gamma mixed models**
Presenter: **Tomas Hobza**, Czech Technical University in Prague, Czech Republic
Co-authors: Yolanda Marhuenda, Domingo Morales
Average incomes, poverty proportions and poverty gaps are additive parameters obtained as averages of given functions of an income variable. As the variable income has an asymmetric distribution, it is not properly modelled via normal distributions. When dealing with this type of variables, a first option is to apply transformations that approach normality. A second option is to use non-symmetric distributions like the gamma distribution. The use of unit-level gamma mixed models is proposed for modelling positive variables and for deriving three types of predictors of small area additive parameters, called empirical best, marginal and plug-in. The mean squared errors of the predictors are estimated by a parametric bootstrap. Some results of simulation experiments studying the behaviour of the small area predictors and the estimator of the mean squared errors are presented. By using data of the Spanish living condition survey of 2013, an application to the estimation of average incomes and poverty proportions in counties of the region of Valencia is given. In the real data application a procedure for estimating the nuisance parameters is proposed.

E0887: **Estimation of the gender pay gap using linear mixed models**
Presenter: **Esther Lopez Vizcaino**, Universidade da CoruÃśa, Spain
Co-authors: Maria Jose Lombardia, Cristina Rueda
An often used methodology to study labor market differences between men and women is to decompose the pay gap into a portion that is due to differences in group characteristics and a portion that cannot be explained by such differences. In this sense, the Blinder-Oaxaca decomposition divides the wage differential between two groups: one part that is explained by group differences in productivity characteristics, such as education or work experience, and a residual part that cannot be accounted for by such differences in wage determinants. This no-explained part is often used as a measure for discrimination, but it also takes in the effects of group differences in unobserved predictors. The Blinder-Oaxaca decomposition has been widely used in the study of labor market discrimination. The objective is to introduce a novel approach to the analysis of wage discrimination with methods that are robust to model (mis-)specification. Following this idea, we apply linear mixed models for the Oaxaca-Blinder decomposition of wage differentials between men and women. Also, we use the small area estimation (SAE) methodology to analyze the wage differentials by economic activities in the region of Galicia (Spain).

E0910: **Small area estimation with partially missing direct estimates**
Presenter: **Jan Pablo Burgard**, Trier University, Germany
Co-authors: Domingo Morales, Anna-Lena Woelwer
A common problem in small area estimation is that direct estimates are not reported for some areas or domains. This typically occurs when the areas or domains of interest are not planned domains in the sampling design. But also too imprecise estimates, e.g. because of too low sample sizes, lead statistical agencies to suppress their publication. This inhibits the use of classical small area estimation methods like the Fay-Herriot model. We propose an empirical best predictor for the prediction of area and domain parameters for the situation of partially missing direct estimates. Further, we propose a REML-like algorithm to estimate model parameters. Besides the parameter estimation we also show a first approach for the estimation of the MSE of the empirical best predictor. The performance of this method is evaluated in a model-based simulation study, showing both advantages and caveats of the newly proposed estimator.

E1127: **Modelling educational poverty by area-level SAE latent Markov models**
Presenter: **Gaia Bertarelli**, University of Pisa, Italy
Co-authors: Caterina Giusti, Monica Pratesi, Francesco Bartolucci, Maria Giovanna Ranalli, Luciana Quattrociocchi
Educational Poverty (EP) is defined as deprivation, for children and adolescents, of the ability to learn, experiment, develop and freely flourish skills, talents and aspirations. EP is a latent trait, namely, only indirectly measurable through a collection of observable variables and indicators purposively selected as micro-aspects, contributing to the latent macro-dimension. It is generally measured by ISTAT by two multidimensional indices, the educational poverty index and the adjusted Mazziotta-Pareto index. A problem with these indices is that they are based on direct estimates, which are reliable only at the broad-areas level, while to intervene on the phenomenon it is important to obtain information at a finer geographical level. We use an adapted version of area-level SAE method that uses a Latent Markov Model (LMM) as linking model. In LMMs the characteristic of interest and its evolution in time is represented by a latent process that follows a discrete Markov chain. Therefore, areas are allowed to change their latent state across time. This model can handle both univariate and multivariate characteristics of interest and can provide a classification of the areas by the intensity of their EP. That is, we can use the area-level data of the single dimension or the value of the composite indices.

EO398 Room I1 MODELS AND THEIR INFERENCES FOR CIRCULAR DATA Chair: Toshihiro Abe

E0984: **Adaptive estimation for mode and anti-mode preserving distribution on the circle**
Presenter: **Takayuki Shiohama**, Tokyo University of Science, Japan
The efficient estimation for the parameters in mode and anti-mode preserving distribution on the circle is considered. For practical purpose, it is difficult to assume that the base density function in advance, which are transformed to be skewed and sharply peaked distributions by inverse functions. However, starting with an initial computationally convenient estimator, one can always gain efficiency by doing a one-step from an initial estimate using a likelihood or other efficient criterion function. The proposed methods are extended to the case with circular-circular regression models with mode and anti-mode preserving error distributions. Some Monte Carlo simulations are illustrated for the performance comparison of our proposed one-step estimators among other estimators.

E0468: **On the use of transformations on the circle**
Presenter: **Domien Craens**, UGent, Belgium
Co-authors: Christophe Ley
The transformation approach is widely used for developing new distributions on the real line. On the circle, however, it has so far received little attention. We will expand on this approach for circular random variables and discuss various properties and applications. Besides the famous Moebius transformation, we shall present new transformations, that add more flexibility to existing distributions such as the von Mises, and the corresponding likelihood ratio tests.

E0711: **Hidden Markov random fields for the spatial segmentation of circular data**
Presenter: **Jose Ameijeiras-Alonso**, University of Santiago de Compostela, Spain
Co-authors: Francesco Lagona, Monia Ranalli, Rosa Crujeiras
The aim is to present a model for providing a spatial segmentation of circular data according to a finite number of latent classes employing a hidden Markov random field. Under this setting, the data are modelled by a finite mixture of parametric densities, whose parameters vary across space according to a latent Markov random field. As such, it can be viewed as an extension of a mixture model to the spatial setting. Motivated by wildfires data in the Iberian Peninsula, a model based on a mixture of Kato-Jones circular densities is suggested. This model takes into account special features of wildfire occurrence data such as multimodality, skewness and kurtosis. The parameters of the model will vary across space according to a latent Potts model, modulated by geo-referenced covariates.

E0774: **Analysis of circular interval-censored data motivated by aoristic data in criminology**
Presenter: **Kees Mulder**, Utrecht University, Netherlands
Co-authors: Stijn Ruiter, Irene Klugkist
Motivated by an application in criminology, methods are developed for parameteric and non-parametric density estimation on the circle when interval-censored data are given. In crime analysis, there is an interest to determine at what time of day crimes have happened, which leads to the analysis of circular data. The objective is to estimate the frequency of crime at any given time of the day, day of the week, or day of the year. However, victims of crimes such as theft and burglary often do not know when exactly the crime occurred, knowing only the time interval during which the crime must have happened. Therefore, the temporal information in crime data is often interval-censored. Current so-called aoristic analysis methods comprise mostly of histogram-like methods for density estimation. We extend these methods from a statistical perspective. It can be shown that standard aoristic analysis results in densities with a variance that is too high. We show how likelihood estimation can be adapted to work with interval-censored data. We fit parametric models based on the von Mises distribution and Bayesian semiparametric models based on the Dirichlet process. The advantage of these models is that they use all available information for the density estimation without the requirement to select an arbitrary bandwidth or something alike. These new methods are available as an R package.

EO180 Room L1 ADVANCES IN MODEL-BASED CLUSTERING Chair: Vincent Vandewalle

E1128: **Clustering multivariate count data via a mixture of Poisson factor models**
Presenter: **Yang Tang**, McMaster University, Canada
Co-authors: Paul McNicholas
Dependencies in multivariate counts are of interest in many applications, but few approaches have been proposed for their analysis. We develop a mixture of Poisson factor models to explore the dependencies in multivariate count data. We assume that the p-dimensional random vector of counts Y is modeled using a q-dimensional vector of latent factors U and U follows an inverse Gaussian distribution. The proposed framework provides a parsimonious and easy to interpret representation of multivariate dependencies in counts. Parameter estimation is carried out and information criteria are used for model selection. The use of these models is demonstrated on real and simulated data.

E0926: **Mixtures of skewed matrix variate bilinear factor analyzers**
Presenter: **Michael Gallaugher**, McMaster University, Canada
Co-authors: Paul McNicholas
Clustering is the process of finding and analyzing underlying group structures in data. In recent years, data as become increasingly higher dimensional and therefore an increased need for dimension reduction techniques for use in clustering. Although such techniques are firmly established in the literature for multivariate data, there is a relative paucity in the area of matrix variate or three way data. Furthermore, these few methods all assume matrix variate normality which is not always sensible if skewness is present. We propose a mixture of bilinear factor analyzers model using four skewed matrix variate distributions, namely the matrix variate skew-t, generalized hyperbolic, variance gamma and normal inverse Gaussian distributions. Both simulated and real data will be used for illustration.

E1067: **Variable selection for model-based clustering of multivariate longitudinal categorical data**
Presenter: **Michael Fop**, University College Dublin, Ireland
Multivariate longitudinal categorical data commonly arise in the health and social sciences, where information about the same units is collected at several time occasions, usually by means of a panel survey. The latent Markov model is often employed for the analysis of such data. In this model, a categorical latent variable is assumed to follow a Markov chain with a finite number of states and to influence the evolution over time of the observed variables. Thanks to the Markov process, subjects are allowed to move between latent states with certain transition probabilities. Hence, at a given time point, model-based clustering can be performed by aggregating subjects into different clusters corresponding to the latent states. Moreover, subjects can move between clusters over the considered time period. In this setting, may be of interest to select which variables impact the transition between the latent states and which variables are most informative to cluster the units at a fixed time point. A method for variable selection and clustering of multivariate longitudinal categorical data is proposed based on the latent Markov model. A general framework is developed for selection of both the relevant clustering variables at a given time occasion and those variables affecting the transition process between the latent states.

E0927: **A targeted multi-partitions clustering**
Presenter: **Matthieu Marbac**, CREST - ENSAI, France
Co-authors: Christophe Biernacki, Mohammed Sedki, Vincent Vandewalle
Clustering is generally not a purpose by itself, because its results are mainly tools used by the statistician for another analysis. Indeed, in many applications, clusters are assessed from a set of observed variables, then these clusters are used to predict other variables which are used or not in clustering. Because the final objective of prediction is not considered during cluster analysis, there is no reason to obtain relevant clusters for the variables to predict. We present a unified approach which simultaneously performs cluster analysis and prediction. This method considers that the variables to clusters arise from a product of finite mixture models which provides multiple partition. Moreover, the variables to predict are considered to be independent to the variables to cluster given the partition. The predictions are achieved by a generalized linear model. Model selection is conducted by optimizing the BIC. This optimization is achieved with a modified version of the EM algorithm which performs model selection and maximum likelihood inference simultaneously.

EO230 Room M1 STATISTICS FOR HILBERT SPACES I Chair: Gil Gonzalez-Rodriguez

E0220: **A general theory for large-scale curve time series via functional stability measure**
Presenter: **Xinghao Qiao**, London School of Economics, United Kingdom
Co-authors: Shaojun Guo
Modelling a large bundle of curves arises in a broad spectrum of real applications. However, studies in functional data analysis rely primarily on the critical assumption of independent curve observations. We introduce a general theory for large-scale curve time series, where the number of functional variables, p, is large relative to the number of observations, n, and the dynamical dependence across observations exists. We propose a functional stability measure for stationary functional processes based on their spectral properties and use it to establish the concentration bounds on the sample covariance matrix function under different functional matrix norms. We also develop concentration results on the relevant estimated terms under the Karhunen-Loeve expansion framework. To illustrate with an example, we consider the vector functional autoregressive models, which characterize the temporal and cross-sectional dependence across multiple curve time series. We develop a regularization approach to estimate the autoregressive coefficient functions under the sparsity constraint. Using our derived concentration inequalities, we investigate the theoretical properties of the regularized estimate in the high-dimensional large p, small n, regime. Finally, we evaluate the sample performance of the proposed method through simulation studies.

E1194: **New tests for equality of several covariance functions for functional data**
Presenter: **Jin-Ting Zhang**, National University of Singapore, Singapore
Two new tests for the equality of the covariance functions of several functional populations are discussed, namely a quasi GPF test and a quasi Fmax test. Unlike several existing tests, they are scale-invariant in the sense that their test statistics will not change if we multiply each of the observed functions by any non-zero function of time. We derive the asymptotic random expressions of the two tests under the null hypothesis and show that under some mild conditions, the asymptotic null distribution of the quasi GPF test is a chi-squared-type mixture whose distribution can be well approximated by a simple scaled chi-squared distribution. We also describe a random permutation method for approximating the null distributions of the quasi GPF and Fmax tests. Simulation studies are presented to demonstrate the finite-sample performance of the new tests against five existing tests. An illustrative example is also presented.

E1196: **About the complexity of a functional data set**
Presenter: **Enea Bongiorno**, Universita del Piemonte Orientale, Italy
Co-authors: Aldo Goia, Philippe Vieu
Consider the problem to state the compatibility of observed functional data with a reference model. Starting from the small ball probability factorization, it is possible to introduce the concept of complexity for functional data and suitable indexes measuring it. At a first stage, a descriptive approach, mainly based on a new graphical tool (namely the log-Volugram), is implemented and fruitfully applied. From an inferential perspective, a hypothesis test is implemented: the test statistic is derived, its asymptotic law is studied, a study of level and power of the test for finite sample sizes and a comparison with a competitor are carried out by Monte Carlo simulations. It turns out that the developed methodologies are fully free from assumptions on model, distribution as well as dominating measure. Applications are provided over financial time series.

E0309: **Modelling function-valued processes with nonseparable covariance structure**
Presenter: **Evandro Konzen**, Newcastle University, United Kingdom
Co-authors: Jian Qing Shi
Separability of the covariance structure is a common assumption for function-valued processes defined on two- or higher-dimensional domains. This assumption is often made to obtain an interpretable model or due to difficulties in modelling a potentially complex covariance structure, especially in the case of sparse designs. We suggest using Gaussian processes with flexible parametric covariance kernels which allow interactions between the inputs in the covariance structure. When we use suitable covariance kernels, the leading eigensurfaces of the covariance operator can explain well the main modes of variation in the functional data, including the interactions between the inputs. The results are demonstrated by simulation studies and by an application to human fertility data.

EO360 Room N1 PERFORMANCE EVALUATION AND DEPENDENCE MODELING FOR EXTREMES Chair: Holger Rootzen

E0367: **Forecasters dilemma: Extreme events and forecast evaluation**
Presenter: **Sebastian Lerch**, Karlsruhe Institute of Technology, Germany
Co-authors: Thordis Thorarinsdottir, Francesco Ravazzolo, Tilmann Gneiting
In public discussions of the quality of forecasts, attention typically focuses on the predictive performance in cases of extreme events. However, the restriction of conventional forecast evaluation methods to subsets of extreme observations has unexpected and undesired effects, and is bound to discredit skillful forecasts when the signal-to-noise ratio in the data generating process is low. Conditioning on outcomes is incompatible with the theoretical assumptions of established forecast evaluation methods, thereby confronting forecasters with what we refer to as the forecasters dilemma. For probabilistic forecasts, proper weighted scoring rules have been proposed as decision-theoretically justifiable alternatives for forecast evaluation with an emphasis on extreme events. Using theoretical arguments, simulation experiments and a real data study on probabilistic forecasts of U.S. inflation and gross domestic product (GDP) growth, we illustrate and discuss the forecasters dilemma along with potential remedies.

E0753: **On a minimum distance procedure to select the optimal sample fraction in extreme value estimation**
Presenter: **Holger Drees**, University of Hamburg, Germany
Co-authors: Anja Janssen, Sid Resnick, Tiandong Wang
Many estimators of the extreme value index and other tail parameters use a certain fraction of largest observations. The data-driven choice of this fraction is a notoriously difficult problem. A previous influential paper suggested fitting a generalized Pareto distribution (GPD) to the top k order statistics for all possible k and choose the value that minimizes the Kolmogorov-Smirnov distance between the fitted GPD and the empirical cdf of the exceedances. By the example of the Hill estimator, we will argue why this minimum distance approach usually leads to an inefficient tail estimator. In particular, a serious underestimation of the optimal sample fraction leads to a largely increased asymptotic variance, which can also be observed in simulations.

E1021: **Generalized Pareto copulas: A key to multivariate extremes**
Presenter: **Simone Padoan**, Bocconi University, Italy
Co-authors: Michael Falk
In recent years many efforts have been made by different authors to characterize what a multivariate generalized Pareto distribution is. Generalized Pareto copulas (GPC) are presented. Any GPC can be represented in a simple, analytical way using a particular type of norm on $\mathbb{R}^d$, called D-norm. The characteristic property of a GPC is its *exceedance stability*. The GPC turns out to be a key to multivariate extreme value theory. We discuss the inferential aspects related to GPC and we show its utility analyzying a real dataset.

E1110: **Bayesian networks based on max-linear structural equations**
Presenter: **Claudia Klueppelberg**, Technical University of Munich, Germany
Bayesian networks based on max-linear structural equations are studied and a summary of their independence properties is provided. In particular we emphasize that distributions for such networks are never faithful to the independence model determined by their associated directed acyclic graph unless the latter is a polytree, in which case they are always faithful. In addition, we consider some of the basic issues of estimation and discuss generalised maximum likelihood estimation of the coefficients. Finally, we argue that the structure of a minimal network asymptotically can be identified completely from observational data.

EO306 Room O1 DEPENDENCE MODELS AND COPULAS Chair: Elisa Perrone

E0243: **Confidence regions for multivariate quantiles**
Presenter: **Maximilian Coblenz**, Karlsruhe Institute of Technology, Germany
Co-authors: Rainer Dyckerhoff, Oliver Grothe
Multivariate quantiles are of increasing importance in applications of hydrology. This calls for reliable methods to evaluate the precision of the estimated quantile sets. Therefore, we investigate two recently developed approaches to estimate confidence regions for level sets. These are extended to multivariate quantiles based on copulas. In a simulation study we check coverage probabilities of the employed approaches. A focus is on small sample sizes. Overall, the adapted approaches show reasonable coverage probabilities. However, not only the bounded copula domain but also the additional estimation of the quantile level pose some problems. A small sample application gives further insight into the employed techniques.

E0712: **A classification point-of-view about conditional Kendall's tau**
Presenter: **Alexis Derumigny**, ENSAE-CREST, France
Co-authors: Jean-David Fermanian
The problem of estimating conditional Kendall's tau is shown to be rewritten as a classification task. Conditional Kendall's tau is a conditional dependence parameter that is a characteristic of a given pair of random variables. The goal is to predict whether the pair is concordant (value of 1) or discordant (value of -1) conditionally on some covariates. We prove the consistency and the asymptotic normality of a family of penalized approximate maximum likelihood estimators, including the equivalent of the logit and probit regressions in our framework. Then, we detail specific algorithms adapting usual machine learning techniques, including nearest neighbors, decision trees, random forests and neural networks, to the setting of the estimation of conditional Kendall's tau. A small simulation study compares their finite sample properties. Finally, we apply all these estimators to a dataset of European stock indices.

E0415: **A multivariate dependence analysis of electricity prices**
Presenter: **Luca Rossini**, Vrije Universiteit Amsterdam, Netherlands
Co-authors: Fabrizio Durante, Francesco Ravazzolo, Angelica Gianfreda
The purpose is to study the joint interdependence between electricity prices, demand and renewable energy sources (RES). Several papers have tried to understand the underling non-linear relationship between electricity prices and fundamental factors. However, only few have considered the dynamics affecting the entire distribution, and more importantly, they generally focused on a bivariate relationship that is looking at price-demand or price-wind interactions. Therefore, we aim at filling this gap inspecting a multivariate dependence structure between all four variables specifying proper marginal distributions later used in multivariate copula models. By using the well-established AR-GARCH framework and two types of copulae, we provide for the first time evidence of a time-varying multivariate dependence structure. Finally, density forecasting performances across copula models are tested and compared to provide operational insights. The interactions between electricity prices, demand and electricity generated by renewable energy sources (wind and solar photovoltaic) are investigated across the 24 hours from 2011 to 2017 in the German market, which has been characterized by an increasing level of RES penetration. Our dataset consists of hourly prices collected from the European Energy Exchange, EEX; and hourly forecasted demand, wind and solar generation, collected from Thomson Reuters.

E0800: **An ordering for extremal dependence**
Presenter: **Christopher Strothmann**, TU Dortmund, Germany
Co-authors: Karl Friedrich Siburg
A preorder of tail dependence is introduced to compare different degrees of extremal dependence. Several properties of this preorder, e.g. maximal and minimal elements, symmetry and its relation to the tail-order function, are investigated. Afterwards, various monotone measures for the tail-dependence order are considered and evaluated according to their relevance for some well-known copula families. Finally, an approach to generate families of copulas which are ordered by their tail-dependence functions is presented.

EO414 Room P1 STATISTICAL THEORY AND COMPUTATION FOR ULTRA HIGH FREQUENCY DATA Chair: Nakahiro Yoshida

E0434: **Further developments of the ratio model of Cox-type intensities for high frequency financial data**
Presenter: **Ioane Muni Toke**, CentraleSupelec, France
A model of ratio of Cox intensities sharing a common baseline intensity has been recently proposed. We have shown that such a model is particularly well designed for high frequency financial data, where the common baseline intensity may represent a rapidly varying market activity. We are, for example, able to estimate relative effects of parameters of interest on the intensity of submission of orders in a limit order book. We now build on this model in several directions. In a first direction, we investigate the variations across time of the influences identified by the ratio model, and show the benefits of the model as an econometric tool. In a second direction, we further investigate limit order book modeling by developing a combined model of Cox- and Hawkes-type intensities (with multiple exponential kernels for the latter type) for the submission of orders. The estimation of such a model with (ultra) high frequency data is a challenging task that is carried out on several stocks and exchanges.

E0908: **Locally stable regression with unknown activity index**
Presenter: **Hiroki Masuda**, Kyushu University, Japan
Typically, transition of large-scale dependent data, such as those sampled at ultra high-frequency, are highly non-Gaussian. One of natural ways of modeling such data would be to use continuous-time stochastic processes driven by a non-Gaussian pure-jump noise. The related existing literature is, however, still far from being well-developed. We present tailor-made quasi-likelihood inference results that can efficiently handle such locally and highly non-Gaussian statistical models with the activity index of the driving noise process being unknown. The model setup includes not only Markovian stochastic differential equations but also a class of semimartingale regression models. Of primary interest are cases where estimation target includes not only the rapidly varying scale structure but also the slowly varying trend one.

E0771: **Bayesian inference for stable Lévy driven stochastic differential equations with high-frequency data**
Presenter: **Kengo Kamatani**, Osaka University, Japan
Co-authors: Ajay Jasra, Hiroki Masuda
The focus is on parametric Bayesian inference for stochastic differential equations (SDE) driven by a pure-jump stable Lévy process, which is observed at high frequency. In most cases of practical interest, the likelihood function is not available, so we use a quasi-likelihood and place an associated prior to the unknown parameters. It is shown under regularity conditions that there is a Bernstein-von Mises theorem associated to the posterior. We then develop a Markov chain Monte Carlo (MCMC) algorithm for Bayesian inference and assisted by our theoretical results, we show how to scale Metropolis-Hastings proposals when the frequency of the data grows, in order to prevent the acceptance ratio going to zero in the large data limit. Our algorithm is presented on numerical examples that help to verify our theoretical findings.

E0484: **Hybrid estimation for an ergodic diffusion plus noise based on ultra high frequency data**
Presenter: **Masayuki Uchida**, Osaka University, Japan
Hybrid estimation of both drift and volatility parameters for an ergodic diffusion processes plus noise based on ultra high frequency data is considered. Adaptive maximum likelihood type estimators (MLEs) of both drift and volatility parameters for a discretely observed ergodic diffusion processes plus noise have been proposed, and the asymptotic properties of the adaptive MLEs have been shown. In order to get the quasi MLE, it is crucial to choose a suitable initial estimator for optimization of the quasi likelihood function. From a computational point of view, we propose initial Bayes type estimators (initial BEs) of both drift and volatility parameters and the adaptive MLEs with the initial BEs, which are called hybrid estimators, are proposed. It is shown that the hybrid estimators with the initial BEs have asymptotic normality and convergence of moments. We also give an example and simulation results.

EO184 Room Q1 ROBUST METHODS FOR HIGH DIMENSIONAL DATA Chair: Peter Rousseeuw

E0212: **Cellwise robust regularized discriminant analysis**
Presenter: **Ines Wilms**, KU Leuven, Belgium
Co-authors: Stephanie Aerts
Quadratic and Linear Discriminant Analysis (QDA/LDA) are the most often applied classification rules under normality. In QDA, a separate covariance matrix is estimated for each group. If there are more variables than observations in the groups, the usual estimates are singular and cannot be used anymore. Assuming homoscedasticity, as in LDA, reduces the number of parameters to estimate. This rather strong assumption is however rarely verified in practice. Regularized discriminant techniques that are computable in high-dimension and cover the path between the two extremes QDA and LDA have been proposed in the literature. However, these procedures rely on sample covariance matrices. As such, they become inappropriate in presence of cellwise outliers, a type of outliers that is very likely to occur in high-dimensional datasets. We propose cellwise robust counterparts of these regularized discriminant techniques by inserting cellwise robust covariance matrices. Our methodology results in a family of discriminant methods that (i) are robust against outlying cells, (ii) provide, as a by-product, a way to detect outliers, (iii) cover the path between LDA and QDA, and (iv) are computable in high-dimension. The good performance of the new methods is illustrated through simulated and real data examples.

E0295: **Sparse principal component analysis based on least trimmed squares**
Presenter: **Yixin Wang**, University of Leuven, Belgium
Co-authors: Stefan Van Aelst
Sparse principal component analysis can be used to obtain stable and interpretable principal components from high-dimensional data. Robust sparse PCA is considered to handle outliers in the data. The new method LTS-SPCA starts from the MLTS-PCA method which provides a robust but non-sparse PCA solution. MLTS-PCA yields the PC subspace corresponding to the proportion of the data which gives the smallest sum of squared residuals. To get sparse solutions, LTS-SPCA then incorporates an l_1-norm penalty on the loading vectors to obtain sparsity. LTS-SPCA searches for the PC directions sequentially. This approach avoids that score outliers in the PC subspace destroy the sparse structure of the loadings. Simulation studies and real data examples show that LTS-SPCA can give accurate estimates, even when the data is highly contaminated. Moreover, compared to existing robust sparse PCA methods, LTS-SPCA can reduce the computation time to a great extent.

E0583: **The minimum regularized covariance determinant estimator**
Presenter: **Tim Verdonck**, KU Leuven, Belgium
Co-authors: Kris Boudt, Peter Rousseeuw, Steven Vanduffel
The Minimum Covariance Determinant (MCD) approach estimates the location and scatter matrix using the subset of given size with lowest sample covariance determinant. Its main drawback is that it cannot be applied when the dimension exceeds the subset size. We propose the Minimum Regularized Covariance Determinant (MRCD) approach, which differs from the MCD in that the subset-based covariance matrix is a convex combination of a target matrix and the sample covariance matrix. A data-driven procedure sets the weight of the target matrix, so that the regularization is only used when needed. The MRCD estimator is defined in any dimension, is well-conditioned by construction and preserves the good robustness properties of the MCD. We prove that so-called concentration steps can be performed to reduce the MRCD objective function, and we exploit this fact to construct a fast algorithm. We verify the accuracy and robustness of the MRCD estimator in a simulation study and illustrate its practical use for outlier detection and regression analysis on real-life high-dimensional data sets in chemistry and criminology.

E0895: **Invariant coordinate selection for outlier detection with application to quality control**
Presenter: **Anne Ruiz-Gazen**, Toulouse School of Economics, France
Co-authors: Aurore Archimbaud, Klaus Nordhausen
Detecting outliers in multivariate data sets is of particular interest in various contexts including quality control in high standards fields such as automotive or avionics. Some classical detection methods are based on the Mahalanobis distance or on robust Principal Component Analysis (PCA). One advantage of the Mahalanobis distance is its affine invariance while PCA is only invariant under orthogonal transformations. For its part, PCA allows some components selection and facilitates the interpretation of the detected outliers. We propose an alternative in a casewise contamination context and when the number of observations is larger than the number of variables, called invariant coordinate selection. Its principle is quite similar in spirit to PCA with invariant components derived from an eigendecomposition followed by a projection of the data on some selected eigenvectors. The decomposition is based on two scatter matrix estimators instead of one for PCA. While principal components are scale dependent, the invariant components are affine invariant for affine equivariant scatter matrices. Moreover, under some elliptical mixture models, the Fisher's linear discriminant subspace coincides with a subset of invariant components in the case where group identifications are unknown. The method will be illustrated on several data sets from the quality control field. The problem of multicollinearity and singular scatter matrices will be also advocated.

EO208 Room O2 CSDA JOURNAL: BAYESIAN METHODS Chair: Alastair Young

E0600: **A fast estimation method in nonparametric additive location-scale model based on Bayesian P-splines**
Presenter: **Philippe Lambert**, Universite de Liege / Universite catholique de Louvain, Belgium
In a previous publication on nonparametric additive location-scale models for interval censored data, it has been explained how Bayesian P-splines could be used in regression models to specify a smooth error density and the joint (possibly) nonlinear effects of covariates on location and dispersion. That methodology extends traditional additive regression models by releasing the parametric constraint on the error distribution and by acknowledging that covariates can affect multiple aspects of the conditional distribution in a non trivial way. These extensions are very attractive and practically useful, but have an important computational cost following from the use of the Metropolis-within-Gibbs algorithm in a richly parameterized model. We show how Laplace based approximations to the marginal posterior distributions of smoothness parameters can be used to set up a quickly converging iterative algorithm to select penalty parameters and to estimate the spline parameters in the pivotal distribution and in the additive components for location and dispersion in a fast and reliable way, as confirmed by simulation results. We conclude the presentation with an application to survey data.

E1008: **Bayesian methods in biomedical imaging**
Presenter: **Michele Guindani**, University of California, Irvine, United States
The use of flexible Bayesian approaches in biomedical imaging is discussed. We will first discuss applications to the analysis of task-related fMRI data in single-subject and multi-subject experiments, where the aim is to account for the heterogeneity in neuronal activity both within- and between- subjects. We will then discuss an application to cancer radiomics, an emerging discipline that promises to elucidate lesion phenotypes and tumor heterogeneity through the analysis of large amounts of quantitative imaging features that can be derived from medical images. We will show how a fully Bayesian probabilistic framework may help characterizing the heterogeneity of adrenal lesions images obtained from CT scans more precisely than a class of machine-learning approaches currently used in cancer radiomics. We further assess whether the subtypes resultant from our analysis are clinically oriented by investigating their correspondence with pathological diagnoses.

E1130: **Bayesian capture-recapture data modelling with behavioural effects**
Presenter: **Luca Tardella**, Sapienza University of Rome, Italy
Co-authors: Danilo Alunni Fegatelli
In the context of capture-recapture sampling we rely on the generalized linear model framework for modelling behavioural effects by regressing the capture occurrence on previous partial capture histories although shortcuts have been embedded to reduce computational complexity whenever possible. In particular, we extend the modelling ideas of using suitable meaningful summaries of individual previous partial histories. This leads to generalizing the Markov dependence in the presence of a non-linear regression function. Theoretical arguments related to the so-called likelihood failure support the use of a Bayesian approach for the estimation of the unknown population size in the presence of behavioral response to capture. Posterior summaries for inferring the population size within the Bayesian logistic regression is carried out exploiting idea of data augmentation in logistic models using the class of Polya-Gamma distributions hence allowing for a general and flexible computational and modelling framework. Special emphasis will be also given to possible alternative Bayesian computation strategies for model selection. Simulated and real data analysis shows the comparative effectiveness of the proposed inferential approach.

E0599: **Frequentist validation and criticism of Bayesian selective inference**
Presenter: **Alastair Young**, Imperial College London, United Kingdom
Co-authors: Daniel Garcia Rasines
As much as frequentist approaches, Bayesian inference is challenged by the problem of selective inference, where the analyst interacts with the data to select what questions about an underlying population should be addressed. A conceptual framework for selection-adjusted Bayesian inference, based on specifying explicitly the selection rule which determines when inference is provided for a particular parameter, is considered. Inference is based on the selection-adjusted posterior distribution of the parameter, obtained from a model that prepends a prior to a truncated data likelihood. We examine the repeated sampling properties of the inference, revealing non-trivial and practically significant asymptotic as well as finite repeated sampling behavior.

EO038 Room Q2 ADVANCES IN BAYESIAN METHODOLOGY Chair: David van Dyk

E0397: **Posterior model selection consistency: Beyond asymptotic optimality**
Presenter: **David Rossell**, Universitat Pompeu Fabra, Spain
An important property for Bayesian model selection is that the posterior probability of the data-generating model (or Kullback-Leibler closest to it) converges to 1, and that the corresponding convergence rate is fast. This guarantees frequentist model selection consistency and asymptotically valid uncertainty quantification. The aim is two-fold. First, we provide a general framework to study consistency for any given model, prior, sample size n and dimension p, and potentially under model misspecification. Second, we deploy it to canonical variable selection and show that there can be a big gap between lessons learnt from asymptotically optimal rates (the large n, even large p paradigm) and practical situations with finite sample size (the small n, large p paradigm). Specifically, this gives interesting insights regarding sparsity/sensitivity tradeoffs, e.g. one may forsake asymptotic optimality to obtain significant gains in power. These gains are noticeable even in simple sparse scenarios, and become more relevant in truly non-sparse settings.

E0655: **Strategies for differential shrinkage in regression with non-orthogonal designs**
Presenter: **Christopher Hans**, The Ohio State University, United States
Thick-tailed mixtures of g priors that mix over a single, common scale parameter have gained traction as a default prior in Bayesian regression settings. Such priors shrink all regression coefficients in the same manner and can negatively impact inference and model comparison in situations where differential shrinkage across regression coefficients is appropriate. We will review two known deficiencies of existing mixtures of g priors that arise under an asymptotic regime that is motivated by the common data analytic setting where one regression coefficient is expected to be much larger than the others. The driver behind these undesirable behaviors is the use of a latent scale parameter that is common to all coefficients. Classes of block hyper-g priors that employ differential shrinkage across groups of coefficients have been proposed to avoid these behaviors, however the theory underlying these priors requires that the regression design matrix has a block orthogonal structure. Extensions are described to the theory underlying the behaviors that are relevant for general, non-orthogonal designs, and introduces new prior distributions for imposing differential shrinkage in this setting. The priors rely on identifying blocks of related predictors that can be prioritized in terms of their relationship with the response. We discuss strategies for analysis that are robust to these modeling choices.

E0957: **Log-linear Bayesian additive regression trees for multinomial logistic and count regression**
Presenter: **Jared Murray**, University of Texas at Austin, United States
Bayesian additive regression trees (BART) have been applied to nonparametric mean regression and binary classification problems in a range of applied areas. To date BART models have been limited to models for Gaussian "data", either observed or latent, and with good reason - the

Bayesian backfitting MCMC algorithm for BART is remarkably efficient in Gaussian models. But while many useful models are naturally cast in terms of observed or latent Gaussian variables, many others are not. We extend BART to a range of log-linear models including multinomial logistic regression and count regression models with zero-inflation and overdispersion. Extending to these non-Gaussian settings requires a novel prior distribution over BART's parameters. Like the original BART prior, this new prior distribution is carefully constructed and calibrated to be flexible while avoiding overfitting. With this new prior distribution and some data augmentation techniques we are able to implement an efficient generalization of the Bayesian backfitting algorithm for MCMC in log-linear BART models. We demonstrate the utility of these new methods with several examples and applications.

E1092: **Functional BART**
Presenter: **Carlos Carvalho**, The University of Texas at Austin, United States

The aim is to introduce functional BART, a new approach for functional response regression–that is, estimating a functional mean response $f(t)$ that depends upon a set of scalar covariates x. Functional BART, or funBART, is based on the Bayesian Additive Regression Trees (BART) model. The original BART model is an ensemble of regression trees; funBART extends this model to an ensemble of functional regression trees, in which the terminal nodes of each tree are parametrized by functions rather than scalar responses. Just like the original BART model, funBART offers an appealing combination of flexibility with user-friendliness: it captures complex nonlinear relationships and interactions among the predictors, while eliminating many of the onerous "researcher degrees of freedom" involved in function-on-scalar regression using standard tools. In particular, functional BART does not require the user to specify a functional form or basis set for $f(t)$, to manually choose interactions, or to use a multi-step approach to select model terms or basis coefficients. Our model replaces all of these choices by a single smoothing parameter, which can either be chosen to reflect prior knowledge or tuned in a data-dependent way.

CI015 Room A0 FINANCIAL TIME SERIES Chair: Alessandra Amendola

C0168: **Multiplicative nonstationary volatility models with exogenous information**
Presenter: **Cristina Amado**, University of Minho, Portugal
Co-authors: Timo Terasvirta

A multiplicative nonstationary volatility model allowing for nonlinear behaviour driven by exogenous information is proposed. The new model extends the time-varying GARCH model by including an additional stochastic variable to allow the conditional variance to change smoothly between regimes. Modelling strategies for the proposed model are developed, and they rely on Lagrange multiplier tests. The estimation of the model is simplified by employing maximisation by parts and the asymptotic properties of the proposed estimators are also studied. Finite-sample properties of these procedures and statistical tests are examined by simulation. An empirical application illustrates the functioning of the model in practice.

C0169: **Modelling dynamic covariance matrices with stochastic volatility latent factors**
Presenter: **Roxana Halbleib**, University of Konstanz, Germany
Co-authors: Giorgio Calzolari

A new method is proposed to model and forecast large dimensional covariance matrices of daily returns by taking advantage of the commonality in their dynamics by means of a latent factor structure with stochastic volatility. As such a model is very difficult to estimate from daily returns, we use the richer information content of intraday returns incorporated in realized covariance matrices, which are consistent estimates of the multivariate daily variation. Our stochastic volatility latent factor model is able to capture the empirical features of daily covariance matrices, such as commonality in dynamics and long persistence in autocorrelation, within a very parsimonious framework: the number of the parameters is of order $O(n)$ compared to the ones of existing multivariate dynamic volatility approaches, which are of order at least $O(n^2)$. The proposed model has a non-Gaussian non-linear state-space representation; we estimate it my means of exact numerical maximum-likelihood using the non-Gaussian filtering approach previously proposed. We prove the accuracy of the parameter estimates within a Monte Carlo study and the usefulness of our approach to forecast high-dimensional covariance matrices within an empirical application to daily realized covariance matrices of the DJIA components.

C0170: **Realized estimators of tail risk measures**
Presenter: **Giuseppe Storti**, University of Salerno, Italy
Co-authors: Ostap Okhrin

A novel estimation method for the quantiles and tail expectation of the distribution of financial returns that exploits information on realized higher order moments built from intra-daily data is proposed. Building on recent results on the joint elicitability of VaR and ES, our approach can be sees as a data driven generalization of standard asymptotic expansions such as the Cornish-Fisher one. The proposed procedure can be used to generate realized proxies of conditional VaR and ES as well as it can be extended to generate predictions of the future values of these risk measures. The empirical performance of the proposed methods will be assessed via Monte Carlo simulations and applications to real stock market data.

CO372 Room A2 MACROECONOMIC FORECASTING Chair: Simon van Norden

C1119: **Business cycle asymmetry and unemployment rate forecasts**
Presenter: **Simon van Norden**, HEC Montreal, Canada

Asymmetries in unemployment dynamics have been observed in the time series of a number of countries, including the United States. Asymmetries in unemployment rate forecast errors are studied. We consider conditions under which optimal forecasts will display asymmetrically-distributed errors and how the degree of asymmetry might vary with forecast horizon. Using data from the U.S. Survey of Professional Forecasters and the Federal Reserve Greenbook, we find substantial evidence of forecast error asymmetry, which tends to increase with the forecast horizons; we also find noteworthy differences in forecasts from these two sources. The results give insight into the ability of professional forecasters to adapt their forecasts to asymmetry in underlying processes.

C1158: **Changes in predictability of the Canadian economy: Evidence from the Bank of Canada staff's forecast**
Presenter: **Rodrigo Sekkel**, Bank of Canada, Canada
Co-authors: Julien Champagne, Guillaume Poulin-Bellisle

An evaluation of the Bank of Canada staff's forecasts for real GDP growth and CPI inflation since 1982 is provided by using a novel database of real-time data and forecasts. We compare the staff's forecasts with commonly-used time series models estimated with real-time data, and study changes in predictability of the Canadian economy following the announcement of the inflation targeting regime in 1991. While a large literature has examined the forecasting performance of the U.S. Federal Reserve staff, few papers have focused on small open economies. Our evidence is particularly interesting, as it includes over 30 years of staff's forecasts, two severe recessions and different monetary policy regimes.

C1179: **The trend unemployment rate in Canada**
Presenter: **Pierre St-Amant**, Bank of Canada, Canada
Co-authors: Marie-Noelle Robitaille, Laurence Savoie-Chabot, Dany Brouillette
The unemployment rate is an important variable for monetary policy. A low (high) unemployment rate suggests that conditions in the labour market are tight (easy), which tends to generate upward (downward) pressures on wages and consumer prices index inflation. But the unemployment rate is only high or low when compared with some reference value. The authors call this value the trend unemployment rate (TUR). Their objective is to determine if a TUR useful for the conduct of Canadian monetary policy can be identified. Various approaches have been proposed to measure the TUR. A frequently chosen approach uses the information provided by inflation within a reduced form Phillips curve. A different approach is to develop models with variables (e.g. payroll taxes and unionization rate) thought to determine the TUR. We consider methods following these two approaches. They assess them based on three criteria. A first criterion is that TUR estimation methods should provide explanations for changes in trend unemployment. A second criterion is that revisions to TUR estimates need to be well-behaved. A third criterion is that a UGAP should help forecast inflation. They use real-time data for our assessment of conformity with the second and third criteria.

C1471: **Evaluating the use of real-time data in forecasting output levels and recessionary events in the US**
Presenter: **Kevin Lee**, University of Nottingham, United Kingdom
Co-authors: Chrystalleni Aristidou, Kalvinder Shields
A modelling framework and evaluation procedure is proposed to judge the usefulness of real-time datasets incorporating past data vintages and survey expectations in forecasting. The analysis is based on 'meta models' obtained using model-averaging techniques and judged by various statistical and economic criteria, including a novel criterion based on a fair bet. Analysing US output data over 1968q4-2015q1, we find both elements of the real-time data are useful with their contributions varying over time. Revisions data are particularly valuable for point and density forecasts of growth but survey expectations are important in forecasting rare recessionary events.

CO066 Room B2 REGIME CHANGE MODELING I Chair: Willi Semmler

C1280: **Testing for forecast rationality under Markov switching**
Presenter: **Florens Odendahl**, Banque de France, France
Co-authors: Barbara Rossi, Tatevik Sekhposyan
Novel tests are proposed for forecast rationality robust to the presence of discrete and recurring switches in forecasting ability, such as Markov switching. Existing forecast rationality tests robust to instabilities are based on non-parametric techniques; relative to the latter, our tests perform better in the presence of discrete switches, rather than smooth changes, under the alternative. Monte Carlo simulations suggest that our tests have better power than existing tests in detecting Markov switching deviations from unbiasedness or efficiency. We investigate whether Blue Chip Financial Forecasts for the Federal Funds target rate are rational. Our test finds evidence against forecast unbiasedness and uncovers the fact that forecasters tend to systematically overestimate the future interest rate especially during periods of monetary easing. The size of the systematic bias component is around 25 basis points.

C1250: **Climate disaster risk, regime switching and monetary policy**
Presenter: **Willi Semmler**, New School for Social Research, United States
Co-authors: Stefan Mittnik
The macroeconomic and financial market impact of rare large disasters has since the Great Recession of the years 2007/9 been studied in much literature. Those disasters are seen to have both large real effects (capital and output losses) and financial market effects (shift of discount rates, risk premia, asset prices, and returns). Such effects of large financial crashes studied in financial economics are also shown to be essential for climate-related rare disasters. Insights of the former are used to study the effects of disaster risks in the macroeconomics of climate change. The empirics of disaster risks in climate economics shows a link between GDP growth, greenhouse gas emission, climate change and climate-related disasters. The proposed model uses calibrated rare large disaster shocks and their effects on output and capital losses and rising risk premia in a multi-phase dynamic decision model. We build on previous works which develop such a multi-phase decision model. The model is solved via the non-linear programming method AMPL which is augmented by an arc parametrization method (APM). The proposed method can deal with regime shifts, arising from large disaster shocks, in a multi-regime model. Such a model is also suitable to include fiscal and financial policies to address the issue of mitigation of and adaptation to disaster risks.

C1441: **Towards a fairer distribution of carbon tax across generations in the DICE model through green bonds**
Presenter: **Sergey Orlov**, International Institute for Applied Systems Analysis, Laxenburg and Lomonosov Moscow State University, Russia
Co-authors: Elena Rovenskaya, Willi Semmler
The introduction of fiscal instruments into DICE model is considered in order to smooth the transition to low-carbon economy and to improve the intergenerational fairness. Namely, we introduce green bonds and taxation for possibility of wealth redistribution in order to pay for the mitigation efforts. The two modeling assumptions are explored: the internal debt and the external debt. While modeling internal debt, we use a portfolio approach for households that allows for them to choose the optimal proportion between bonds and capital investments. We obtain via theoretical investigation and model simulation that the external debt may help improvement of intergenerational fairness with respect to optimal DICE scenario. The internal debt does not however lead to improvement of intergenerational fairness but provides the incentives for investing in CO2 abatement and helps to smooth a carbon tax.

C1269: **Portfolio under-diversification: Equity sector bias**
Presenter: **Ibrahim Tahri**, PIK (Potsdam Institute for Climate Impact Research), Germany
A potential motive behind investors' preference for holding fossil fuel (FF) assets relative to clean energy (RE) assets might stem from an asymmetry of information vis-a-vis potential payoffs of RE sectors compared those of the FF sectors or/and simply an underestimation of the risks of carbon assets depreciation. This suggests there might exist a bias in the portfolio choice of the investor. An underdiversified portfolio can be rationalized through a combined learning-investment model. A key characteristic in this type of models is the role of endogenous information choice in creating bias, where thanks to some general equilibrium forces there are increasing returns to information, which in turn leads to full specialization in learning and influences the optimal asset allocations of the investor. We attempt to provide a theoretical support to describe the potential presence of sector equity bias, similar to the home equity bias puzzle, in the energy sector. For this purpose, we rely on a new strand of literature which combines information theory to portfolio choice theory that has been used to explain the portfolio bias.

CO380 Room D2 ECONOMETRICS OF NETWORK MODELS WITH APPLICATIONS Chair: Laurent Pauwels

C1289: **Modelling latent network stochastic volatility spillovers**
Presenter: **Laurent Pauwels**, University of Sydney, Australia
Co-authors: Michael McAleer, Manabu Asai

Volatility spillovers and linkages of financial portfolios are modelled by using novel latent network stochastic volatility (NetSV) models that capture the latent linkages across the financial assets. Financial theory points to latent information linkages as the origin of volatility spillovers. These linkages are created from common information, which impacts the expectations of financial traders, and also information spillovers that arise from their hedging behaviour. The theory is extended to network and stochastic volatility models, which are assumed to be random and latent. The networks provide an interpretable and tractable model of the volatility linkages. New Bayesian algorithms are developed to identify latent random networks within the context of multivariate stochastic volatility models. Upon identifying the network, the volatility spillover effects across markets and dynamic optimal hedge ratios can be estimated and tested statistically. The latent network approach reduces the estimation burden of the spillover effects that are typically encountered in multivariate volatility models with high-dimensional parameters.

C1310: **Identification and estimation of a partially linear regression model using network data**
Presenter: **Eric Auerbach**, Northwestern University, United States

A regression model is studied in which one covariate is an unknown function of a latent driver of link formation in a network. Rather than specify and fit a parametric network formation model, a new method is introduced based on matching pairs of agents with similar columns of the squared adjacency matrix, the ij-th entry of which contains the number of other agents linked to both agents i and j. The intuition behind this approach is that for a large class of network formation models the columns of this matrix characterize all the identifiable information about individual linking behavior. We first describe the model and formalize this intuition. We then introduce estimators for the parameters of the regression model and characterize their large sample properties.

C1388: **Modeling social networks using linear preferential attachment**
Presenter: **Phyllis Wan**, Erasmus University Rotterdam, Netherlands
Co-authors: Tiandong Wang, Richard Davis, Sid Resnick

Preferential attachment is an appealing mechanism for modeling power-law behavior of degree distributions in social networks. We consider fitting a directed linear preferential attachment model to network data under three data scenarios: 1) When the full history of the network growth is given, MLE of the parameter vector and its asymptotic properties are derived. 2) When only a single-time snapshot of the network is available, an estimation method combining method of moments with an approximation to the likelihood is proposed. 3) When the data are believed to have come from a misspecified model or have been corrupted, a semi-parametric approach to model heavy-tailed features of the degree distributions is presented, using ideas from extreme value theory. We illustrate these estimation procedures and explore the usage of this model through simulated and real data examples.

C1406: **Quantile connectedness: Modelling tail behaviour in the topology of financial networks**
Presenter: **Matthew Greenwood-Nimmo**, University of Melbourne, Australia
Co-authors: Tomohiro Ando, Yongcheol Shin

A new technique is developed to estimate vector autoregressions by quantile regression. A factor structure is used to remove cross-section correlation in the residuals such that the system can be estimated on an equation-by-equation basis using existing quantile regression toolboxes. We use our model to study credit risk spillovers among a panel of 18 sovereigns and their respective financial sectors between January 2006 and February 2012. We show that idiosyncratic credit risk shocks do not propagate strongly at the median but that powerful spillovers occur in both tails. Furthermore, rolling sample analysis reveals marked time-varying tail-dependence. These important features of credit risk transmission are obscured in models estimated using conventional conditional mean estimators.

CO166 Room E2 NEW METHODS FOR HEAVY TAILS, COPULAS AND CRYPTOCURRENCIES Chair: Artem Prokhorov

C1180: **Cryptocurrencies: Intrinsic value, bubbles and heavy-tailedness**
Presenter: **Rustam Ibragimov**, Imperial College London and Innopolis University, United Kingdom
Co-authors: Christine Parlour, Johan Walden

A parsimonious model is developed for the price of a virtual currency. In the model, the virtual currency's price depends on the surplus it generates by decreasing frictions of trade, its exposure to sudden negative shocks, e.g., in the form of regulation, and potentially the presence of rational bubbles. Price and trading volume dynamics of the virtual currency differ in several important ways from those of stocks and other asset classes. Importantly, the model implies heavy-tailed power law distributions for the virtual currency's price in the case when it contains a bubble component, and semi-heavy-tailed log-normal distributional tails when the price is based on intrinsic value. In empirical tests, the recent market dynamics of Bitcoin and other cryptocurrencies are well explained by our model.

C1614: **A CUSUM test for tail behavior of GARCH(1,1) models**
Presenter: **Eunju Hwang**, Gachon University, Korea, South
Co-authors: JunHyeong Kim

A GARCH model has been a very popular time series model for the volatility of financial returns. A CUSUM test is proposed for detecting the tail behavior of a stationary GARCH(1,1) model, more particularly, for testing whether the tail index of the model is changed or not. The CUSUM test statistic is constructed using the empirical distribution function with sample extremes and its limiting distribution is shown to be a Brownian bridge. The proof is based on the weak dependence structure and on the existence of the phantom distribution function of the stationary GARCH model. This test can be used for general weakly-dependent time series models. A Monte-Carlo study is conducted to see the performance of power and size of the CUSUM test in GARCH(1,1) models with heavy-tailed noises, adopting tail index of the noises to be changed. Real data applications are given with financial data such as KOSPI.

C1336: **A new approach to credit rating**
Presenter: **Artem Prokhorov**, University of Sydney and Innopolis University, Australia
Co-authors: Stan Uryasev, Giorgi Pertaia

Current credit ratings are shown to be inadequate, as they fail to capture the tail mass of the risk distribution. We propose a new approach based on the concept of buffered probability of exceedance.

C1749: **Investment approach to the blockchain market segments**
Presenter: **Tomasz Slonski**, WrocÅĆaw University of Economics, Poland
Co-authors: Aleksander Mercik

The primary goal is to divide digital token market into different sectors classified according to types of application. The new classification

recognizes the underlying value of a token which is based on a function it serves in particular application which create the base for detailed financial analysis of all sectors. Consequently, sectors are put under examination with respect to their size, market trends and unique investment features (volatility and rates of return). The novelty of our research is to propose up-to-date blockchain project classification and perform market segments analysis in its entirety rather than focusing on cryptocurrency leaders. We carefully scrutinize 350 blockchain projects based on Ethereum platform and group them into 21 sectors. The analysis of investment features of each sector designates the most efficient sectors: blockchain infrastructure, media and publishing and ether itself. Additionally, in some sectors series of returns are characterized by autocorrelation which indicates inefficiency and calls for more refined construction of an investment portfolio. We find significant correlation between rates of return of different sectors, which confirms that investors can decompose total risk into markets systematic risk and specific risk of each sector. The overall result suggests that qualified, active investors can find many opportunities on growing blockchain market.

CO212 Room F2 MACRO-FINANCIAL LINKAGE Chair: Wenying Yao

C0946: **The information content of inflation swap rates for the long-term inflation expectations of professionals**
Presenter: **Ahmed Hanoma**, Free University Berlin, Germany
Co-authors: Dieter Nautz
Long-term inflation expectations taken from the survey of professional forecasters are a major source of information for monetary policy. Unfortunately, they are published only on a quarterly basis. The daily information content of inflation-linked swap rates for the next survey outcome is investigated. Using a mixed data sampling approach, we find that professionals account for the daily dynamics of inflation swap rates when they submit their long-term inflation expectations. We propose a daily indicator of professionals' inflation expectations that outperforms alternative indicators that ignore the high-frequency dynamics of inflation swap rates. To illustrate the usefulness of the new indicator, we provide new evidence on the re-anchoring of U.S. inflation expectations.

C0746: **Signed spillover effects building on historical decompositions**
Presenter: **Vladimir Volkov**, University of Tasmania, Australia
Co-authors: Mardi Dungey, Pierre Siklos
The spillover effects of interconnectedness can be further decomposed into both the sources of shocks and whether they amplify or dampen volatility conditions in the target market. We show how to use historical decompositions to rearrange the information from a VAR to include the sources, direction and signs of spillover effects building on the unsigned forecast error variance decomposition approach. We apply the methodology to a panel of CDS spreads of sovereigns and financial institutions for the period 2003-2013 and identify how these entities contribute to global systemic risk.

C0884: **The role of market indices in forecasting stocks volatility: A HAR framework using a mixed sampling approach**
Presenter: **Marwan Izzeldin**, Lancaster University Management School, United Kingdom
Co-authors: Ingmar Nolte, Vasileios Pappas, Rodrigo Hizmeri
The aim is to examine the value added in forecasting high-frequency stock data using a Heterogeneous Autoregressive (HAR) model augmented with market indices (SPY and the S&P 500). Our empirics are based on high-frequency data of 10 representative stocks, the S&P 500 and SPY market indices for the period 2000 to 2016. We allow for different sampling frequencies on both sides of the HAR regression specification, different market regimes and signed realised variances and covariances. We find that, irrespectively of the specification adopted, the Market-Augmented HAR (*M-HAR*) specification always brings significant forecasting gains over the conventional HAR. Despite the high correlation between the S&P 500 and SPY realised variances and covariances, both indices have different statistical features and patterns (i.e. periodicity, persistence, continuity and leverage) that results in differentiated forecasting gains. The S&P 500 adds more than SPY at all sampling frequencies. Choosing an optimal sampling frequency is essential to maximise gains and these tend to vary across with the index in use. The gains from *M-HAR* specification are regime sensitive where the highest gains are observed in the pre and post-crisis regimes. Adding the index during the crisis episode adversely affects the forecasts.

C0630: **Testing the rank of cojumps in high-frequency data with market microstructure noise**
Presenter: **Wenying Yao**, Deakin University, Australia
Co-authors: Lars Winkelmann
A test for the rank of a cojump matrix is proposed. The matrix consists of estimated jump sizes at specific intraday time points. High-frequency trading and the market microstructure can distort a rank test. To obtain noise-robust statistics, we use a pre-average estimator of the cojump matrix. We derive an asymptotic chi-square test for its rank. Simulation shows that the proposed test has good size and power properties. We use the new test to investigate the factor structure of cojumps at macroeconomic announcement times.

CO336 Room G2 EMPIRICAL APPLICATIONS IN ECONOMICS AND FINANCE Chair: Michael Ellington

C0259: **Stock market volatility spillovers: Evidence for Latin America**
Presenter: **Luis Melo**, Banco de la Republica, Colombia
Co-authors: Santiago Gamba, Jose Eduardo Gomez, Jorge Hurtado
A previous framework is extended to construct volatility spillover indexes using a DCC-GARCH framework to model the multivariate relationships of volatility among assets. We compute spillover indexes directly from the series of asset returns and recognize the time-variant nature of the covariance matrix. Our approach allows for a better understanding of the movements of financial returns within a framework of volatility spillovers. We apply our method to stock market indexes of the United States and four Latin American countries. Our results show that Brazil is a net volatility transmitter for most of the sample period, while Chile, Colombia and Mexico are net receivers. The total spillover index is substantially higher between 2008Q3 and 2012Q2, and shock transmission from the United States to Latin America substantially increased around the Lehman Brothers episode.

C0263: **Interval prediction of electricity prices: A robust approach**
Presenter: **Luigi Grossi**, University of Verona, Italy
Co-authors: Lisa Crosato, Fany Nan
A doubly robust approach is introduced in order to model the volatility of electricity spot prices, minimizing the misleading effects of the extreme jumps that characterize this particular kind of data on the predictions. With respect to the mainstream literature on electricity price forecasting, which highlights the importance of predicting spikes, the attention is moved to the correction of the impact that spikes have on the estimation of the prices and, in particular, on their volatility. Volatility of electricity prices has often been estimated through GARCH type models which can be strongly affected by the presence of extreme observations. Although the presence of spikes is a well-known stylized effect observed on electricity markets, robust volatility estimators have not been so far applied. We try to fill this gap by suggesting a robust procedure to the study of the dynamics of electricity prices. The conditional mean of de-trended and seasonally adjusted prices is modeled though a robust estimator of SETAR processes based on a polynomial weighting function, while a robust GARCH is used for the conditional variance. The robust GARCH estimator

relies on the extension of the forward search. The robust SETAR-GARCH model is applied to the Italian electricity markets using data in the period spanning from 2013 to 2015.

C0349: **The institutional blockholders influence on corporate investment: Evidence from emerging markets**
Presenter: **Carlos Pombo**, Universidad de los Andes, Colombia
Co-authors: Mauricio Jara-Bertin
The relationship between firm investment ratios and institutional blockholders for a sample of 6,300 publicly traded firms in 16 large emerging markets for the 2004-2014 period is examined. Results show that independent, long-term, and local institutional investors boost investment ratios, which is consistent with the monitoring role and blockholder voice intervention hypotheses. The presence of institutional blockholders, regardless of their monitoring involvement, reduces firm cash flow sensitivity ratios and thus reduces firms financial constraints. Minority institutional investors complement the positive effect of blockholders investors. However, the effect on financial constraints decreases as the quality of the country's institutions increases.

C0555: **The effects of productivity shocks and job destruction in a changing world**
Presenter: **Michael Ellington**, University of Liverpool, United Kingdom
Co-authors: Chris Martin, Bingsong Wang
A "second generation" Bayesian time-varying parameter VAR model with stochastic volatility, combined with a simple search frictions model of the labour market, is used to explore the changing relationships between labour productivity, unemployment, vacancies and real wages using US data from 1962-2016. We find marked changes in labour market linkages suggesting that the key mechanisms vary throughout time. One key finding is that the responsiveness of wages to identified shocks gradually increases throughout our sample; whilst the sensitivity of unemployment and vacancies to these shocks remains stable. We show that a simple search frictions model is unable to match our empirical results when changing structural parameters. This suggests that the search frictions model requires augmenting in order to capture true US labour market dynamics.

CO216 Room H2 TERM STRUCTURE OF INTEREST RATES Chair: Laura Coroneo

C0320: **The negative interest rate policy and the yield curve**
Presenter: **Dora Xia**, Bank for International Settlements, Switzerland
Co-authors: Jing Cynthia Wu
The aim is to extract the market's expectations about the ECB's negative interest rate policy from the euro area's yield curve and to study its impact on the yield curve. To capture the rich dynamics taking place at the short end of the yield curve, we introduce two policy indicators that summarise the immediate and longer-horizon future monetary policy stances. The ECB has cut interest rates four times under zero. We find that the June 2014 and December 2015 cuts were expected one month ahead but that the September 2014 cut was unanticipated. Most interestingly, the March 2016 cut was expected four months ahead of the actual cut.

C0515: **The information in the joint term structures of bond yields**
Presenter: **Andrew Meldrum**, Board of Governors of the Federal Reserve System, United States
While standard no-arbitrage term structure models are estimated using nominal yields from a single country, a growing literature estimates joint models of yields in multiple countries or nominal and real yields from a single country. However, it is argued that in two of the most common applications joint modeling appears to be unnecessary. Joint models of U.S. and German nominal yields do not offer economically significant advantages in fitting the cross-section of yields, or predicting future yields. We obtain similar results for joint models of U.S. nominal and real yields. Thus, we lose little if we simply estimate separate models of those yields.

C0689: **Predicting interest rates in real-time**
Presenter: **Laura Coroneo**, University of York, United Kingdom
The aim is to analyse the predictive ability of real-time macroeconomic information for the yield curve of interest rates. We specify a mixed-frequency real-time macro-yield model that incorporates interest rate surveys and that treats macroeconomic factors as unobservable components that we extract simultaneously with the yield curve factors. Using U.S. data from 1972 to 2016, we find that real-time macroeconomic information was helpful to predict the yield curve of interest rates up to December 2008 but, after this date, interest rate surveys have stronger predicting ability.

C0329: **A flexible short-rate based four factor arbitrage-free term structure model with an explicit monetary policy rule**
Presenter: **Ken Nyholm**, European Central Bank, Germany
An arbitrage-free four-factor term structure model is derived which facilitates direct parameterization of the short-term interest rate process. The interplay between macroeconomic variables and the term structure via a monetary policy reaction function is therefore directly supported. We show that the proposed model is a constrained member of the canonical GDTSM family. The model's loading structure closely resembles a previous model, but it relies only on a single nonlinear shape parameter, and it is therefore easy to estimate. An empirical application to US data covering the period from 1961 to 2017 demonstrates that the proposed model fits yields well, and that an embedded policy rule, including industrial production and the inflation rate, is statistically significant and economically meaningful during this time-period.

CO625 Room I2 STRUCTURAL VAR MODELS Chair: Simone Maxand

C0707: **Identification of SVAR models via independent component analysis: A comparative study**
Presenter: **Alessio Moneta**, Scuola Superiore Sant'Anna, Italy
Co-authors: Gianluca Pallante
Independent Component Analysis (ICA) is a statistical method that transforms a set of random variables in least dependent linear combinations. Under the assumption that the observed data are mixtures of non-Gaussian and independent processes, ICA is able to recover the underlying components that generated the data, up to a scale and order indeterminacy. Its application to structural vector autoregressive (SVAR) models is straightforward because it allows to recover the impact of independent structural shocks on the observed series from the estimated residuals. We compare the performances of three different ICA techniques: fastICA algorithm, minimization of Cramer-von-Mises (CvM) distance, and minimization of distance covariance (DCov). We investigate through Monte Carlo experiments the ability of these procedures of recovering structural impulse response functions from a VAR model. Using a p-generalized normal distribution, we let the underlying processes approach or diverge from a Gaussian distribution. Our results suggest a relatively better performance of DCov, on average, when normality is approached. Despite the relatively larger bias of the estimates, fastICA algorithm is relatively more robust to distributional assumptions. We also present an empirical illustration using Japanese data to study the real effects of unconventional monetary policies.

C0803: **Causality assessment in panel vector autoregressive models: A novel approach based on structural VARs**
Presenter: **Helmut Herwartz**, Georg-August-University Goettingen, Germany
Structural shocks in dynamic systems are hidden and often identified with reference to a priori economic reasoning. Based on assumptions of heteroskedastic or non-Gaussian systems, recent advances in identifying structural vector autoregressive models point to the potential of data based identification schemes. In small samples data based identification might lack applicability of complicated algorithms, or might suffer from limited power of diagnostic testing. A panel approach is suggested to test specific causal structures for a cross section of economies that builds upon principles of Hodges Lehmann estimation.

C0885: **Set identification in non-Gaussian SVARs: A refinement of the sign restriction approach**
Presenter: **Simone Maxand**, University of Helsinki, Finland
Co-authors: Helmut Herwartz
In identifying structural vector autoregressive (SVAR) models, theory based sign restrictions have become a prominent means to reduce a universe of potential models to a set of structural relations which fulfil supposedly consensual economic beliefs. Typically, sign identified SVARs are obtained from tracing the symmetric effects of unit impulses hitting a system of interest in isolation. Noticing an important recent literature that targets at the identification of unique non-Gaussian structural shocks, set identification by means of isolated impulses is at the risk to neglect higher order dependence and possible asymmetries among orthogonalized model residuals. Accounting for higher order dependence in set identification of non-Gaussian SVARs, we (i) take advantage of conditional moment profiles to design 'stylized unit shocks', and (ii) rely on selected 'empirical shocks'.

C1003: **Estimating non-causal VAR using all-pass filters**
Presenter: **Bernd Funovits**, University of Helsinki, Finland
A new estimator for possibly non-causal structural VAR systems driven by non-Gaussian i.i.d. shocks is derived. Since the root location cannot be identified from the second order spectral density, third and fourth order (cumulant) spectral densities are used to conclude on whether the determinantal roots of the AR matrix polynomial are inside or outside the unit circle. Multivariate all-pass filters are used to obtain all possible combinations of root locations (based on an initial estimate). Subsequently, higher order spectra are used to conclude on the root location. Measures of estimation uncertainty are derived through bootstrapping methods. The estimation procedure is implemented in the R-package varAllPass.

CO190 Room M2 ECONOMICS OF CRYPTOCURRENCIES Chair: Marco Lorusso

C0752: **Forecasting cryptocurrencies under model and parameter instability**
Presenter: **Stefano Grassi**, University of Rome 'Tor Vergata', Italy
Co-authors: Leopoldo Catania, Francesco Ravazzolo
The predictability of cryptocurrencies time series is studied. We compare several alternative univariate and multivariate models in point and density forecasting of four of the most capitalized series: Bitcoin, Litecoin, Ripple, and Ethereum. We apply a set of crypto predictors and rely on dynamic model averaging to combine a large set of univariate dynamic linear models and several multivariate vector autoregressive models with different forms of time variation. We find statistically significant improvements in point forecasting when using combinations of univariate models and in density forecasting when relying on the selection of multivariate models. Both schemes deliver sizeable directional predictability.

C0775: **Cryptocurrencies and monetary policy**
Presenter: **Marco Lorusso**, Northumbria University, United Kingdom
Co-authors: Francesco Ravazzolo
A Dynamic Stochastic General Equilibrium (DSGE) model is developed to evaluate the economic repercussions of cryptocurrencies. We assume that the representative household maximizes its utility given by consumption, leisure and both government currency and cryptocurrency holdings. The model includes entrepreneurs that determine the supply of cryptocurrency in the economy. We also consider a central bank setting the nominal interest rate following a general augmented Taylor-type interest-rate rule. In particular, the nominal rate responds not only to the interest rate in the previous period and to deviations of output and inflation from their steady-state values, but also to nominal money growth in government currency and cryptocurrency. We calibrate our model using US data. Our impulse response analysis shows the effects of a "traditional" shock to household's demand for real balances of government currency as well as to a "new" shock to household's demand for real balances of cryptocurrency. Moreover, we evaluate the response of main macroeconomic fundamentals to productivity shocks for production of cryptocurrency. Finally, we quantify the importance of the shocks demand/supply shocks of cryptocurrecny through a variance decomposition analysis.

C1295: **Quantitative risk management for cryptocurrencies**
Presenter: **Francesco Ravazzolo**, Free University of Bozen/Bolzano, Italy
Co-authors: Leopoldo Catania, Stefano Grassi
Cryptocurrencies have recently gained a lot of interest from investors, central banks and governments worldwide. The lack of any form of political regulation and their market far from being efficient, requires new forms of regulation in the near future. From an econometric viewpoint, the process underlying the evolution of the cryptocurrencies' volatility has been found to exhibit at the same time differences and similarities with other financial time-series, e.g. foreign exchanges returns. We analyse how quantitative risk management techniques need to be implemented when dealing with cryptocurrencies time-series. We focus on the estimation and backtesting of the Value-at-Risk (VaR) and Expected Shortfall (ES) risk measures and report advices for quantitative risk managers and investors. Our results indicate that naive approaches generally used by practitioners, like variance estimation via exponential smoothing, can be extremely dangerous when dealing with cryptocurrencies.

C1300: **Quantity theory of money and cryptocurrencies**
Presenter: **Ladislav Kristoufek**, Institute of Information Theory and Automation, Czech Academy of Sciences, Czech Republic
Dynamics of major cryptocurrencies is studied in the light of two fundamental economic laws - the law of one price and the equation of exchange. We perform the analysis in two steps. First, utilizing the unprecedented data availability of Bitcoin (and cryptocurrencies in general) statistics, we are able to construct its theoretical appreciation with respect to the US dollar. If the (hard-) fork of March 2013 is taken as a starting point of Bitcoin stability, our results suggest that the cryptocurrency was well within the theory-implied appreciation until the end of 2016. It is only the time starting in 2017 that can be considered as a period of large deviation from fundamental equilibrium dynamics implied by the economic laws. And second, we build on these promising results and utilize the economic laws for the relationship between Bitcoin and other major cryptocurrencies. We show that most of the currency pairs are well described within the logic of the economic laws. Both these main results can be seen as evidence of existing fundamental value of cryptocurrencies.

CO552 Room N2 TIME SERIES ECONOMETRICS II **Chair: Josu Arteche**

C0501: **The estimation and testing of the fractional cointegration order based on the frequency domain: A robust approach**
Presenter: **Valderio Anselmo Reisen**, DEST-CCE-UFES, Brazil
Co-authors: Igor Souza, Glaura Franco, Pascal Bondon

The aim is to estimate the degree of cointegration in bivariate series. A test statistic for the non-cointegration based on the determinant of the spectral density matrix for the frequencies close to zero is proposed. Series are assumed to be $I(d)$, $0 < d < 1$, with parameter d supposed to be known. In this context, the order of integration of the error series is $I(d-b)$, $b \in [0,d]$. The proposed estimator for b is obtained by performing a regression of logged determinant on a set of logged Fourier frequencies. Under the null hypothesis of non-cointegration, the expressions for the bias and variance of the estimator were derived and its consistency property was also obtained. The asymptotic normality of the estimator, under Gaussian and non-Gaussian innovations, was also established. The robustness to the presence of outliers is also addressed using M-periodograms. Their performance was investigated using Monte Carlo simulations.

C0971: **Modelling persistence change in fractionally integrated models**
Presenter: **Luis Filipe Martins**, ISCTE-IUL, Portugal
Co-authors: Josu Arteche

In recent years a vast literature documenting changes in the historical behaviour of economic and financial time series has been put forward. The popular parsimonious long-memory ARFIMA model describes both short and long memories simultaneously. There has been proposed parametric local stationary long-memory models. A new approach is proposed to model persistence change in fractionally integrated models. The model's statistical properties, estimation and inference is also studied.

C0915: **A model for count time series with periodic two orders autoregressive structure**
Presenter: **Pascal Bondon**, CentraleSupelec, France
Co-authors: Paulo Prezotti, Valderio Anselmo Reisen, Marton Ispany, Faradiba Sarquis

A new model for count time series with conditional Poisson or geometric distribution with a periodic two orders autoregressive structure is introduced. This model is an extension of the Periodic Integer Autoregressive model of order 1 (PINAR(1) model). Stochastic properties of the model such as mean, variance, marginal and joint distributions are discussed. Moment-based and conditional maximum likelihood estimates of the parameters are presented. An alternative numerical estimation procedure, which involves less computational effort, is proposed and its performance is investigated through Monte Carlo simulations. The usefulness of the model is illustrated by an application to real data set.

C0675: **Forecasting a latent variable: Application to VaR in stochastic volatilty models**
Presenter: **Josu Arteche**, University of the Basque Country UPV/EHU, Spain
Co-authors: Javier Garcia

The existence of latent variables contaminated with an added noise is quite common in many areas, a typical example being the volatility component in Stochastic Volatility (SV) models. Prediction of these latent variables can be either in-sample, or, in a more etymological sense, out-of-sample. The former is related with signal extraction whereas the latter implies prediction of future values. We focus on out-of-sample predictions, analysing several forecasting techniques for prediction of latent variables. There is a controversy over whether these techniques should be applied on the contaminated series of observables or on in-sample predictions of the latent variable obtained by signal extraction. Some light is shed on this issue by implementing a Monte Carlo analysis with different forecasting techniques applied in models with low frequency behaviour, which are common in SV modelling. Their applicability to forecast the volatility in Stochastic Volatility models and its use for Value at Risk evaluation is also examined by an application to a daily series of SP500 returns.

Saturday 15.12.2018 16:20 - 18:00 Parallel Session I – CFE-CMStatistics

EI005 Room Sala Convegni GRAPHICAL AND GEOMETRICAL STATISTICS Chair: Miguel de Carvalho

E0999: **Visualising and exploring models interactively with Shiny**
Presenter: **Catherine Hurley**, Maynooth University, Ireland
Two new interactive tools for model exploration are discussed. Both are implemented as Shiny-based R packages ERSA and condvis which are available on CRAN. ERSA is aimed at the student and teacher of linear regression models. It uses linked displays of model summary tables and drill-down parallel coordinate plots of case-based information for families of models. With a point and click interface, students explore the effects of dropping predictors and cases on model fits and their visualisations. The second package, condvis, provides a way of examining model behaviour for fits ranging from linear models, to Bayesian fits to black-box models. It displays fits on sections of data space, together with observed data near the section. Sections are either chosen interactively by the user, or calculated to reveal some feature of the model. Such visualisation offers a way of examining model behaviour, checking model stability and juxtaposing multiple model fits, leading to improved understanding and possibly discovery of improved fits.

E1237: **On the geometry of Bayesian inference**
Presenter: **Bradley Barney**, University of Utah, United States
Co-authors: Miguel de Carvalho, Garritt Page
A geometric interpretation to Bayesian inference is provided which allows the introduction of a natural measure of the level of agreement between priors, likelihoods, and posteriors. The starting point for the construction of our geometry is the observation that the marginal likelihood can be regarded as an inner product between the prior and the likelihood. A key concept in our geometry is that of compatibility, a measure which is based on the same construction principles as Pearson correlation, but which can be used to assess how much the prior agrees with the likelihood, to gauge the sensitivity of the posterior to the prior, and to quantify the coherency of the opinions of two experts. Estimators for all the quantities involved in our geometric setup are discussed, which can be directly computed from the posterior simulation output. Some examples are used to illustrate our methods, including data related to on-the-job drug usage, midge wing length, and prostate cancer.

EO599 Room Aula 4 NON-CONVEX OPTIMIZATION PROBLEMS IN STATISTICS Chair: Sahand Negahban

E0725: **From shallow to deep: Theoretical insights into training of neural networks**
Presenter: **Mahdi Soltanolkotabi**, University of Southern California, United States
Neural network architectures (a.k.a. deep learning) have recently emerged as powerful tools for automatic knowledge extraction from data, leading to major breakthroughs in applications spanning visual object classification to speech recognition and natural language processing. Despite their wide empirical use the mathematical success of these architectures remains a mystery. One challenge is that training neural networks correspond to extremely high-dimensional and nonconvex optimization problems and it is not clear how to provably solve them to global optimality. While training neural networks is known to be intractable in general, simple local search heuristics are often surprisingly effective at finding global/high quality optima on real or randomly generated data. We will discuss some results explaining the success of these heuristics. We will discuss results characterizing the training landscape of single hidden layer networks demonstrating that when the number of hidden units are sufficiently large then the optimization landscape has favorable properties that guarantees global convergence of (stochastic) gradient descent to a model with zero training error. Second, we introduce a de-biased variant of gradient descent called Centered Gradient Descent (CGD). We will show that unlike gradient descent, CGD enjoys fast convergence guarantees for arbitrary deep convolutional neural networks with large stride lengths.

E0732: **Optimal link prediction with matrix logistic regression**
Presenter: **Quentin Berthet**, University of Cambridge, United Kingdom
Co-authors: Nicolai Baldin
The problem of link prediction is considered based on partial observation of a large network, and on side information associated to its vertices. The generative model is formulated as a matrix logistic regression. The performance of the model is analysed in a high-dimensional regime under a structural assumption. The minimax rate for the Frobenius-norm risk is established and a combinatorial estimator based on the penalised maximum likelihood approach is shown to achieve it. Furthermore, it is shown that this rate cannot be attained by any (randomised) algorithm computable in polynomial time under a computational complexity assumption.

E0968: **Improved sample size conditions for non-convex matrix completion**
Presenter: **Xiaodong Li**, UC Davis, United States
Recent years have witnessed great advances in the statistical analysis of nonconvex optimization in applications of machine learning, statistics, signal processing, etc, and matrix completion is one of the most illustrative models to demonstrate diverse properties of nonconvex optimization approaches. Nonconvex optimization methods are usually favored in practice due to their computational conveniences and empirical successes, but in theory the sample size conditions established so far are usually much worse than those established for convex methods. We introduce two results that improve the state-of-the-art sample size conditions. First, we study nonconvex matrix completion from a perspective of assumption-free approximation: with no assumptions on the underlying positive semidefinite matrix in terms of rank, eigenvalues or eigenvectors, we established the low-rank approximation error based on any local minimum of the proposed objective function. As interesting byproducts, corollaries of our main theorem improve the state-of-the-art results for nonconvex matrix completion with no spurious local minima. Second, we improve the state-of-the-art sample size conditions for implicit regularization of vanilla gradient descents for nonconvex matrix completion.

E1142: **Recovering a hidden Hamiltonian cycle via linear programming**
Presenter: **Yihong Wu**, Yale University, United States
The problem of hidden Hamiltonian cycle recovery is introduced, where there is an unknown Hamiltonian cycle in an n-vertex complete graph that needs to be inferred from noisy edge measurements. The measurements are independent and distributed according to P_n for edges in the cycle and Q_n otherwise. This formulation is motivated by a problem in genome assembly, where the goal is to order a set of contigs (genome subsequences) according to their positions on the genome using long-range linking measurements. Computing the maximum likelihood estimate in this model reduces to the Traveling Salesman Problem (TSP). Despite the NP-hardness of TSP, we show that a simple linear programming (LP) relaxation, namely the fractional 2-factor (F2F) LP, recovers the hidden Hamiltonian cycle with high probability as $n \to \infty$ at the exact information-theoretic optimal threshold. Departing from the usual proof techniques based on dual witness construction, the analysis relies on the combinatorial characterization (in particular, the half-integrality) of the extreme points of the F2F polytope. Evaluation of the algorithm on real data shows improvements over existing approaches.

EO046 Room Aula 5 EMERGING TRENDS IN PREDICTIVE INFERENCE Chair: Bertrand Clarke

E0335: **Recent ideas in tree-based inference**
Presenter: **Lucas Mentch**, University of Pittsburgh, United States
Tree-based methods and their ensemble extensions remain a popular tool in the statistical machine learning domain. In addition to their demonstrated robust predictive accuracy, a variety of ad hoc tools are available to assist in understanding the model fit and underlying processes. In recent years, a flurry of theoretical developments investigating the consistency and asymptotic distributions of predictions from such methods has helped to pull these tools further within the domain of statistics. We will highlight a number of these developments and discuss how those results pave the way for more traditional statistical analyses to be performed within these normally black-box procedures. We focus on particular on generating confidence intervals for predictions, the development of formal hypothesis tests for variable importance, efficient variable screening procedures, as well as a recent proposal based on classical permutation tests that allows such procedures to scale to high-dimensional settings and to be performed at many locations throughout the feature space simultaneously. Simulation results and demonstrations on real data will also be provided.

E0499: **Random forest prediction intervals**
Presenter: **Dan Nettleton**, Iowa State University, United States
Co-authors: Haozhe Zhang, Joshua Zimmerman, Daniel Nordman
Random forests are among the most popular machine learning techniques for prediction problems. When using random forests to predict a quantitative response, an important but often overlooked challenge is the determination of prediction intervals that will contain an unobserved response value with a specified probability. We propose new random forest prediction intervals that are based on the empirical distribution of out-of-bag prediction errors. These intervals can be obtained as a by-product of a single random forest. Under regularity conditions, we prove that the proposed intervals have asymptotically correct coverage rates. Simulation studies and analysis of 60 real datasets are used to compare the finite-sample properties of the proposed intervals with quantile regression forests and recently proposed split conformal intervals. The results indicate that intervals constructed with our proposed method tend to be narrower than those of competing methods while still maintaining marginal coverage rates approximately equal to nominal levels.

E0967: **Quantifying genomic connectedness and whole-genome prediction accuracy using bootstrap aggregation sampling**
Presenter: **Gota Morota**, Virginia Polytechnic Institute and State University, United States
Prediction of complex traits has been a focus of quantitative genetics since the beginning of the 20th century. The advancement of whole-genome prediction has sparked a renewed interest in this topic. In statistics, connectedness is a measure germane to estimable comparisons. In the whole-genome prediction era, the concept of genetic connectedness can be extended to measure a connectedness level between training and testing sets. Recent studies have shown that connectedness across sets increased with the increasing proportion of related individuals, and this increase was associated with the improved accuracy of prediction. However, prior studies on comparing the degree of connectivity mainly used model-based formulas of prediction error variance computed from best linear unbiased prediction, leaving an open question about the possibility of computing empirical connectedness. Therefore, we derived prediction error variance by using bootstrap aggregation sampling and investigated the relationship between empirical connectedness measures and prediction accuracy in the cross-validation framework. We also demonstrated the potential of non-parametric relationship matrices to quantify genomic connectedness and prediction accuracy in the presence of non-additive gene actions.

E1378: **High dimensional discriminant analysis for structurally dependent data**
Presenter: **Taps Maiti**, Michigan State University, United States
Co-authors: Yingjie Li
Linear discriminant analysis (LDA) is one of the most classical and popular classification techniques. However, it performs poorly in high-dimensional classification. Many sparse discriminant methods have been proposed to make LDA applicable in high dimensional case. One issue of those methods is the structure of the covariance among features is ignored. We propose a new procedure for high dimensional discriminant analysis for structurally correlated data. Specifically, we will discuss spatially structured data. Penalized maximum likelihood estimation (PMLE) is developed for feature selection and parameter estimation. Tapering technique is applied to reduce computation load. The theory shows that the method proposed can achieve consistent parameter estimation, features selection, and asymptotically optimal misclassification rate. Extensive simulation study shows a significant improvement in classification performance under spatial dependence.

EO354 Room Aula B STATISTICAL METHODS FOR ANALYZING WEARABLE DEVICE DATA Chair: Russell Shinohara

E0338: **Potential batch effects and biases in the UK Biobank accelerometer data**
Presenter: **John Muschelli**, Johns Hopkins University, United States
Data from wearable technology, especially wrist-worn accelerometers is becoming more available. Over 100,000 people in a subset of the UK Biobank data set wore accelerometers for approximately 7 days in the study. We explore the potential biases in the data, including differences in the cohort with data and those without. We also discuss potential calibration and data normalization issues that are within the data that are device-dependent. We apply standard batch-effect correction methods, such as Tukeys biweight, to determine how these affect the overall results in a standard mortality analysis.

E0519: **Identifying circadian chronotypes using accelerometers**
Presenter: **Julia Wrobel**, Columbia University, United States
Co-authors: Jeff Goldsmith, Vadim Zipunnikov
Circadian rhythms are 24-hour biological processes that influence health on both the macroscopic and molecular level. Activity profiles produced by accelerometers can be used to understand circadian rhythm and detect chronotypes, which are subject-level differences in timing of circadian cycles. The focus is on understanding differences in the timing and intensity of activity, using a technique called registration. Registration aligns accelerometer curves by separating them into components of amplitude and phase variability. After alignment, the amplitudes show population level activity patterns that are consistent with well-documented diurnal patterns, and the phases contain information on subject-specific wake and sleep times. Previous work outlined a novel nonparametric method for registering exponential family functional data. We expand that method to understand circadian patterns in accelerometer curves, developing a 4-parameter approach that emphasizes interpretability of phase and amplitude components. After alignment the amplitudes can be described in terms of two parameters that identify the overall activity level, and whether each subject is likely to have a mid-day energy dip. The phases can be described in terms of parameters that identify the shift and duration of a subjects wake period. We validate our method using accelerometer data from the Baltimore Longitudinal Study of Aging. Code is publicly available as part of the registr package.

E0738: **Physical activity versus inactivity versus sleep: Isotemporal substitution effects**
Presenter: **John Staudenmayer**, UMass-Amherst, United States
It has long been recognized that how people allocate their time between physical activity, sedentary behavior, and sleep influences health. Large representative epidemiological surveillance studies such as the National Health and Nutrition Examination Survey (NHANES) recently have added

data to that discussion by both assessing health outcomes and using accelerometers to estimate how people spend their time at different levels of physical activity, inactivity, and sleep. Those measurements and estimates present a modeling challenge since the physical activity, inactivity, and sleep covariates add up to a constant (24 hours) for each participant. An ordinary linear regression model approach to this type of data has received a lot of attention in the epidemiology literature. When $p-1$ of the covariates are used in a model a regression coefficient estimates the effect of increasing its covariate by one unit on the health outcome while decreasing the left out covariate by one unit; i.e. an isotemporal substitution effect. A novel non-parametric approach to this problem will be developed.

E0956: **Multilevel variance components model in minute-level accelerometry measures for twin studies**
Presenter: **Haochang Shou**, University of Pennsylvania, United States
The emergence of mobile technologies, such as physical activity assessed via wearable actigraphy devices has provided an unprecedented opportunity to obtain objective evaluations of multiple physiological systems in real-time over weeks or months. However, the complexity of the devices and the high-dimensionality of the data also pose many analytic challenges to time-dependent measures. Most of the current approaches are based on summary statistics of activity that neglect the important time effects. We developed multilevel functional data analysis approaches that integrate multiple domains of complex measurements and reduce the dimensionality of the data while accounting for correlations in the repeated observations. In particular, motivated by the physical activity data observed from Brisbane adolescent twin study, we extended the traditional ACE model for a single univariate trait to functional outcomes based on an earlier work of structural functional principal component analysis (SFPCA). The method simultaneously: 1) handle various levels of correlation in the data; 2) identify interpretable traits via dimensionality reduction based on principal components; and 3) estimate relative variances that are attributed by additive genetic, shared environmental and unique environmental effects. Within-family similarities of those complex measures could also be effectively quantified.

EO210 Room Aula Magna RECENT DEVELOPMENTS IN COMPLEX COHORT STUDIES Chair: Yi Li

E0321: **Rank-consistency of the generalized Bradley-Terry model with link misspecification**
Presenter: **Yongdai Kim**, Seoul National University, Korea, South
The estimation of ranks has received much attention in the areas of information retrieval and online game design, and the Bradley-Terry model is popularly used. We show that the popularity of Bradley-Terry model is due to not only its flexible modeling and easy computation but also its good asymptotic properties even when the model is misspecified. We provide the sufficient conditions under which the Bradley-Terry model yields a consistent estimate of ranks with partial preferences data when the true generative model of ranks belongs to the class of the Thurstone model. By numerical experiments, we illustrate that the proposed sufficient conditions are important and practically useful.

E0477: **Local signal detection on irregular domains via bivariate spline**
Presenter: **Lan Xue**, Oregon State University, United States
Local signal detection is useful in many scientific areas such as imaging processing and speech recognition, for extracting meaningful patterns from noisy signals. We study estimation and local signal detection for spatial data distributed over irregular domains. In particular, we use bivariate splines defined on triangulations to approximate unknown signals on a complex domain nonparametrically. Subsequently, we propose a penalized polynomial spline method that simultaneously detects the null regions with signals and estimates the patterns on non-null regions. A smoothing proximal gradient (SPG) algorithm is used to find the estimator efficiently. In theory, the proposed estimator is shown to be consistent in estimating the true underling patterns. Furthermore, it is also able to detect the null signal region with probability approaching one. The numeric performance of the proposed method is evaluated through simulation studies and real data analysis. This validation shows that the proposed method and algorithm efficiently detect local signals on complex domains.

E0966: **The Lq-norm learning for ultrahigh-dimensional survival data: An integrative framework**
Presenter: **Hyokyoung Grace Hong**, Michigan State University, United States
In the era of precision medicine, survival outcome data with high-throughput predictors are routinely collected from many biomedical studies. Models with an exceedingly large number of covariates are either infeasible to fit or likely to incur low predictability because of overfitting. Variable screening is key in identifying and removing irrelevant attributes. Recent years have seen a surge in screening methods, but most of them rely on some particular modeling assumptions. Motivated by a study on detecting gene signatures for multiple myeloma patients' survival, we propose a model-free L_q-norm learning procedure, which includes the well-known Cramer–von Mises and Kolmogorov criteria as two special cases. The work provides an integrative framework for detecting predictors with various levels of impact, such as short- or long-term impact, on censored outcome data. The framework naturally leads to a scheme which combines results from different q to reduce false negatives, an aspect often overlooked by the current literature. We show that our method possesses sure screening properties. The utility of the proposal is confirmed with simulation studies and an analysis of the multiple myeloma study.

E1276: **Supervised dimensionality reduction for exponential family data**
Presenter: **Yoonkyung Lee**, Ohio State University, United States
Supervised dimensionality reduction techniques, such as partial least squares and supervised principal components, are powerful tools for making predictions with a large number of variables. The implicit squared error terms in the objectives, however, make it less attractive to non-Gaussian data, either in the covariates or the responses. Drawing on a connection between partial least squares and the Gaussian distribution, we show how partial least squares can be extended to other members of the exponential family similar to the generalized linear model for both the covariates and the responses. Unlike previous attempts, our extension gives latent variables which are easily interpretable as linear functions of the data and is computationally efficient. In particular, it does not require additional optimization for the scores of new observations and therefore predictions can be made in real time.

EO072 Room A1 RECENT DEVELOPMENTS IN IMAGING GENETICS Chair: Michele Guindani

E0618: **Tensor-on-tensor regression**
Presenter: **Eric Lock**, University of Minnesota, United States
A framework is proposed for the linear prediction of a multi-way array (i.e., a tensor) from another multi-way array of arbitrary dimension, using the contracted tensor product. This framework generalizes several existing approaches, including methods to predict a scalar outcome from a tensor, a matrix from a matrix, or a tensor from a scalar. We describe an approach that exploits the multiway structure of both the predictors and the outcomes by restricting the coefficients to have reduced CP-rank. We propose a general and efficient algorithm for penalized least-squares estimation, which allows for a ridge (L_2) penalty on the coefficients. The objective is shown to give the mode of a Bayesian posterior, which motivates a Gibbs sampling algorithm for inference. We illustrate the approach with an application to predict magnetic resonance spectroscopy metabolite profiles from genomic data.

E0625: **A consistent estimator of variance explained by genome-wide and whole-brain analyses**
Presenter: **Wesley Thompson**, Institute of Biological Psychiatry, Denmark

A typical analysis of genome-wide association studies (GWAS) and whole-brain voxel or vertex analyses is characterized by a large number of univariate regressions, wherein an outcome is regressed on thousands to millions of markers or voxels, one at a time. Assuming a linear model linking the markers to the outcome, an estimator is proposed for the variance explained for the trait, defined as the fraction of the variance of the trait explained by the markers or voxels in the study. The estimator is easy to compute, highly interpretable, and is consistent as the number of markers and the sample size increase. Importantly, it can be computed from summary statistics typically reported in GWAS and thus not requiring access to the original data. The estimator takes full account of the correlation between genomic markers or voxels. We also provide an analytical form for the standard errors of the GWASH estimator. We also sketch how this method may be extended to incorporate genetic and imaging data simultaneously.

E0626: **EEG spectral and heritability analysis using a nested Dirichlet process**
Presenter: **Mark Fiecas**, University of Minnesota, United States

A novel approach is presented for conducting spectral analysis on resting-state EEG (RS-EEG) data collected from the Minnesota Twin Family Study (MTFS). Typically, spectral analysis methods treat time series from each subject separately, and independent spectral densities are fit to each time series. In certain scenarios, such as our EEG data collected on twins, it is reasonable to assume that time series may have similar underlying characteristics, and borrowing information across subjects can significantly improve estimation. However, there are currently very few methods that share information across subjects when estimating spectral densities. We develop a Bayesian nonparametric modeling approach for estimating EEG spectra. In our methodology, we use Bernstein polynomials and a Dirichlet process (DP) to estimate each subject-specific spectrum. In order to estimate the spectra for the entire sample, we nest this model using a nested DP process. Thus, the top level DP cluster subjects with similar spectral densities and the bottom-level dependent DP fits a functional curve to the subjects within that cluster. We illustrate our methodology by conducting spectral analysis on resting state EEG data collected from the MTFS. The MTFS collected resting-state EEG and behavioral information from 379 monozygotic and 199 dizygotic twin pairs.

E1022: **A review of statistical methods in imaging genetics**
Presenter: **Linglong Kong**, University of Alberta, Canada

Simultaneously extracting and integrating rich and diverse heterogeneous information in neuroimaging and/or genomics from these big datasets could transform our understanding of how genetic variants impact brain structure and function, cognitive function, and brain-related disease risk across the lifespan. Such understanding is critical for diagnosis, prevention, and treatment of numerous complex brain-related disorders (e.g., schizophrenia and Alzheimer). However, the development of analytical methods for the joint analysis of both high-dimensional imaging phenotypes and high-dimensional genetic data, called big data squared (BD^2), presents major computational and theoretical challenges for existing analytical methods. Besides the high-dimensional nature of BD^2, various neuroimaging measures often exhibit strong spatial smoothness and dependence and genetic markers may have a natural dependence structure arising from linkage disequilibrium. We review some recent developments of various statistical techniques for the joint analysis of BD^2, including massive univariate and voxel-wise approaches, reduced rank regression, mixture models, and group sparse multi-task regression. By doing so, we hope that this review may encourage others in the statistical community to enter into this new and exciting field of research.

EO152 Room Aula A STATISTICS AND STOCHASTIC ANALYSIS FOR COMPLEX RANDOM SYSTEMS Chair: Hiroki Masuda

E0328: **Global jump filters and quasi likelihood analysis for volatility**
Presenter: **Nakahiro Yoshida**, University of Tokyo, Japan

New estimation schemes for volatility parameters of a semimartingale with jumps are proposed. In order to detect jumps, construction of a suitable filter that correctly discriminates intervals having jumps among all observation intervals is critical. The jump-filters proposed so far are based on time-locally constructed tests of jumps, and they suffer restrictions on the distribution of jumps. The proposed jump-filters take advantage of global information in the data to detect jumps more accurately. The quasi likelihood analysis (QLA) for volatility parameter estimation is formulated by using the newly proposed jump filters. Stable convergence to a mixed normal distribution of the QLA estimators (the quasi maximum likelihood estimator and the quasi Bayesian estimator) and moment convergence of the error are established. Numerical simulations show that our global method obtains better estimates of the volatility parameter than the previous local method.

E0331: **Point process regression model in the yuima project**
Presenter: **Lorenzo Mercuri**, University of Milan, Italy

The point process regression models have been introduced previously for modeling the limit order book. These processes can be seen as a generalization of the multivariate Hawkes model due to the presence of covariates in the intensity process. Moreover, the flexibility of the package yuima allows the users to define a specific dynamics for the intensity process. We review classes and methods for the estimation and the simulation of these models. We also discuss some useful advances with respect to the existing theoretical literature that are used in the implementation schemes.

E0678: **Local asymptotic mixed normality for integrated diffusion processes**
Presenter: **Teppei Ogihara**, The Institute of Statistical Mathematics, Japan

Statistical inference for integrated diffusion processes is studied and asymptotic properties of this model in a high-frequency limit are considered. This model arises when we observe a process after passage through an electric filter, and is also related to the modeling of the stochastic volatility in finance. This model has been previously studied and the local asymptotic mixed normality (LAMN) has been shown when the latent diffusion process is one-dimensional. The LAMN property is important in asymptotic statistical theory and enables us to discuss asymptotic efficiency of estimators. We extend their results of the LAMN property to multi-dimensional diffusion processes which may have a feedback from the integral process. Then, we can apply these results to the Langevin equation which is a model for molecular motion. We also consider the construction of an efficient estimator.

E0749: **Testing the residual sparsity of a high-dimensional continuous-time factor model**
Presenter: **Yuta Koike**, University of Tokyo, Japan

Investigation of the correlation structure of the residual process of a factor model is an important problem to assess systematic risk factors unexplained in the model. Such a problem is also important in high-dimensional covariance estimation because approximate factor models are often employed to reduce the curse of dimensionality, especially in the context of financial applications. A multiple testing procedure is developed to detect correlated pairs in the residual processes of a continuous-time factor model for multiple assets observed at a high-frequency in a high-dimensional setting such that the number of assets is possibly larger than the sample size.

EO366 Room Aula C EXPLORING THE LIMITS OF STATISTICAL LEARNING TECHNIQUES Chair: Markus Pauly

E0227: **Modular regression: A lego system for building distributional regression models with tensor product interactions**
Presenter: **Thomas Kneib**, University of Goettingen, Germany
Co-authors: Nadja Klein, Stefan Lang, Nikolaus Umlauf

Semiparametric regression models offer considerable flexibility concerning the specification of additive regression predictors including effects as diverse as nonlinear effects of continuous covariates, spatial effects, random effects, or varying coefficients. Recently, such flexible model predictors have been combined with the possibility to go beyond pure mean-based analyses by specifying regression predictors on potentially all parameters of the response distribution in a distributional regression framework. We introduce a generic concept for defining interaction effects in such semiparametric distributional regression models based on tensor products of main effects. These interactions can be anisotropic, i.e. different amounts of smoothness will be associated with the interacting covariates. We investigate identifiability and the decomposition of interactions into main effects and pure interaction effects (similar as in a smoothing spline analysis of variance) to facilitate a modular model building process. Inference is based on Markov chain Monte Carlo simulations with iteratively weighted least squares proposals under constraints to ensure identifiability and effect decomposition.

E0470: **A cautionary tale on using imputation methods for inference in matched pairs design**
Presenter: **Burim Ramosaj**, Ulm University, Germany
Co-authors: Lubna Amro, Markus Pauly

Imputation procedures in biomedical fields have turned into statistical practice, since further analyses can be conducted ignoring the former presence of missing values. In particular, non-parametric imputation schemes like the random forest or a combination with the stochastic gradient boosting have shown favorable imputation performance compared to the more traditionally used MICE procedure. However, their effect on valid statistical inference has not been analyzed so far. This gap is closed by investigating their validity for inferring mean differences in incompletely observed pairs while opposing them to a recent approach that only works with the given observations at hand. Our findings indicate that machine learning schemes for (multiply) imputing missing values heavily inflate type-I-error in small to moderate matched pairs, even after modifying the test statistics using Rubin's multiple imputation rule. In addition to an extensive simulation study, an illustrative data example from a breast cancer gene study has been considered.

E0834: **Tuning and tunability: Importance of hyperparameters of machine learning algorithms**
Presenter: **Philipp Probst**, LMU Munich, Germany
Co-authors: Anne-Laure Boulesteix, Bernd Bischl

Modern machine learning algorithms for classification or regression such as gradient boosting, random forest and neural networks involve a number of parameters that have to be fixed before running them. Such parameters are commonly denoted as hyperparameters. Users of these algorithms can use defaults of the hyperparameters that are specified in the employed software package, set them to alternative specific values or use a tuning strategy to optimize them with respect to performance for the specific dataset at hand. We formalize the problem of tuning from a statistical point of view and suggest general measures quantifying the tunability of hyperparameters and of algorithms. They are calculated for six of the most common statistical learning algorithms. Our results may help users and software developers to set defaults appropriately, to decide whether it is worth to conduct a possibly time consuming tuning strategy, to focus on the most important hyperparameters and to choose adequate hyperparameter spaces or even prior distributions for tuning strategies like sequential model-based optimization. This is one step in the automation of the model building process which consists of several steps such as feature creation and selection, tuning, stacking, etc. and which is partly already available in implementations like auto-sklearn, AutoWeka and H_2O AutoML. Ideally the time of this process can be estimated and restricted before execution.

EO685 Room B1 GRAPHICAL MARKOV MODELS IV Chair: Elena Stanghellini

E1399: **Markov properties of determinantal point processes**
Presenter: **Kayvan Sadeghi**, University of Cambridge, United Kingdom

Determinantal point processes (DPPs) have been widely used in machine learning for statistical modelling. We discuss the conditional independence structure of DPPs. In particular, we show that the induced independence model by DPPs can be naturally captured by bidirected graphs. In addition, we show that the context-specific induced independence models by DPPs (conditioning on variables being all equal to 1) act in the same way as the independence model induced by Gaussian distribution. This leads to context-specific DPP undirected as well as directed acyclic graphical models.

E1418: **Causal transmission in reduced-form models**
Presenter: **Vassilios Bazinas**, International Monetary Fund, United States

A method is proposed to explore the causal transmission of a catalyst variable through two endogenous variables of interest. The method is based on the reduced-form system formed from the conditional distribution of the two endogenous variables given the catalyst. The method combines elements from instrumental variable analysis and Cholesky decomposition of structural vector autoregressions. We give conditions for uniqueness of the causal transmission.

E1702: **On integer linear programming approach to learning decomposable graphical models**
Presenter: **Milan Studeny**, Institute of Information Theory and Automation of the CAS, Czech Republic

The decomposable graphical models, described by chordal undirected graphs, are crucial in famous local computation method, used widely in probabilistic graphical models. The basic ideas of the integer linear programming approach to learning these graphical models will be recalled. We propose to represent them by special zero-one vectors, which idea leads to the study of a special polytope, called chordal graph polytope. The focus will be on a conjecture; what are all facet-defining inequalities for this polytope.

E1759: **Graphical models for circular data**
Presenter: **Anna Gottard**, University of Firenze, Italy
Co-authors: Agnese Panzera

Graphical models have been successfully used to characterise conditional independence structures among random variables. For instance, in proteomics, we would now like to analyse the structure of proteins in terms of their characterising angles. For this, the first task is to study conditional independence in multivariate distributions on angles. One example is the multivariate von Mises distribution, also known as the multivariate sine distribution. We review the existing literature on graphical models for circular data and propose possible extensions both for model specifications and inference.

EO300 Room C1 NEW DEVELOPMENT IN FUNCTIONAL DATA AND DENSITY ESTIMATION Chair: Yuedong Wang

E0188: **Covariance estimation and principal component analysis for spatially dependent functional data**
Presenter: **Yehua Li**, University of California at Riverside, United States
Spatially dependent functional data are considered which are collected under a geostatistics setting, where locations are sampled from a spatial point process and a random function is observed at each location. Observations on each function are made on discrete time points and contaminated with measurement errors. The error process at each location is modeled as a non-stationary temporal process rather than white noise. Under the assumption of spatial isotropy, we propose a tensor product spline estimator for the spatio-temporal covariance function. If a coregionalization covariance structure is further assumed, we propose a new functional principal component analysis method that borrows information from neighboring functions. Under a unified framework for both sparse and dense functional data, where the number of observations per curve is allowed to be of any rate relative to the number of functions, we develop the asymptotic convergence rates for the proposed estimators. The proposed methods are illustrated by simulation studies and a motivating example of the home price-rent ratio data in the San Francisco metropolitan area.

E0182: **Semi-parametric density models**
Presenter: **Yuedong Wang**, University of California - Santa Barbara, United States
Maximum likelihood estimation within a parametric family and nonparametric estimation are two traditional approaches for density estimation. Sometimes it is advantageous to model some components of the density function parametrically while leaving other components unspecified. We propose estimation methods for a general semiparametric density model and develop computational procedures under different situations. We also present simulation results and real data examples.

E0731: **Flexible regression for probability densities in Bayes spaces**
Presenter: **Almond Stoecker**, LMU Munich, Germany
Co-authors: Eva-Maria Maier, Sonja Greven
A flexible regression framework is presented for functional compositional data, i.e. functional additive regression models (FAMs) for the case that functional response variables or covariates are probability density functions (PDFs). The special nature of PDFs - in particular their property to integrate to one - prohibits direct application of usual functional regression models. Instead, we formulate FAMs for PDFs in a Bayes Hilbert space. The isometry given by the so called centered log-ratio transform allows us to carry over the flexibility of previous FAMs to Bayes space models. Thus, we are able to provide a wide range of different types of categorical, scalar or functional covariate effects. For instance, in an application to the annual birth distributions in Germany from 1950 to 2016, we include a smooth scalar effect for the year and a categorical effect for sex to model the birth PDFs, while also accounting for the cyclic nature of the response PDFs. For estimation, we consider two different procedures: a penalized least squares approach and component-wise gradient boosting, which also yields inherent model selection.

E1144: **Fréchet regression and Wasserstein covariance for random density data**
Presenter: **Alexander Petersen**, University of California Santa Barbara, United States
Co-authors: Hans-Georg Mueller
Samples of density functions appear in a variety of disciplines, including connectivity distributions of voxel-to-voxel correlations of fMRI signals or distributions of voxel-specific attenuation coefficients from CT scans across subjects. The nonlinear nature of density space necessitates adaptations and new methodologies for the analysis of random densities. We define our geometry using the Wasserstein metric, an increasingly popular choice in theory and application, and investigate two modeling problems. First, when densities appear as responses in a regression relationship with vector covariates, we consider the Fréchet regression model, which provides a general purpose methodology for response objects in a generic metric space. Importantly, we enlarge the scope of this regression framework for density data by placing distributional assumptions on the residual processes (in this case, random optimal transport maps) that allow for further inference beyond estimation, specifically submodel testing. Second, when multiple random densities are observed for each subject, we propose the Wasserstein covariance matrix, yielding a scalar summary measure of covariance for each pair of random densities. Using the fMRI connectivity distributions as an example, we find that the Wasserstein covariance matrix provides an interpretable summary of dependence across regions that also reveals key distinguishing features between normal and Alzheimer's subjects.

EO408 Room D1 SOME NEW TRENDS IN HYPER-PARAMETER CALIBRATION Chair: Adrien Saumard

E0398: **Comparison methods for bandwidth selection**
Presenter: **Claire Lacour**, Paris-Est University / INRIA, France
Co-authors: Pascal Massart, Vincent Rivoirard, Suzanne Varet
The problem of estimating a density with kernel estimators is considered. A classical issue is the choice of the bandwidth. We focus on the Goldenshluger-Lepski selection method, which is based on pairwise comparisons between estimators with respect to some loss function. The method also involves a penalty term than typically needs to be large enough in order that the method works (in the sense that one can prove some oracle type inequality for the selected estimator). In the case of the quadratic loss, we study the procedure for different values of the tuning parameter. We give a minimal value of the penalty, beyond which the procedure fails, that brings to light a phase transition phenomenon for penalty calibration. Moreover, we highlight a degenerate case, where all the estimators are compared to the overfitted one. This leads to a new method which is in some sense intermediate between Lepski's method and penalized empirical risk minimization. We provide some theoretical results which lead to some fully data-driven selection strategy. We will also show the numerical performance of the method.

E0720: **V-fold cross-validation improved for nonparametric regression**
Presenter: **Amandine Dubois**, CREST-ENSAI, France
Co-authors: Adrien Saumard
The framework where the properties of model selection procedures are best theoretically understood is that of estimating a function, such as a regression function or a density, by minimizing the risk on finite-dimensional models, corresponding to developments on functional bases. In this case, the hyper-parameter which needs to be tuned is the dimension of the models that are considered. The methods commonly used are based on the unbiased risk estimation principle. The basis is the idea that the validity of this principle is essentially asymptotic. In the least squares regression framework, a modification of the V-fold penalty will be proposed, that surpasses the limits of the unbiased risk estimation principle. In a nutshell, it is more efficient to estimate a quantile of the risk of the estimators rather than its mean. An experimental study will highlight the performances of this procedure in comparison with classical V-fold cross-validation.

E0973: **Cross-validation improved by aggregation: Agghoo**
Presenter: **Guillaume Maillard**, Universita Paris Sud, France
Co-authors: Matthieu Lerasle, Sylvain Arlot
Cross-validation is widely used for selecting among a family of learning rules. A related method, called aggregated hold out (Agghoo), is studied which mixes cross-validation with aggregation; Agghoo can also be related to bagging. We provide the first theoretical guarantees on Agghoo, ensuring that one can use it safely: at worse, Agghoo performs like the hold out, up to a constant factor. For the hold out, oracle inequalities

were known in the case of bounded losses, as in binary classification. The approach can be extended to most classical risk minimization problems, including regression with least squares loss or others: it works particularly well with Lipschitz losses such as the Huber loss or quantile regression. In all these settings, Agghoo verifies an oracle inequality. However, simulation studies suggest that real performance is often much better than what theory can currently prove. In particular, there is a large gain from aggregation that current bounds derived from the hold out are incapable of capturing. As a result, Agghoo appears to be competitive with standard cross-validation in practice.

E1547: **Early stopping rules reproducing kernel Hilbert spaces**
Presenter: **Alain Celisse**, Lille University, France
The main focus is on the nonparametric estimation of a regression function by means of reproducing kernels and iterative learning algorithms (gradient descent or Tikhonov regularization). First, we exploit the general framework of filter estimators to provide a unified analysis of these different algorithms. Second, we introduce an early stopping rule derived from the so-called discrepancy principle. Its behavior is compared with that one of other existing stopping rules and analyzed. An oracle type inequality is derived to quantify the finite sample performance of the proposed stopping rule. The practical performance of the procedure is also empirically assessed from simulation experiments.

EO252 Room E1 RECENT ADVANCES IN SKEWNESS Chair: Nicola Loperfido

E0273: **Symmetric tensor rank and projection pursuit**
Presenter: **Nicola Loperfido**, University of Urbino, Italy
A tensor is a multi-way array representing a multilinear operator, up to a choice of bases. It is symmetric if it is invariant under permuting indices. A symmetric tensor is decomposable if its elements may be represented as products of a fixed number of elements belonging to the same vector. The symmetric rank of a symmetric tensor (also known as the symmetric tensor rank) is the minimum number of symmetric, decomposable tensors which need to be added together to get the tensor itself. Symmetric tensor rank plays a relevant role in projection pursuit, as for example in skewness-based projection pursuit, kurtosis-based projection pursuit, projection pursuit regression and projection pursuit density estimation.

E0387: **Analytic solution of a portfolio optimization problem in a mean-variance-skewness model**
Presenter: **Zinoviy Landsman**, University of Haifa, Israel
Co-authors: Udi Makov, Tomer Shushi
In portfolio theory, it is well-known that in most of the cases stocks follows a non-symmetric and unimodal distributions. Therefore, many researches have suggested considering the skew-normal distribution an accurate model in quantitative finance. From the fact that the portfolio of stocks is non-symmetric, the celebrated mean-variance theory fails to provide an optimal portfolio selection rule, which comes from the fact that the mean-variance model does not take into account the skewness of the stocks. We provide a novel approach that solves such a problem of optimal portfolio selection with non-symmetric stocks, by putting it into a framework of mean-variance-skewness measure. For example, we show an analytical portfolio optimization solution to the exponential utility of the well-known skew-normal distribution or, even more general, scale mixtures of skew-normal distribution. Moreover, our optimal solutions are explicit and has closed-forms, and therefore they can be investigated in comparison to other portfolio selection rules, such as the standard mean-variance model. The results are then illustrated numerically.

E0377: **Projection pursuit under skew-normal vectors: An approach oriented to skewness stochastic comparisons**
Presenter: **Jorge Martin Arevalillo**, UNED, Spain
Co-authors: Hilario Navarro Veguillas
The skewness based projection pursuit problem for vectors that follow a multivariate skew-normal (SN) distribution is revisited. The issue, which is a standard in multivariate data analysis, has a close connection with the canonical transformation of SN vectors. We elaborate on the implications of such a connection in order to define a skewness ordering that allows stochastic comparisons within the family of multivariate SN distributions. The proposed ordering relies on the convex transform ordering between the only skewed component derived from the canonical representation of SN vectors under comparison. Its relationship with standard multivariate skewness measures is also examined and some highlights showing its usefulness in the statistical practice are provided as well.

E1284: **Objective Bayesian analysis of the multivariate regression model with skew-t errors**
Presenter: **Antonio Parisi**, University of Rome Tor Vergata, Italy
Co-authors: Brunero Liseo
In the last decades, several specifications of skew-Student t distributions have been proposed as empirical models for data characterized by skewness and/or extra-kurtosis. The flexibility provided by these models generally comes at the cost of several inferential difficulties, mostly in the multivariate setup. Under a Bayesian perspective, we consider a (possibly multivariate) regression model in which the error vector term has a multivariate skew-elliptical distribution. More specifically, a skew-t distribution or a special case of it: the Student-t, the skew-normal or the Gaussian distribution. For this regression model, we propose a set of objective priors and a specifically designed Monte Carlo sampler. The new version of the R package mvst contains functions that implement the described methods, in particular to estimate the model and to obtain pseudo-random draws. The package also allows the choice among the skew-t model and the nested ones via the Bayes factor. Moreover, it allows to generalize the reference model in several ways; as an example, a stochastic frontier model can be estimated by a change in the elicitation of the prior distribution for the shape parameter.

EO138 Room F1 STABILITY VERSUS NON-STABILITY Chair: Marie Huskova

E0353: **Quantile lasso with changepoints in panel data models**
Presenter: **Matus Maciak**, Charles University, Czech Republic
Panel data models are quite modern statistical tools and they are commonly used in all kinds of econometric problems. In our approach we consider panel data models with changepoints, and atomic pursuit methods are utilized to detect and estimate these changepoints in the model. In order to obtain robust estimates and, also, to have a more complex insight into the data, we adopt the quantile lasso approach and the final model is obtained in a fully data-driven manner in just one modelling step. The main theoretical results are presented and some inferential tools for changepoint significance are proposed. The presented methodology is applied for a real data scenario and some finite sample properties are investigated via s simulation study.

E0369: **On the performance of weighted bootstrapped kernel deconvolution density estimators**
Presenter: **William Pouliot**, University of Birmingham, United Kingdom
A weighted bootstrap approach is proposed which can improve on current methods to approximate the finite sample distribution of normalized maximal deviations of the kernel density estimators in the case of *ordinary smooth* errors. Using results from the approximation theory for weighted bootstrap empirical processes, we establish an unconditional weak limit theorem for the corresponding weighted bootstrap statistics. Because the proposed method uses weights that are not necessarily confined to be uniform (as in Efron's original bootstrap), it provides the practitioner with additional flexibility for choosing the weights. As in immediate consequence of our results, one can construct uniform confidence bands, or perform

goodness-of-fit tests, for the underlying density. We have also carried out some numerical examples which show that, depending on the bootstrap weights chosen, the proposed methods has the potential to perform better than the current procedures in the literature.

E0792: **A Bayesian circular changepoint method to identify changes in daily activity levels in the elderly**
Presenter: **Simon Taylor**, Lancaster University, United Kingdom
Co-authors: Rebecca Killick
According to Age UK there are 11.6m over 65's, 3.64m who live alone and over 25% need help with at least one daily activity. A growing body of research indicates that changes in daily routine signal a change in health and well-being. Motivated by an industrial collaboration with Howz, we are using motion sensors in the home to automatically detect these changes as a key step in improving outcomes for elderly people and vulnerable members of society who live alone. Traditional changepoint methods to identify changes in activity levels view time as linear and thus are able to identify the day/night routine on a day-to-day basis. The typical assumption of independence of segments results in estimated changepoints and parameters that are unlikely to be consistent from day-to-day. As changes in routine happen on longer time scales, traditional methods make determining the normal daily patterns more challenging. We demonstrate a new changepoint method in the Bayesian framework that views time as circular in order to estimate the time-of-day changepoint events between different activity levels by pooling together information from across multiple days. These daily patterns can then be monitored for significant changes in daily changepoint locations and/or parameter estimates within segments.

E1048: **Structural breaks in panel data models with stationary regressors**
Presenter: **Marie Huskova**, Charles University, Czech Republic
Co-authors: Adam laf
The aim is to test and detect structural breaks in the panel data setup with stationary regressors when both T (time dimension) N (number of panels) tend to infinity. Test procedures for no break versus there is a break and estimators of the time of structural break are developed. The asymptotic behavior of the test statistics and break point estimators will be presented. Theoretical results are accompanied by simulations and the application in the framework of the four factors CAPM model for monthly returns of the US mutual fund during the period covering the subprime crises.

EO498 Room G1 NEW MODELLING APPROACHES FOR COMPLEX SURVIVAL DATA Chair: Giuliana Cortese

E0898: **Combining low-dimensional clinical and high-dimensional molecular data in a survival prediction model**
Presenter: **Riccardo De Bin**, University of Oslo, Norway
Combining low-dimensional clinical and high-dimensional molecular sources of information in a survival prediction model is not straightforward. Several issues arise due to their difference in nature: the former is characterized by few predictors whose prediction value is usually well-validated in the biomedical literature; the latter by a large number of predictors and a low signal to noise ratio. Different strategies have been proposed to efficiently combine the two sources of data, mainly aiming at fully exploiting the clinical information notwithstanding the noise linked to the molecular part. It is shown how these strategies work in practice, with a special focus on their performances when used within a statistical boosting procedure. Merging the powerfulness of a machine learning algorithm and the interpretability of a statistical model, boosting is one of the most interesting approaches to use when dealing with both low and high-dimensional sources of data. The results are illustrated through two real data examples.

E1027: **Spatial survival models for analysis of exocytotic events on human beta-cells recorded by TIRF imaging**
Presenter: **Thi Huong Phan**, Department of Statistical Sciences, University of Padua, Italy
In cell biology, exocytosis is a fundamental event observed in human beta-cells from high-resolution microscopy images. Studying the rate and spatial locations of exocytosis events, and predicting its survival probability, are of great interest in biomedical research as it helps to discover the cellular processes related to insulin-secretion dysfunctioning in diabetic patients. The main objective is to investigate the relationship between the exocytosis rate and syntaxin level observed during the experiments, while studying the possible spatial correlations within each cell. A Gaussian frailty survival model is proposed where individual spatial correlation is investigated through several different parametric families while independence clustering structure is preserved in the block pattern of frailty variance-covariance matrix. For estimation of model parameters, two common likelihood-based inferential approaches are firstly investigated: Monte-Carlo Expectation-Maximization (MCEM) approach and a penalized partial likelihood (PPL) approach. Finally, we propose a new approach where the marginal likelihood is estimated by pairwise likelihoods and quadrature approximation (QPLH). Their drawbacks and advantages are discussed in simulation studies, and also major results of the data application are presented, showing that exocytosis rates are spatially correlated and depend on their distance within each cell.

E0656: **Diagnostics and predictions for joint models of survival and multivariate longitudinal data**
Presenter: **Marcella Mazzoleni**, University of Milano Bicocca, Italy
Co-authors: Mariangela Zenga
The joint models analyse the effect of longitudinal covariates onto the risk of an event. The longitudinal and the survival sub-models compose the joint models. The survival sub-model is a proportional hazard model, while the longitudinal sub-model is a linear multivariate mixed model. An Expectation-Maximisation (EM) algorithm which maximises the joint likelihood function is implemented. For testing the goodness of fit of the algorithm some diagnostics elements will be presented, such as residuals for both sub-model and the estimated survival function. Moreover, the dynamic predictions for the survival part of the model based on the longitudinal covariates will be shown.

E0685: **Monte Carlo modified profile likelihood in survival models for clustered censored data**
Presenter: **Claudia Di Caterina**, University of Padova, Italy
The main focus of the analysts who deal with clustered survival data is usually not on the clustering variables, and hence the group-specific parameters are treated as nuisance. If a fixed effects formulation is preferred and the total number of clusters is large relative to the single-group sizes, classical parametric frequentist techniques relying on the profile likelihood are often misleading. The use of alternative tools, such as modifications to the profile likelihood or integrated likelihoods, for making accurate inference on a parameter of interest can be complicated by the presence of nonstandard modelling assumptions. We propose to employ Monte Carlo simulation in order to approximate the modified profile likelihood in general regression settings for survival data with unspecified censoring mechanism. Particularly, a nonparametric bootstrap is encompassed in the procedure to estimate the latter.

EO302 Room I1 THE STEIN METHOD AND APPLICATIONS IN STATISTICS Chair: Andreas Anastasiou

E0314: **Statistical inference with Stein discrepancies**
Presenter: **Francois-Xavier Briol**, University of Warwick, United Kingdom
A major class of problems which cannot be directly tackled by maximum likelihood estimation include models whose likelihood include an unknown normalisation constant. These include many graphical models, where computing the normalisation constant would require summing or integrating all possible combinations of the graph, and latent-variable models, where the normalisation constant is an integral over all latent variables. This has led to the development of a wide range of methods which approximate this unknown normalisation constant. We propose to make use of Stein's method to develop estimators which bypass this issue and only require access to the score functions. We make use of information geometry to provide results on the consistency and asymptotic distribution of these estimators, then study their robustness to corrupted data.

E1151: **Chi-square approximation by Stein's method with application to Pearson's statistic**
Presenter: **Robert Gaunt**, The University of Manchester, United Kingdom
The Stein method for chi-square approximations is reviewed and some recent developments are discussed. We apply this theory to bind the distributional distance between Pearsons statistic and its limiting chi-square distribution, measured using smooth test functions. In combination with the use of symmetry arguments, Steins method yields explicit bounds on this distributional distance of order $1/n$. This bound also has the correct dependence on the cell classification probabilities, and we obtain a Kolmogorov distance bound which shares this property.

E1166: **Stein's method applied to some statistical problems**
Presenter: **Jay Bartroff**, University of Southern California, United States
Two instances of applying Stein's method to statistical problems are discussed. First, in the setting of group sequential testing methods where accumulating data is evaluated intermittently in stages, an existing multivariate Berry-Esseen bound based on Stein's method is applied to the joint distribution of MLEs of a vector parameter at each group analysis to obtain explicit bounds to its limiting normal distribution. The setting is a general parametric regression setup which allows the i-th observation to be the i-th subject's (say) response regressed on their covariates. Second, new and improved concentration inequalities are obtained via Stein's method for a class of multivariate occupancy models whose marginal distributions are lattice log concave and satisfy some other weak conditions. Examples with a statistical flavor in this class include degree counts in an Erdos-Renyi random graph, the number of neighbors and the volume covered by multi-way intersections in germ-grain models, bin occupancy counts in the multinomial model, and population sizes under multivariate hypergeometric sampling. In these models the new method provides concentration inequalities having the Poisson tail rate, many of which improve on those achieved by competing methods.

E1197: **Stein kernels and information**
Presenter: **Yvik Swan**, Universite de Liege, Belgium
Co-authors: Gesine Reinert, Marie Ernst
A general concept of Stein kernels is presented and the corresponding covariance identities are developed. We propose a notion of Stein-Fisher information. Among many possible applications, we extract a general tool for goodness-of-fit testing based on recent works concerning "kernelized stein discrepancy for goodness-of-fit tests". We insist mainly on the discrete framework for the examples.

EO266 Room L1 CLUSTERING AND SKETCHING IN STATISTICS AND COMPUTATION Chair: Stephane Chretien

E1283: **Co-clustering: A versatile way to perform clustering in high dimension**
Presenter: **Christine Keribin**, INRIA - Universite Paris-Sud, France
Co-authors: Christophe Biernacki
Standard model-based clustering is known to be very efficient for low dimensional data sets, but it fails for properly addressing high dimension (HD) ones, where it suffers from both statistical and computational drawbacks. In order to counterbalance this curse of dimensionality, some proposals have been made to take into account redundancy and features utility, but related models are not suitable for too many variables. We advocate that co-clustering, an unsupervised mixture model learning method to define simultaneously groups of rows (individuals) and groups of columns (variables) on a data matrix, is of particular interest to perform HD clustering of individuals even if it is not its primary mission. Indeed, column clustering is recasted as a strategy to control the variance of the estimation, the model dimension being driven by the number of groups of variables instead of the number of variables itself. However, the statistical counterpart of this important variance reduction brings naturally some important model bias. The purpose is to access (first in an empirical manner) the trade-off bias-variance of the co-clustering strategy in scenarii involving HD fundaments (correlated variables, irrelevant variables). We show the ability of co-clustering to outperform simple mixture row-clustering, even if co-clustering clearly corresponds to a misspecified model situation, revealing a promising manner to efficiently address (very) HD clustering.

E1535: **Clustering using low rank matrix estimation**
Presenter: **Stephane Chretien**, NPL, United Kingdom
The problem of unsupervised clustering of high dimensional data, e.g.images, time series, gene expression data, etc. is considered. Such problems have attracted much interest in mathematical learning research, because of its wide applicability, from image segmentation, automatic medical diagnosis, marketing, data quality assessment, outlier detection, etc. We show how clustering can be addressed using a very simple low rank nonnegative matrix estimation problem, which can be solved efficiently using Burer-Monteiro type factorisation techniques. We then apply this clustering technique cluster-based reduced-order modelling (CROM), a recent technique generalising the Ulam-Galerkin method classically used for non-linear dynamical system analysis in order to determine a finite-rank approximation of the Perron-Frobenius operator.

E1331: **Sketching algorithms for the Galerkin finite element method**
Presenter: **Nick Polydorides**, University of Edinburgh, United Kingdom
Some methodologies are discussed for sampling the Galerkin system of equations arising during the numerical solution of elliptic partial differential equations on high-dimensional models. We show that the assembly of the coefficients matrix and right-hand side vector in the resulting linear Galerkin system can be formulated as a high-dimensional sum of sparse, low-rank matrices which we attempt to approximate by sampling. It turns out that the optimal sampling distribution involves the parameters of the PDE and some geometric properties of the numerical model. To reduce the computational complexity in solving the sketched system we explore two approaches: projecting the Galerkin equations onto a low-dimensional subspace as well as casting the problem as a least squares problem. For both cases we provide error bounds based on the optimal sampling distributions and sampling budgets, and illustrate the performance of the algorithms using numerical experiments.

E1521: **A data-driven approach to transfer operators in nonlinear dynamics using neural networks**
Presenter: **Luigi Marangio**, Universite Bourgogne Franche-Comte / Universita di Pisa, Italy
Co-authors: Christophe Guyeux

Non linear dynamical systems arise almost everywhere in science: natural phenomenas, from biology to economics, can be described as a non linear dynamics in some suitable space. In several contexts finding a good dynamical model for the phenomena which is studied is a hard task. Recently, a new trend is taking hold: rather than observing a phenomena trying to model it with (partial differential) equations, this new approach aims to compute (usually) a matrix that should describe, up to some approximation error, the dynamic underling a big set of data (data-driven approach). Functional analysis provide a powerful tool to understand the statistical properties of dynamics, the so-called transfer operator: an infinite dimensional operator associated to a dynamical system, describing how the dynamics governs the evolution of initial probability densities instead of initial points. Under suitable assumptions, it is reasonable to approximate it with a matrix, which can be computed from the data. The goal is to shortly resume the data-driven techniques developed until now, clarify the mathematical theory involved in these approaches, to present a neural network based algorithm for the computation of a finite approximation of the transfer operator, and finally to apply this algorithm to a data set arising from fireman activity.

EO256 Room M1 STATISTICS FOR HILBERT SPACES II Chair: Gil Gonzalez-Rodriguez

E0876: **Functional regression modeling for agricultural data**
Presenter: **Hidetoshi Matsui**, Shiga University, Japan
Co-authors: Keiichi Mochida
In crop cultivation, it is considered that there are strong relations between information on crop yields and the habitat environment such as temperature and sunlight. In particular, many data sets for the environments are measured over time, and it is desirable to properly handle this information. We report a method for constructing a statistical model that represents the relationship between the data for habitat environment and the crop yield. Time course observations for environments are treated as functional data, and then a functional regression model is considered. We investigate the effect of the environments on the crop yield from the estimated model.

E0463: **Estimation of points of impact in nonparametric regression with functional predictors**
Presenter: **Dominik Liebl**, University Bonn, Germany
Co-authors: Dominik Poss, Alois Kneip
In many applications only specific locations or time-points of functional predictors have an impact on the outcome. The selection of such points of impact constitutes a particular variable selection problem, since the high correlation in the functional predictors violates the basic assumptions of existing high-dimensional variable selection procedures. We introduce a method to estimate points of impact in nonparametric regression models. We propose a threshold-based and a fully data-driven estimator and show that the point of impact estimators can be estimated with a super-consistent convergence rate for a large class of functional data processes. The finite sample properties of our estimators are assessed by means of a simulation study. Our methodology is motivated by a psychological case study in which the participants were asked to continuously rate their emotional state while watching an affective online video on the persecution of African albinos.

E0762: **Multivariate functional additive mixed models**
Presenter: **Sonja Greven**, LMU Munich, Germany
Co-authors: Alexander Volkmann, Almond Stoecker, Fabian Scheipl
Functional data are often multivariate, i.e. they simultaneously measure different functional aspects of a process. So far, few regression methods have been developed to efficiently handle the full amount of information provided by multivariate functional data. We develop a multivariate functional additive mixed model (MFAMM). The dependency structure between the different dimensions is incorporated using multivariate functional principal component analysis. The model accounts for correlation within the functions, between the multivariate functional dimensions as well as potentially further between-function correlation - which is often induced by the study design - via multivariate functional random intercepts. Multivariate functional data generated in a speech production study with a crossed study design are analyzed. The analysis is more parsimonious compared to fitting independent univariate models to the data and generates insight into the dependency structure between acoustic and articulatory processes. Application results also suggest that estimated confidence regions might be more efficient for the MFAMM than for the univariate approach.

E1204: **Optimal function-on-scalar regression over complex domains**
Presenter: **Matthew Reimherr**, Pennsylvania State University, United States
Co-authors: Bharath Sriperumbudur, Hyun Bin Kang
The problem of estimating function-on-scalar regression models is considered when the functions are observed over multi-dimensional domains and with potentially multivariate output. We establish the minimax rates of convergence and present an estimator based on reproducing kernel Hilbert spaces that achieves this optimal rate. We conclude with a numerical study and an application to 3D facial imaging.

EO050 Room N1 STATISTICAL ANALYSIS OF EXTREMES IN FINANCE AND INSURANCE Chair: Gilles Stupfler

E0236: **Extremiles: A new perspective on asymmetric least squares**
Presenter: **Abdelaati Daouia**, UMR5314 TSE-R CNRS, France
Co-authors: Irene Gijbels, Gilles Stupfler
Quantiles and expectiles of a distribution are found to be useful descriptors of its tail in the same way as the median and mean are related to its central behavior. A valuable alternative class to expectiles, called extremiles, is considered which parallels the class of quantiles and includes the family of expected minima and expected maxima. The new class is motivated via several angles, which reveals its specific merits and strengths. Extremiles suggest better capability of fitting both location and spread in data points and provide an appropriate theory that better displays the interesting features of long-tailed distributions. We discuss their estimation in the range of the data and beyond the sample maximum. Some motivating examples are given to illustrate the utility of estimated extremiles in modeling noncentral behavior. There is in particular an interesting connection with coherent measures of risk protection.

E0237: **Sequential monitoring of the tail behavior of dependent data**
Presenter: **Dominik Wied**, University of Cologne, Germany
Co-authors: Yannick Hoga
A sequential monitoring procedure for changes in the tail index and extreme quantiles of beta-mixing random variables is constructed which can be based on a large class of tail index estimators. The assumptions on the data are general enough to be satisfied in a wide range of applications. In a simulation study empirical sizes and power of the proposed tests are studied for linear and non-linear time series. Finally, we use our results to monitor Bank of America stock log-losses from 2007 to 2012 and detect changes in extreme quantiles without an accompanying detection of a tail index break.

E0418: **Estimation of conditional extreme risk measures from heavy-tailed elliptical random vector**
Presenter: **Antoine Usseglio-Carleve**, Institut Camille Jordan, France

In recent years, the question of estimating extreme quantiles, or more generally extreme risk measures has seen many advances. We consider an elliptical random vector, and focus on the extreme quantiles of a component conditioned by all the others. For that purpose, we start by recalling the main properties of elliptical distributions, especially the consistency property. Then we introduce a heavy-tail assumption on the marginal distributions. Once the frame has been defined, we propose in a first time an asymptotic relationship between conditional and unconditional quantiles, based on two parameters, called extremal parameters. Under our assumption, we easily provide estimators for both extremals parameters, and give their asymptotic distribution. Then, using the regular variation properties induced by our assumption, we deduce estimators for extreme conditional quantiles, and give some simple conditions for consistency. On the other hand, a stronger assumption and other conditions are required for asymptotic normality. A simulation study with a Student vector is proposed, in order to compare our estimators with theoretical results. The choice of the sequences for the tail index and kernel estimators and quantile level is also discussed through boxplots. We conclude by showing that many extreme risk measures may be deduced from extreme quantiles, like Haezendonck-Goovaerts risk measures, or Lp-quantiles. A financial data example is finally proposed.

E0530: **Computing value-at-risk via peaks-over-threshold generalized Pareto distribution**
Presenter: **Yi He**, Monash University, Australia
Co-authors: Liang Peng

The value-at-risk of financial loss in the tail is computed by fitting a generalized Pareto distribution to exceedance over a high but not divergent threshold. Such a model is inferred for both independent observations and time series data. We show that asymptotic variance for the maximal likelihood estimator depends on the choice of threshold, the tail index of the distribution, and the parameters of time-series model, which all make it quite challenging to quantify the uncertainty of high-level value-at-risk measure. To make the inference practically feasible, we then propose a smooth empirical likelihood based method for constructing a confidence interval for the value-at-risk based on either independent errors or AR-GARCH errors. The finite sample performance of the derived confidence intervals is demonstrated through numerical studies before applying to real data.

EO088 Room O1 DEPENDENCE MODELS AND COPULAS I Chair: Fabrizio Durante

E0388: **On extreme value copulas and concordance measures**
Presenter: **Piotr Jaworski**, University of Warsaw, Poland

The concordance measures, like for example Kendall tau, Spearman rho or Blomquist beta, are the main numerical characterization of Bivariate Extreme Value (BEV) copulas. We will provide the bounds for the sets of BEV copulas with a fixed concordance measure. Furthermore, we are going to show that for any two continuous symmetric concordance measures, κ_1 and κ_2, there exists a mapping Ψ, such that the sets of BEV copulas with $\kappa_1 = x$ and $\kappa_2 = \Psi(x)$, "almost" coincide with each other.

E0447: **On Kendall's tau for order statistics**
Presenter: **Sebastian Fuchs**, TU Dortmund, Germany
Co-authors: Klaus D Schmidt

Using Kendall's tau of the corresponding copulas, we compare the dependence structure of a random vector $(X_1,\dots,X_d)$ with identical univariate marginals and that of its order statistic $(X_{1:d},\dots,X_{d:d})$. Although the corresponding copulas are in general not comparable with respect to pointwise or concordance order, it turns out that the value of Kendall's tau of the copula for the order statistic is always at least as large as that of the copula for the random vector. In the case where the univariate marginals are not only identical but also independent, we further calculate Kendall's tau for $(X_{1:d},\dots,X_{k:d})$ with $2 \le k \le d$ and show that this value is identical with that for $(X_{d-k+1:d},\dots,X_{d:d})$.

E0795: **Maximum asymmetry of copulas revisited**
Presenter: **Noppadon Kamnitui**, University of Salzburg, Austria
Co-authors: Juan Fernandez Sanchez, Wolfgang Trutschnig

Motivated by the nice characterization of copulas A for which $d_\infty(A,A^t)$ is maximal, we study maximum asymmetry with respect to the conditioning-based metric D_1. Despite the fact that $D_1(A,A^t)$ is generally not straightforward to calculate, it is possible to provide both, a characterization and a handy representation of all copulas A maximizing $D_1(A,A^t)$. This representation is then used to prove the existence of copulas with full support maximizing $D_1(A,A^t)$. A comparison of D_1- and d_∞-asymmetry including some surprising examples is provided.

E0454: **A note on duality theorems in mass transportation**
Presenter: **Pietro Rigo**, University of Pavia, Italy

In the framework of mass transportation, let $(\mathcal{X},\mathcal{F},\mu)$ and $(\mathcal{Y},\mathcal{G},\nu)$ be probability spaces and $c : \mathcal{X}\times\mathcal{Y} \to \mathbb{R}$ a measurable cost function such that $f_1 + g_1 \le c \le f_2 + g_2$ for some $f_1, f_2 \in L_1(\mu)$ and $g_1, g_2 \in L_1(\nu)$. Define $\alpha(c) = \inf_P \int c\,dP$ and $\alpha^*(c) = \sup_P \int c\,dP$, where inf and sup are over the probabilities P on $\mathcal{F}\otimes\mathcal{G}$ with marginals μ and ν. A few duality theorems for $\alpha(c)$ and $\alpha^*(c)$, not requiring μ or ν to be perfect, are proved. As an example, suppose $\mathcal{X}$ and $\mathcal{Y}$ are metric spaces and at least one of μ and ν is separable. Then, duality holds for $\alpha(c)$ (for $\alpha^*(c)$) provided c is upper-semicontinuous (lower-semicontinuous). Moreover, duality holds for both $\alpha(c)$ and $\alpha^*(c)$ if the sections $x \mapsto c(x,y)$ and $y \mapsto c(x,y)$ are continuous. This improves the existing results if c has continuous sections and the cardinalities of $\mathcal{X}$ and $\mathcal{Y}$ do not exceed the continuum. Finally, the duality problem is investigated in a finitely additive setting. In this case, if c is bounded, some results by Ruschendorf are generalized.

EO679 Room P1 BRANCHING PROCESSES: THEORETICAL, APPLIED AND COMPUTATIONAL ISSUES I Chair: Miguel Gonzalez Velasco

E1290: **Self-regulating two-sex branching processes**
Presenter: **Cristina Gutierrez Perez**, University of Extremadura, Spain
Co-authors: Miguel Gonzalez Velasco, Rodrigo Martinez Quintana

The standard two-sex branching process is a discrete-time process which models the evolution of a two-sex population that evolves without any restriction and in which there are females and males that form couples in order to produce females and males offspring. This process has been modified in several directions (immigration, control, varying environment, etc.) in order to obtain more applicable models, but all of them verify the dichotomy extinction-explosion usual in many branching processes. Sometimes this fact makes these models unsuitable for applications to real biological situations. We introduce a self-regulating two-sex branching process with the aim of studying the evolution of a population in which the total population size is limited by the environment. For that, we define a discrete-time two-sex branching process where at each generation only some selected couples survive to produce offspring. We introduce a random control function at mating time to make this selection. This function will depend on the number of couples initially formed in the generation and on an associated survival probability which will be related to the carrying capacity of the environment. For this model, we show the behavior of the process in long term by means of simulations and we present some results about the extinction of the population.

E1323: **ABC methodology for controlled branching processes**
Presenter: **Ines M. del Puerto**, University of Extremadura, Spain
Co-authors: Miguel Gonzalez Velasco, Carmen Minuesa Abril
Controlled branching processes are stochastic growth population models in which the number of individuals with reproductive capacity in each generation is controlled by a random control function. The purpose is to examine the Approximate Bayesian Computation (ABC) methods and to propose appropriate summary statistics for them in the context of these processes. This methodology enables to approximate the posterior distribution of the parameters of interest satisfactorily without explicit likelihood calculations and under a minimal set of assumptions. In particular, the tolerance rejection algorithm, the sequential Monte Carlo ABC algorithm, and a post-sampling correction method are provided. The accuracy of the proposed methods are illustrated and compared with a "likelihood free" Markov chain Monte Carlo technique by the way of a simulated example developed with the statistical software R.

E1352: **Branching random walks in non homogeneous environments**
Presenter: **Elena Yarovaya**, Lomonosov Moscow State University, Russia
Nowadays it is commonly accepted that branching random walks are crucially useful in investigations of stochastic systems with birth, death and migration of their elements. The principal attention will be paid to the properties of branching random walks on multidimensional lattices. We will be mainly interested in the problems related to the limiting behavior of branching random walks such as existence of phase transitions under change of various parameters, the properties of the limiting distribution of the particle population, existence and the shape of the propagating fronts of particles, etc. The answer to these and other questions heavily depend on numerous factors which affect the properties of a branching random walk. Therefore, we will try to describe, how the properties of a branching walk depend on the fact of non homogeneity of the branching media, on the number and mutual disposition of the branching sources, and also on such properties of a branching walk as its symmetry and finiteness or infiniteness of the variance of jumps. We present also some results of simulation of branching random walks and discuss how they may be applied to numerical estimation of various characteristics describing the properties of the phase transitions.

E1386: **On periodic branching processes with immigration**
Presenter: **Marton Ispany**, University of Debrecen, Hungary
Recently, there has been considerable interest in integer-valued time series models for analyzing data sets which consist of counts of events, objects or individuals. Several integer-valued time series models proposed in the literature are based on the branching model. However, these models do not account the periodic characteristic observed in some real series. We consider a branching process with immigration (BPI) in time varying environment where the environment of the process is periodic of period S, i.e., the mean and variance functions of the offspring and immigration distributions are periodic function with period S. We can interpret the process as the size of a population, e.g., the number of traffic accidents, hospital admissions, attacks on computer systems, transactions in transaction processing systems. A classification theorem is proved for periodic BPIs. Statistical properties of the process such as mean, variance, autocovariance function and marginal distributions are discussed. Moment-based conditional least squares and conditional maximum likelihood estimates of the parameters are presented for the subcritical and critical cases, respectively. Numerical estimation procedures are proposed and their performances are investigated through Monte Carlo simulations.

EO611 Room Q1 RECENT ADVANCES IN ROBUST MODELLING Chair: Eva Cantoni

E0833: **Robust MM estimation for imperfect regression discontinuity designs**
Presenter: **Ben Hansen**, University of Michigan, United States
Co-authors: Adam Sales
In a regression discontinuity design (RDD), assignment to treatment versus a control condition is determined by the value of a particular baseline variable, R. In one recent RDD, R is the average of a student's grades in his first year at university; the treatment condition is academic probation, forced upon a student if his R falls below a threshold; and downstream effects of the academic probation regime are estimated using ordinary least squares. Some cutting-edge RDD methods contrast limits of $\mathrm{E}(Y|R=r)$ as r approaches a cut-point, c, from either side; others avoid passing to limits by supposing that in sufficiently narrow neighborhoods of the threshold, there is random assignment. Both frameworks are difficult to reconcile with examples such as the academic probation study, where tests for manipulation of the running variable find the experimental analogy to be at its weakest in the immediate vicinity of the cut-point. Our method addresses these challenges with a combination of tactics: weakening the local randomization assumption; decontaminating the sample with the aid of specification tests; and robust MM-estimation. The MM-estimators' insensitivity to contamination imparts a unique robustness to leading validity threats facing RDDs.

E1064: **A robust version of GAMLSS**
Presenter: **Rosalba Radice**, Cass Business School, United Kingdom
Co-authors: Eva Cantoni, Giampiero Marra, William Aeberhard
A robust version of generalised additive models for location, scale and shape is discussed where any parameter of the distribution can be specified as function of additive predictors allowing for several types of covariate effects (e.g., linear, non-linear, random and spatial effects). The estimation approach permits all models parameters to be estimated robustly by limiting the influence of deviating datapoints on each log-likelihood contribution. We evaluate the empirical performance of the proposed method through simulation experiments. We also illustrate the use of this approach on functional magnetic resonance imaging measurements for a human brain subject to a particular experimental stimulus.

E0401: **Robust semiparametric inference with missing data**
Presenter: **Xavier de Luna**, Umea University, Sweden
Co-authors: Eva Cantoni
Semiparametric inference with missing outcome data is based on partially specified models which are not of direct interest (e.g., model for missingness mechanism). Different classes of estimators exist, which are more or less robust to misspecification of these models. Another type of threat to the validity of the inference occurs in situations where some observations are contaminated (generated by some nuisance distribution). Classical semiparametric inference is not robust to such contamination, and a single observation may have an arbitrary large effect on bias as measured by the influence function. We introduce inverse probability weighted, double robust and outcome regression estimators of location and scale parameters, which are robust to contamination in the sense that their influence function is bounded. We give asymptotic properties and study finite sample behaviour. Our simulated experiments show that contamination can be more serious a threat to the quality of inference than model misspecification. An interesting aspect of our results is that the auxiliary outcome model used to adjust for ignorable missingness is also useful to protect against contamination. We also illustrate through a case study how both adjustment to ignorable missingness and protection against contamination are achieved through weighting schemes, which can be contrasted to gain further insights.

E0768: **Robust causal inferences in small area estimation**
Presenter: **Setareh Ranjbar**, HEC Lausanne, Switzerland
Co-authors: Nicola Salvati, Barbara Pacini
When doing impact evaluation and making causal inferences in many cases, it is important to acknowledge the heterogeneity of the treatment effects

for different domains. Where certain geographic, socio-demographic, or socio-economic unplanned domains may benefit from a program/policy intervention, others may be worse off. If the domain for which we are interested in the impact, is small with regards to its sample size (or even zero in some cases), then the evaluator has entered the small area estimation (SAE) dilemma. In addition small area estimators are intrinsically very sensitive to the presence of outliers due to the small sample sizes. Therefore it is important to develop or make use of robust methods in SAE. Based on the modification of inverse propensity weighting and the robust small area estimators, we propose new methods that allows one to robustly estimate the area specific average treatment effects for the unplanned domains. The Mean Squared Error (MSE) of the proposed predictors are analytically approximated for the situations that propensity scores are taken as known and a bias calibration method is also provided. By means of these methods we can provide a map of policy impacts that can help to better target the treatment group(s).

EO621 Room C2 STATISTICAL METHODS FOR RISK MANAGEMENT IN FINANCE AND INSURANCE Chair: Hideatsu Tsukahara

E0390: **Asymptotically normal estimators of the ruin probability for Levy insurance risks**
Presenter: **Yasutaka Shimizu**, Waseda University, Japan

A statistical inference for ruin probability from a certain discrete sample of the surplus is discussed under a Levy insurance risk. The surplus model is a spectrally negative Levy process with diffusion terms and possibly infinite activity jumps. We assume that the observations are discrete equidistant records of the surplus, and large jumps (claims). We consider the Laguerre series expansion of the ruin probability and give an estimator of the partial sum, which is an approximation of the ruin probability in L2-sense. Under a high-frequent observation scheme, we show the asymptotic normality of the proposed estimator with the estimable asymptotic variance. This estimator enables not only a point estimation of ruin probability, but also an interval estimation and a testing hypothesis.

E0650: **On Hawkes processes**
Presenter: **Matthias Kirchner**, ETH RiskLab Switzerland, Switzerland
Co-authors: Paul Embrechts

Event streams are shown to have become an increasingly important data category. The mathematical counterparts of empirical event stream data are point processes on the real line. Hawkes point processes form a most successful model class. We motivate and explain the model. In finance and insurance, one typically considers multitype event stream data. e.g., different orders in limit order books, price jumps of different stocks, or credit defaults in different industries. We illustrate how in these kinds of Hawkes process applications, graphical description of data and models turns out to be useful.

E0758: **Comparison of EVT methods for GARCH-EVT approach applied to financial time series**
Presenter: **Hibiki Kaibuchi**, SOKENDAI The Graduate University of Advanced Studies, Japan
Co-authors: Yoshinori Kawasaki

Managing extreme event risk in finance and insurance is vital in our modern society. It is known that the statistically justifiable modeling and prediction of rare events are challenging because the historical data on extreme events are inherently scarce. In order to prevent or prepare for unfavorable scenarios, the approaches based on extreme value theory (EVT) have been devised. The aim is to estimate conditional extreme quantiles (Value at Risk) using GARCH-EVT framework. For that, we: (i) pre-whiten the financial time series with a GARCH-type model for forecasting volatility; (ii) apply the semi-parametric bias-corrected tail estimators under β-mixing condition to the residuals from the GARCH analysis instead of the Peaks-Over-Thresholds (POT) method under IID condition. The results are illustrated on simulated data and on a financial real dataset.

E0784: **Value-at-risk and expected shortfall of stock portfolio using skew-*t* copulas**
Presenter: **Toshinao Yoshiba**, Institute of Statistical Mathematics, Japan

In financial portfolio risk management, student-*t* copula is frequently used to capture the tail dependence of risk factors. Azzalini-Capitanio (AC) and Generalized Hyperbolic (GH) skew-*t* copulas are considered to incorporate asymmetric tail dependence of risk factors. The estimated parameters of the AC skew-*t*, GH skew-*t*, student-*t*, normal copulas by maximum likelihood are examined for the daily returns of three TOPIX sector indices. After the skewness of skew-*t* copulas is validated both for the unfiltered returns and for the filtered returns by GARCH and EGARCH models, the effect of the skewness on the value-at-risk and expected shortfall of the selected stock portfolio is investigated.

EO246 Room O2 CSDA JOURNAL: CLUSTERING AND MIXTURE MODELS Chair: Peter Rousseeuw

E0873: **Dealing with missing data in model-based clustering through a MNAR model**
Presenter: **Christophe Biernacki**, Inria, France
Co-authors: Gilles Celeux, Julie Josse, Fabien Laporte

Since the 90s, model-based clustering is largely used to classify data. Nowadays, with the increase of available data, missing values are more frequent. Traditional ways to deal with them consist in obtaining a filled data set, either by discarding missing values or by imputing them. In the first case, some information is lost; in the second case, the final clustering purpose is not taken into account through the imputation step. Thus, both solutions risk to blur the clustering estimation result. Alternatively, we defend the need to embed the missingness mechanism directly within the clustering modeling step. There exists three types of missing data: missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). In all situations logistic regression is proposed as a natural and flexible candidate model. In particular, its flexibility property allows us to design some meaningful parsimonious variants, as dependency on missing values or dependency on the cluster label. In this unified context, standard model selection criteria can be used to select between such different missing data mechanisms, simultaneously with the number of clusters. Practical interest of our proposal is illustrated on data derived from medical studies suffering from many missing data.

E1314: **Total sum of squares decomposition for mixtures of regressions**
Presenter: **Salvatore Ingrassia**, University of Catania, Italy
Co-authors: Antonio Punzo

A three-term decomposition of the total sum of squares is proposed for mixtures of linear regressions whose parameters are estimated by maximum likelihood, via the expectation-maximization algorithm, under normally distributed errors in each mixture component. In particular, three types of mixtures of regressions are considered: with fixed covariates, with concomitant variables, and with random covariates. A ternary diagram is also suggested to make easier the joint interpretation of the three terms of the proposed decomposition. Furthermore, local and overall coefficients of determination are respectively defined to judge how well the model fits the data group-by-group but also taken as a whole. Artificial data are considered to find out more about the proposed decomposition, including violations of the model assumptions. Finally, an application to real data illustrates the use and the usefulness of these proposals.

E1572: **Some thoughts on simulation studies to compare clustering methods**
Presenter: **Christian Hennig**, UCL, United Kingdom

Simulation studies are often used to compare different clustering methods, be it with the aim of promoting a new method, or for investigating the quality of existing methods from a neutral point of view. A number of aspects of designing and running such studies will be discussed, including the definition and measurement of clustering quality, the choice of models to generate data from, aggregation and visualisation of results, and also limits of what we can learn from such studies. Some aspects will be relevant also for the design of simulation studies in wider areas of statistics. The material was partly developed in collaboration with the IFCS cluster benchmarking task force.

E0505: **Automated learning of mixtures of factor analyzers with missing information**
Presenter: **Tsung-I Lin**, National Chung Hsing University, Taiwan

The mixtures of factor analyzers (MFA) model is a powerful tool which provides a unified approach to dimensionality reduction and model-based clustering of heterogeneous data. In seeking the most appropriate number of factors (q) of a MFA model with the number of components (g) fixed a priori, a two-stage procedure is commonly performed by first carrying out parameter estimation over a set of prespecified numbers of factors, and then selecting the best q according to certain penalized likelihood criteria. When the dimensionality of data grows higher, such a procedure can be computationally prohibitive. To overcome this obstacle, we develop an automated learning scheme to effectively merge parameter estimation and selection of q into a one-stage algorithm. The proposed new learning procedure that allows for much less computational cost is called the automated MFA (AMFA) algorithm, and our development is also extended to accommodate missing values. In addition, we explicitly derive the score vector and the empirical information matrix for inferring standard errors associated with the estimated parameters. The potential and applicability of the proposed method are demonstrated through a number of real datasets with genuine and synthetic missing values.

EO420 Room Q2 BAYESIAN MODELLING AND COMPUTATION Chair: Bernardo Nipoti

E0402: **New insights on Bayesian graphs and neural networks from distributional properties**
Presenter: **Julyan Arbel**, Inria, France
Co-authors: Florence Forbes, Mariia Vladimirova, Hongliang Lu

Ongoing work on Bayesian modeling of (1) graphs and (2) neural networks is considered. Part (1) is devoted to Bayesian nonparametric modeling of data structured as a graph. In such a setting, the usual assumption of exchangeability does not hold. We rely on a Potts component in the prior in order to account for graph dependencies. Such a prior induces a distribution on partitions akin to the celebrated Chinese restaurant process. We derive distributional properties which highlight the Potts contribution to the clustering mechanism. Part (2) focuses on distributional results of Bayesian neural networks. We derive some new non-asymptotic results highlighting that the deeper the network layer, the heavier-tailed the unit distribution.

E0493: **Conjugate Bayes for probit regression via unified skew-normals**
Presenter: **Daniele Durante**, Bocconi University, Italy

Regression models for dichotomous data are ubiquitous in statistics. Besides being useful for inference on binary responses, such methods are also key building-blocks in more complex formulations, covering density regression and nonparametric classification, among others. Within the Bayesian setting, inference typically proceeds by updating the Gaussian priors for the coefficients with the likelihood induced by probit or logit regressions for the binary responses. In this updating, the apparent absence of a tractable posterior has motivated a variety of computational methods, including MCMC routines and algorithms which approximate the posterior. Despite being routinely implemented, current MCMC methodologies face mixing or time-efficiency issues in large p and small n studies, whereas approximate routines fail to capture the skewness typically observed in the posterior. It is shown that the posterior distribution for the probit coefficients has a unified skew-normal kernel, under Gaussian priors. This result allows fast and accurate Bayesian inference for a wide class of applications, especially in large p and small-to-moderate n studies where state-of-the-art computational methods face substantial issues. These notable advances are quantitatively outlined in a genetic study, and further motivate the development of a wider class of conjugate priors for probit regression along with a novel independent sampler from the unified skew-normal posterior.

E0579: **WARP: Wavelets with adaptive recursive partitioning for multi-dimensional data**
Presenter: **Li Ma**, Duke University, United States
Co-authors: Meng Li

Traditional statistical wavelet analysis carries out modeling and inference under a given, predetermined wavelet transform. This approach can quickly lose efficiency for multi-dimensional data (e.g., observations measured on a multi-dimensional grid), because a predetermined transform does not exploit the structure of the underlying function in a problem-specific manner. We overcome this challenge by making the wavelet transform adaptive to the structure of the data. By exploiting a connection between permutations on the index space of multi-dimensional functions and recursive partitions on that space, we show that the desired adaptivity in the wavelet transform can be achieved through Bayesian hierarchical modeling on the space of such recursive partitions. When one applies this framework to Haar wavelets, exact Bayesian inference under the model can be achieved analytically through recursive message passing with an efficient computational complexity linear in the sample size. We also provide recipes for incorporating block shrinkage into the framework as well as for applying it to other wavelet bases. We demonstrate via numerical experiments that with the enhancement under this framework even simple Haar wavelets can achieve excellent performance in 2D and 3D image reconstruction. We investigate the source of the gain by quantitatively comparing the efficacy of energy concentration under our adaptive wavelet transforms to that of classical fixed wavelet transforms.

E0929: **(Exact) Bayesian inference for hidden Markov models via duality and approximate filtering distributions**
Presenter: **Guillaume Kon Kam King**, University of Torino, Italy
Co-authors: Omiros Papaspiliopoulos, Matteo Ruggiero

Filtering hidden Markov models, or sequential Bayesian inference on the hidden state of a signal, is analytically tractable only for a handful of models. Examples are finite-dimensional state space models and linear Gaussian systems (Baum-Welch and Kalman filters). Recently, a principled approach has been proposed for extending the realm of analytically tractable models, exploiting a duality relation between the hidden process and an auxiliary process. Then, the solution of the filtering problem consists in a finite mixture of distributions. We study the computational effort required to implement this strategy for two parametric and nonparametric models: the Cox-Ingersoll-Ross process, the K-dimensional Wright-Fisher process, the Dawson-Watanabe process and the Fleming-Viot process. In all cases, the number of components involved in the filtering distributions increases rapidly with the number of observations. Although this could render the algorithm impractical for long observation sequences and undermine its practical relevance, the mathematical form of the filtering distributions suggest that the number of components which contribute most to the mixture remains small. This suggests several efficient natural approximation strategies. We assess the performance of these strategies in terms of accuracy, speed and prediction, benchmarked against the exact solution.

CI013 Room A0 EMPIRICAL MACROECONOMICS Chair: Michael Owyang

C0160: **Estimating the aggregate effects of border adjustments**
Presenter: **Nora Traum**, HEC Montreal, Canada
Co-authors: Matteo Cacciatore

An estimated, quantitative international business-cycle model is demonstrated to be able to account simultaneously for international cross-country comovements, exchange rate dynamics and trade flows. We estimate the model using Bayesian methods with U.S. and trade-weighted aggregates for the rest of the world. We show the estimated model implies a substantial influence of U.S. sourced disturbances on the rest of the world, helping reconcile cross-country comovement between the model and data. We then combine Bayesian prior and posterior analyses to quantify the effects of the U.S. adopting a border-adjustment tax (BAT). While the model is a priori uncertain about the size and sign of the effects of a BAT on GDP, posterior estimates significant, small contraction.

C0161: **Measuring GDP growth data uncertainty**
Presenter: **Ana Galvao**, University of Warwick, United Kingdom
Co-authors: James Mitchell

Economic statistics are prone to data uncertainty since they are subject to both sampling and non-sampling errors. GDP estimates are regularly revised as new information is received and methodological improvements are made. Although the Office for National Statistics, in the UK, emphasise that initial estimates of GDP values will be revised, it is the Bank of England that provide quantitative estimates of the likely uncertainty around past GDP values. We find that the revision mean and measurement error volatility are time varying for both UK and US GDP growth. We evaluate the Bank of England's predictive densities for revised GDP growth values; and show that their point predictions better anticipate mature ONS growth estimates than the ONS's own first releases. Their density estimates are on average well-calibrated, but this masks changes in predictive performance. We propose a measure of data uncertainty that removes the forecastable component of data revisions as predicted by the Bank of England's backcast densities. We show that data uncertainty jumps at the onset of the 2008/2009 recession and contribute to macroeconomic uncertainty.

C0162: **Tax progressivity**
Presenter: **Michael Owyang**, Federal Reserve Bank of St Louis, United States
Co-authors: Laura Jackson Young, Christopher Otrok, Nora Traum

The empirical analysis of tax policy has generally been limited to the study of the effect of changes in tax revenue. This approach may ignore important distributional effects. We consider the effect of revenue neutral increases in tax progressivity. Greater progressivity implies a higher tax burden for higher incomes and a lower tax burden for lower incomes. We develop an empirical measure of progressivity jointly with a measure of the tax revenue level. We find that a shock to the level of taxes produces similar effects as has been found in the previous literature. Moreover, we find that a positive shock to tax progressivity increases output, suggesting that lowering taxes on the low end of the income distribution is more expansionary than raising taxes on the high end. We then consider the channels through which tax progressivity can affect output, consumption, and inequality.

CO084 Room B2 REGIME CHANGE MODELING II Chair: Willi Semmler

C1256: **Nonlinear credit dynamics, regime switches in the output gap and credit shocks evaluated through local projections**
Presenter: **Francesco Simone Lucidi**, Sapienza University, Italy
Co-authors: Willi Semmler

As much research has recently shown, credit cycles are linked to financial instability with large effects on the real economy. On the other hand, central banks tried to stimulate the economy with credit policies. Empirically, in several euro area countries, one can observe an inverse long-run relation of credit spreads and the output gap - credit spread falling with a positive output gap and rising with a negative output gap. We build a small scale nonlinear quadratic model (NLQ) to study how credit flows and credit spreads are linked to the two regimes of the output gap. Then, we empirically estimate the impulse responses for an exogenous credit supply shock through local projections and propose a new external instrument to identify the dynamic causal effects of the structural shock. The theoretical model demonstrates that credit dynamics may lead the system to a new regime, allowing for expansions but also triggering instability through sudden jumps in credit spreads. In the empirical part, we allow dynamic multipliers to smoothly pass from periods of sustained growth to periods of deep recession so to catch empirically the nonlinear features of the theoretical model. We find that credit supply shocks have a significant impact on real activities, credit spreads, stock prices and house prices. Furthermore, the revealed regime dependence of such effects may provide further information to monetary authorities in setting credit-oriented policies.

C1260: **Synchronization patterns in the European Union**
Presenter: **Mattia Guerini**, OFCE - SciencesPo, France
Co-authors: Mauro Napoletano, Duc Thi Luu

A novel approach is proposed for investigating the synchronization of business cycles. We apply it to a EUROSTAT database that describes the manufacturing industrial production in the EU and that covers the 2000-2016 period at a monthly frequency. We employ the method of Random Matrix Theory (RMT), which is now commonly used for the study of cross-correlations between stock-indexes. Our empirical exercise traces the evolution of the European synchronization patterns and identifies the emergence of synchronization clusters between different EU economies. Two main conclusions are drawn. First, synchronization in the Euro area increased during the first decade of the century, reached its peak during the crisis period, but decreased in the aftermath of the great recession, reverting to the levels observable at the beginning of the century. Second, different clusters of countries that coordinate well among them are identified: while in the early years of the century the clustering broke along the east-west dimension, the recession brought about a structural transformation and nowadays the break is evident alongside the north-south axis. We conclude that the recent a-synchronization process might be harmful for the EU because of the heterogeneous responses from the common policies that it might entail.

C1264: **Regime shifts in currency markets with bounded rationality and limits to arbitrage**
Presenter: **Soumya Datta**, South Asian University, India

The purpose is to examine the impact of limits to arbitrage in currency markets with heterogeneous boundedly rational agents, under both complete and incomplete information. Departing from existing literature, we show that limits to arbitrage become more restrictive for larger deviation from the fundamentals. This leads to multiple equilibria. Under complete information, either everyone or no one attempts arbitrage. Under incomplete information, however, there might be a continuum of agents attempting arbitrage. In this case, a regime shift between fundamental and non-fundamental steady states is possible. We also find that the prior beliefs are not important in case of small noises and uniform priors; however, for larger noises and nonuniform priors, prior beliefs might acquire a greater role. In this case, in a departure from efficient markets, large fund managers might have a disproportionately large power in influencing the outcome.

C1242: **To what extent globalization affects inflation: The role of global value chains**
Presenter: **Jacek Kotlowski**, Narodowy Bank Polski, Poland
Co-authors: Aleksandra Halka, Jan Hagemejer
The aim is to investigate to what extent the globalization process reflected by the increasing participation in the global value chains influences the inflation in various economies. We use the broad panel data set based on World Input-Output Database (WIOD) to check whether the growing trade in intermediate goods first results in lowering domestic inflation and second leads to decrease of exchange rate pass-through (ERPT). We also decompose the exchange rate pass-through to consumer prices into direct and indirect (via domestic producer prices) channel. We find that the growing participation in the GVC results in the reduction of the ERPT as well as the in lowering of the level of overall inflation.

CO060 Room E2 PENALIZED, NONPARAMETRIC, SPATIAL AND CONTAMINATED MODELS Chair: Artem Prokhorov

C0244: **A ridge to homogeneity**
Presenter: **Stanislav Anatolyev**, CERGE-EI and New Economic School, Czech Republic
In some heavily parameterized econometric models, one may benefit from shrinking a subset of parameters towards a common target. We consider L_2 shrinkage towards an equal parameter value that balances between unrestricted estimation (i.e., allowing full heterogeneity) and estimation under equality restriction (i.e. imposing full homogeneity). The penalty parameter of such ridge regression estimator is tuned using one-leave-out cross-validation. The reduction in predictive mean squared error tends to increase with the dimensionality of the parameter set. We illustrate the benefit of such shrinkage with a few stylized examples. We also work out, both theoretically and empirically, a heterogenous linear panel data setup and compare several estimators and corresponding confidence intervals.

C1093: **Fixed effects estimation in single-index models**
Presenter: **Daniel Henderson**, University of Alabama, United States
Single index models with fixed effects are estimated. We show that our estimators are consistent and asymptotically normal. Monte Carlos simulations support the asymptotic development. An empirical example is given to show how the methods work in practice.

C1474: **Parametric stochastic frontier models and probability statements with spatial errors**
Presenter: **Ian Wright**, University of Miami, United States
Co-authors: William Horrace, Christopher Parmeter
The presence of spatial correlation in the error terms for stochastic frontier models yields an intractable likelihood where the number of integrals grows with the sample size. Sequential conditioning is used in order to factor the joint distribution which produces a likelihood that has only a single integral irrespective of the sample size. This leads to a likelihood function which is numerically more feasible to solve. Additionally, certain probability statements are generalized to account for spatial dependence. Maximum likelihood estimates are discussed.

C0224: **A contaminated extended Weibull distribution and its applications**
Presenter: **Boris Choy**, University of Sydney, Australia
An extension of the class of contaminated Weibull distribution for statistical modelling is proposed. The extra parameter in the proposed distribution allows for more flexibility in the estimation of moments and the modelling of the tail behaviour. The properties of the distribution are presented and compared with the existing contaminated Weibull distribution. Bayesian simulation-based methods will be used for statistical inference. We shall demonstrate the adequacy of the proposed distribution in modelling positively-valued data in various applications.

CO298 Room F2 EMPIRICAL STUDIES OF FINANCIAL MARKETS WITH HIGH-FREQUENCY DATA Chair: Teruo Nakatsuma

C0341: **Efficient, transaction and mid prices: Disentangling sources of high frequency market microstructure noise**
Presenter: **Simon Clinet**, Keio University, Japan
Co-authors: Yoann Potiron
An extension of the Roll model is considered where the trade direction, i.e. whether the trade is buyer or seller initiated, is multiplied by the dynamic quoted half-spread. Employing tick-by-tick maximum likelihood estimation on S&P 500 constituents, we find that the efficient price is quite close yet not equal to the mid price and lies systematically on the side of the trade, suggesting that there is additional information in trade direction. Moreover, we document that the variability in the model is relatively small, indicating that the combination of the trade direction and quoted bid-ask spread plays the major role in explaining market microstructure noise. Among different observable high frequency financial characteristics of the underlying stocks, this variability is best explained by the tick-to-spread ratio, implying that discreteness is the second source of noise. We determine the bid-ask bounce effect as the third source of noise.

C0640: **Factor multivariate realized stochastic volatility model**
Presenter: **Yuta Yamauchi**, University of Tokyo, Japan
Co-authors: Yasuhiro Omori
Although modelling time-varying volatility and correlations of multivariate asset returns is one of most important problems in the financial risk management, it has been difficult to obtain stable inference of covariance of asset returns due to the high dimensionality of parameters in dynamic covariance structure. One major solution to reduce the number of parameters is to introduce factor structure assuming that a small number of common factors describe the dynamics of time-varying covariance matrices as discussed in the factor stochastic volatility models. We propose parsimonious modelling of multivariate asset returns based on dynamic factor stochastic volatility models with leverage effect, which allow dynamic latent factors. Firstly, to stabilize the estimation and prediction of time-varying parameters, we incorporate additional observations based on intraday asset returns and market indices. We use realized measures for covariance of asset returns based on intraday asset returns such as realized covariances, and latent factors of asset returns such as market indices. Secondly, we reduce the number of parameters of leverage effect, omitting the leverage effect between each asset and each volatility of asset. We only introduce leverage effect between each latent factor and each volatility of asset.

C0718: **Bayesian analysis of intraday stochastic volatility models with skew heavy-tailed error and smoothing spline seasonality**
Presenter: **Teruo Nakatsuma**, Keio University, Japan
Co-authors: Makoto Nakakita
The aim is to extend the stochastic volatility (SV) model for application with intraday high frequency data of asset returns. It is well-known that intraday high frequency data of asset returns exhibit not only stylized characteristics (e.g., volatility clustering, heavy-tailed distribution) but also cyclical fluctuation in return volatility, which is called intraday seasonality. In a typical trading day, the volatility tends to be higher immediately after the opening or near the closing, but it tends to be lower in the middle of the trading hours. Our modeling strategy is two-fold. First, we model the intraday seasonality of return volatility with a B-spline polynomial and estimate it along with the stochastic volatility simultaneously. Second, we incorporate a possibly skew and heavy-tailed error distribution into the SV model by assuming that the error distribution belongs to a family

of generalized hyperbolic (GH) distribution such as Laplace, variance-gamma and Student's t. We develop an efficient Markov chain Monte Carlo (MCMC) sampling algorithm for Bayesian inference of the proposed model and apply it to intraday data of Japanese stocks.

C0755: **Stochastic conditional duration model with intraday seasonality and limit order book information**
Presenter: **Tomoki Toyabe**, Keio University, Japan
Co-authors: Teruo Nakatsuma
Intraday financial transactions are irregularly spaced, and their durations exhibit positive autocorrelation and intraday seasonality. In the literature, the former is formulated as a time-dependent duration model such as the stochastic conditional duration (SCD) model while the latter is dealt with by filtering out any cyclical fluctuations in time series of durations with a spline smoothing method before the duration model is estimated. We propose a Bayesian approach to model both autocorrelation and intraday seasonality in durations simultaneously. In our new approach, the autocorrelation structure of durations is captured by the SCD model while the intraday seasonality is approximated with B-spline smoothing. Moreover, in our new approach, it is straightforward to include limit order book information (e.g., bid-ask spread) as a covariate in the SCD model. The resultant model is regarded as a non-linear non-Gaussian state space model, for which a Bayesian approach is suitable. We developed an efficient Markov chain sampling scheme for the posterior analysis of the proposed model and applied it to high-frequency transaction data in the Tokyo stock exchange.

CO076 Room G2 NON-CAUSAL AND NON-GAUSSIAN TIME SERIES MODELS Chair: Alain Hecq

C0647: **Seasonal bubbles and volatility models**
Presenter: **Tomas del Barrio Castro**, University of the Balearic Islands, Spain
Co-authors: Alain Hecq, Sean Telg
Economic and financial bubbles usually emerge when prices strongly exceed the asset's intrinsic value namely the fundamental price. The reasons explaining this phenomenon are many (speculation, excessive monetary liquidity, moral hazard, extrapolation, etc.) and economic history is full of extraordinary examples from the Tulipomania to the more recent Brazilian inflation in the 80's, the dot.com, Bitcoin or the subprime crisis. The existence of a bubble is often evaluated retrospect however, after noticing on graphs the existence of a nonlinear pattern with an explosive episode that bursts at its climax. Recently, it has been shown that a linear stationary noncausal process, that is to say an autoregressive process in reverse time, is able to generate features such as bubble patterns. Our interest is to study the behavior of alternative bubble patterns arizing in the mixed causal and noncausal model (MAR) namely a dynamic specification with both leads and lags. It emerges indeed that complex polynomial roots in the MAR model makes the process similar to a process with time varying clustered volatility. We call this patterns seasonal or periodic bubbles. We show that one only needs to detect those roots in either lead or lag polynomial and as such the pseudo-causal representation does not provide additional effect.

C0664: **Predicting bubble collapse using non-causal models**
Presenter: **Elisa Voisin**, Maastricht University, Netherlands
Locally explosive episodes have long been observed in financial and economic time series and such dynamic features are well captured by mixed causal-non-causal autoregressive models. They incorporate both lags and leads of the variable of interest and are characterised by heavy-tailed distributions. The aim is to evaluate various methods' accuracy of predicting the probabilities of a crash during a bubble. Overall, three distributions are considered, a Cauchy and two Student-t, one with finite and the other one with infinite variance. Furthermore, numerical and, when possible, statistical approaches are employed. Conditional predictive densities do not always admit a closed-form expression; while for Cauchy, results can be obtained directly, for Student-t ones, approximations need to be made. To do so, sample- and simulations-based methods are used. The empirical analysis only considers MAR(r,1) processes, and uses the MAR(0,1) as a benchmark.

C0948: **Using quantile regressions for identifying mixed causal-noncausal models**
Presenter: **Alain Hecq**, Maastricht University, Netherlands
Co-authors: Li Sun
Mixed causal-noncausal models are gaining attention, for instance, in modelling bubbles and asymmetric cycles in economic and financial time series. The aim is to extend a previous model selection approach for purely causal and noncausal models to mixed models with both lags and leads components. The estimation of mixed causal-noncausal models is carried out using quantile autoregressions in both direct and reverse time. Our framework is able to choose among several parameter sets those that minimize the sum of absolute rescaled residuals. Monte Carlo simulations and an application on real data illustrate the feasibility of our approach.

C1279: **Now-casting financial volatility with long memory: A non-Gaussian and non-linear state space approach**
Presenter: **Yuze Liu**, University of Cologne, Germany, Germany
Recently, a simple model for now-casting daily financial log-volatility has been proposed. In contrast to existing approaches, it uses current and past information. It obeys an ARMA representation for log-squared returns and is related to the well-known SV model. This model outperforms the EGARCH model and the SV model. However, there are some important limitations. First, the normality assumption on daily returns is critical. Second, the ML estimation under a Gaussian approximation is biased and inefficient in finite-samples. Third, the typical long memory feature is not captured. These issues are tackled by considering a flexible non-Gaussian and non-linear ARMA representation of the log-transformed squared returns. The strong dependence and leverage in volatility are captured by an asymmetric ARFIMA model. It is estimated via the suitable Kitagawa state-space filter. It implements the numerical exact ML estimation under non-Gaussian and non-linear distributions. The employed estimation framework is flexible enough to cover the non-normality of daily returns explicitly. In an extensive Monte Carlo study, bias-reduction and efficiency gains are investigated. The volatility now-casting performance is evaluated by means of MSE and QLIKE. In an empirical application, volatility connectedness of US bond markets is studied in comparison to the commonly applied GARCH(1,1) model.

CO623 Room H2 ECONOMETRIC ANALYSIS OF THE BUSINESS CYCLE Chair: Matteo Luciani

C0480: **Estimating and accounting for the output gap with large Bayesian vector autoregressions**
Presenter: **Benjamin Wong**, Monash University, Australia
The aim is to demonstrate how Bayesian shrinkage can address problems with utilizing large information sets to calculate trend and cycle via a multivariate Beveridge-Nelson (BN) decomposition. We illustrate our approach by estimating the U.S. output gap with large Bayesian vector autoregressions that include up to 138 variables. Because the BN trend and cycle are linear functions of historical forecast errors, we are also able to account for the estimated output gap in terms of different sources of information, as well as particular underlying structural shocks given identification restrictions. Our empirical analysis suggests that, in addition to output growth, the unemployment rate, CPI inflation, and, to a lesser extent, housing starts, consumption, stock prices, real M1, and the federal funds rate are important conditioning variables for estimating the U.S. output gap, with estimates largely robust to incorporating additional variables.

C0743: **Measuring US aggregate output and output gap using large datasets**
Presenter: **Matteo Luciani**, Federal Reserve Board, United States
Co-authors: Matteo Barigozzi
The US aggregate output and output gap are estimated by using a non-stationary dynamic factor model estimated on a large dataset of macroeconomic indicators, together with a non-parametric trend-cycle decomposition. Two main results emerge from our analysis: first, since 2010, output growth was on average 0.4 percentage point higher than measured by GDP. Our measure's higher growth has been concentrated on the first quarter of the year, suggesting that weakness in the GDP's first-quarter growth over the past several years may be due to mis-measurement rather than problems with seasonal adjustment. Second, according to our output gap estimate, while growth in the years before the financial crisis was heavily boosted by temporary factors, hence not sustainable, growth after the financial crisis was mainly driven by permanent factors, thus suggesting that as of 2017:Q4 there is still slack in the economy.

C1120: **Advances in nowcasting economic activity**
Presenter: **Juan Antolin Diaz**, Fulcrum Asset Management, United Kingdom
Co-authors: Thomas Drechsel, Ivan Petrella
Dynamic factor models (DFM) have become the workhorse model for nowcasting economic activity. Exploiting recent advances in Bayesian computational methods, we extend the DFM framework along four dimensions. First, we model low-frequency movements in the growth rate and the volatility of the variables. Second, we allow for heterogeneous lead-lag patterns in the responses of the variables to the common factor. Third, we introduce automatic outlier detection by modeling fat tailed observations in the variables. Fourth, we endogenously model seasonal fluctuations, which is particularly useful whenever there is suspicion that "residual seasonality" is present. We then put our modeling innovations to the test in a comprehensive out-of-sample evaluation exercise using fully real-time unrevised data for seven countries. As the model is re-estimated each time new information arrives, the sheer scale of the exercise requires massive computational power and is made possible thanks to the use of cloud computing. Paying special attention to the production of well-calibrated density forecasts, we show how low frequency movements, dynamic heterogeneity, outliers and seasonality are pervasive features of macroeconomic data and their modeling advances our understanding of the real-time assessment of macroeconomic conditions.

C1254: **E pluribus, pauca: Measuring different dimensions of slack in the US economy with an agnostic model**
Presenter: **Gianni Amisano**, Federal Reserve Board, United States
Co-authors: Matteo Barigozzi, Matteo Luciani
Bayesian inference on the US output gap is taken up. This is done by using a model with a large number of macro indicators. The model parses each series into 3 components: 1) a common cyclical component; 2) a common trend component; 3) an idiosyncratic term to be considered as measurement error. Both the cyclical component and the trend components can be either scalars or vectors and we use statistical criteria to determine their dimensions. The cyclical component loading on GDP and GDI with loading normalized to one is what we call "output gap", and it is of particular relevance for the conduct of monetary policy, since it synthesize the current state of the US economy. The model consists of five different blocks: 1) a product/income block; 2) a price block; 3) a wage block; 4) a labor/employment block; 5) an interest rates block, containing interest rates at different maturities. Each block loads on a vector of stationary factors. The identification of cyclical factors is obtained by placing restriction on loading coefficients. We have a strong a priori that the number of cyclical factors is equal to one. This prior will be verified against the data. Each block also loads on non stationary factors that are modeled as random walks with stochastic drift affected by linear constraints. The model is estimated by using Bayesian techniques.

CO615 Room I2 ADVANCES IN FINANCIAL TIME SERIES AND ECONOMETRICS Chair: Helena Veiga

C0715: **Bayesian nonparametric Bernstein copula for uncorrelated dependent MGARCH errors**
Presenter: **Martina Danielova Zaharieva**, Erasmus University Rotterdam, Netherlands
Co-authors: Concepcion Ausin
The proposed model is a Copula-MGARCH, in which the dependency structure of the uncorrelated error term is modeled by a nonparametric copula. The idea of the dependent but uncorrelated MGARCH error term is extended by introducing a Bayesian Bernstein copula based on a Dirichlet process mixture. Hence, the time-varying dependence structure is modeled trough the dynamic linear correlation captured by the MGARCH model and the remaining (nonlinear) dependence trough the flexible nonparametric Bernstein copula, for which we introduce a stick-breaking representation model. We design a full Bayesian MCMC algorithm and provide a simulation study. Finally, we propose an application to portfolio optimization.

C0819: **Data cloning estimation for asymmetric stochastic volatility models**
Presenter: **J Miguel Marin**, University Carlos III, Spain
Co-authors: Patricia de Zea Bermudez, Helena Veiga
The focus is on applying data cloning to the estimation of asymmetric stochastic volatility models with flexible distributions that are able to capture the leptokurtosis and skewness of the distribution of standardized returns. Data cloning is a general technique to compute maximum likelihood estimates, along with their asymptotic variances, by means of the computation of the posterior distributions by using a MCMC methodology. Using an intensive simulation study and high frequency data for two financial time series of returns, the asymmetric stochastic volatility models are estimated and evaluated using the new proposal and a benchmark. Its performance is compared in terms of parameters' estimation, volatility and out-of-sample forecasts with a well-known Bayesian procedure. The results point out gains in efficiency and accuracy of the new estimators of parameters and volatility.

C1124: **Estimating threshold stochastic volatility models using integrated nested Laplace approximations**
Presenter: **Patricia de Zea Bermudez**, FCiencias.ID, Portugal
Co-authors: J Miguel Marin, Haavard Rue, Helena Veiga
Volatility of financial time series is well captured by stochastic volatility models. These models have been extended to be able to represent the asymmetric response of volatility to negative and positive shocks of the same magnitude. In the literature, there are two well-known extensions. One of these extensions is the threshold stochastic volatility (TSV) model which considers the asymmetric response of the volatility by allowing the parameters of the log-volatility equation to be different depending on the sign of lagged returns. The aim is to apply integrated nested Laplace approximations (INLA) in this framework. INLA is very fast and efficient. It replaces MCMC simulations with accurate deterministic approximations. Proper although not very informative priors are used, as well as PC priors. Finally, an application will be presented using a financial data set.

C0838: **Quantile consumption-capital asset pricing model**
Presenter: **Helena Veiga**, BRU-IUL (Instituto Universitario de Lisboa), Portugal
Co-authors: Abderrahim Taamouti, Sofia Ramos, Chih-Wei Wang
The Capital Asset Pricing Model (CAPM) is a statement about the conditional mean of asset returns and doesn't say anything about other levels

(quantiles) of the conditional distribution of asset returns. We examine whether the Euler equation defined by a consumption-based stochastic discount factor can also be expressed in terms of conditional quantile rather than the conditional mean. We show that replacing the standard expected utility optimization problem, which is expressed in terms of the conditional expectations, with a quantile utility optimization problem, using quantile utility function leads to an Euler equation in which the asset price is a function of the quantile of the stochastic discount factor and of payoff. Under the assumption that consumption growth rate process is log-elliptically distributed, the latter result implies that the quantiles of asset returns are functions of consumption volatility. The empirical evidence validates our theoretical results and show that consumption volatility is a driving factor of quantiles of stock market returns.

CO182 Room M2 ADVANCES IN SVARS Chair: Luca Fanelli

C0283: **Bootstrap inference in proxy SVARs**
Presenter: **Kurt Lunsford**, Federal Reserve Bank of Cleveland, United States
Proxy structural vector autoregressions (SVARs) identify structural shocks in vector autoregressions (VARs) with external proxy variables that are correlated with the structural shocks of interest but uncorrelated with other structural shocks. We provide asymptotic theory for proxy SVARs when the VAR innovations and proxy variables are jointly alpha-mixing. We also prove the asymptotic validity of a residual-based moving block bootstrap (MBB) for inference on statistics that depend jointly on estimators for the VAR coefficients and for covariances of the VAR innovations and proxy variables. These statistics include structural impulse response functions (IRFs). Conversely, wild bootstraps are invalid, even when innovations and proxy variables are either independent and identically distributed or martingale difference sequences, and simulations show that their coverage rates for IRFs can be badly mis-sized.

C0425: **Bayesian structural VAR models: An extended approach**
Presenter: **Michele Piffer**, Queen Mary, University of London, United Kingdom
Co-authors: Martin Bruns
An approach is provided to derive the posterior distribution of SVAR models directly on the structural parameters. Our framework allows for restrictions not only on the contemporaneous relations among variables, but also on impact impulse responses. In so doing, we enrich the tool available for the researchers who aim to set identify structural VAR models. Applying the methodology to simulated data on the New Keynesian model, we find that our approach recovers the true responses more tightly than the popular indirect orthogonal reduced form approach. We then apply our procedure to the identification of fiscal shocks.

C0446: **Proxy-SVAR as a bridge for identification with mixed frequency data**
Presenter: **Andrea Giovanni Gazzani**, Bank of Italy, Italy
Co-authors: Alejandro Vicondoa
High frequency identification around key events has recently solved many puzzles in empirical macroeconomics. A novel methodology, the Bridge Proxy-SVAR, is proposed to identify structural shocks in Vector Autoregressions (VARs) by exploiting high frequency information in a more general framework. Our methodology comprises three steps: (I) identify the structural shocks of interest in high frequency systems; (II) aggregate the series of high frequency shocks at a lower frequency; (III) use the aggregated series of shocks as a proxy for the corresponding structural shock in lower frequency VARs. Both analytically and through simulations, we show that our methodology significantly improves the identification of VARs. In an empirical application on US data, a properly identified monetary policy news shock leads to a fall in output and prices.

C1084: **Heteroskedastic proxy-SVARs**
Presenter: **Luca Fanelli**, University of Bologna, Italy
Identification strategies are discussed for Structural Vector Autoregressions (SVARs) which combine the use of external instruments, the so-called proxy-SVAR or SVAR-IV approach, with the heteroskedasticity found in the data, the so-called identification-via-heteroskedasticity. The focus is on the case in which r valid instruments are used to identify g (larger of equal than one) structural shocks of interest, with r larger or equal than g, and there are m structural breaks in the VAR error covariance matrix which give rise to $m+1$ volatility regimes. It is shown that the combination of the two approaches enhances the identification possibilities for practitioners and produces overidentified, testable models, denoted HP-SVARs. Two types of heteroskedasticity are considered. In one case, the structural breaks do not affect the on-impact coefficients hence the impulse response functions (IRFs) are constant across volatility regimes. In the other case, the structural breaks affect the on-impact coefficients and IRFs are regime-dependent. General identification results for HP-SVARs are derived for these two cases. Estimation can be carried out through maximum likelihood.

CO134 Room P2 ECOSTA JOURNAL PART A: ECONOMETRICS II Chair: Tommaso Proietti

C1514: **Accurate subsampling intervals of principal components factors**
Presenter: **Esther Ruiz**, Universidad Carlos III de Madrid, Spain
Co-authors: Javier de Vicente
In the context of Dynamic Factor Models (DFMs), one of the most popular procedures for factor extraction is Principal Components (PC). Measuring the uncertainty associated to PC factor estimates should be part of interpreting them. However, the asymptotic distribution of PC factors could not be an appropriate approximation to the finite sample distribution for the sample sizes and cross-sectional dimensions usually encountered in practice. The main problem is that parameter uncertainty is not taken into account. We show that several bootstrap procedures proposed in the context of DFM with goals related to inference are not appropriate to measure the uncertainty of PC factor estimates. We propose an asymptotically valid subsampling procedure designed with this purpose. The finite sample properties of the proposed procedure are analyzed and compared with those of the asymptotic and alternative extant bootstrap procedures. The results are empirically illustrated obtaining confidence intervals of the underlying factor in a system of Spanish macroeconomic variables.

C1513: **Estimating treatment effects in regression discontinuity designs with multiple assignment variables**
Presenter: **Chung-Ming Kuan**, National Taiwan University, Taiwan
Co-authors: Yu-Chin Hsu
While treatment assignment is often determined by one threshold value, many empirical studies have shown the pervasiveness of regression discontinuity (RD) designs with more than one assignment variable. Moreover, the literature has focused on the average treatment effect and overlooked the interesting perspectives provided by treatment effects at different quantiles of the outcome distribution. We propose new approaches for RD designs with multiple assignment variables. The approaches allow nonparametric estimation and could be applied to estimating average treatment effects and quantile treatment effects. Based on our Monte Carlo simulation study, we suggest that the performance of the existing approaches is sensitive to the interaction terms in data generating processes as well as large variations in assignment variables. Our new approaches produce robust and more accurate estimates compared to the existing approaches with respect to all scenarios.

C1723: **Smooth FREE COMFORT: Mixing IID and GARCH-based frameworks for optimal portfolio performance**
Presenter: **Marc Paolella**, University of Zurich, Switzerland
Co-authors: Jeffrey Naef, Pawel Polak
The fundamental goals associated with active portfolio management include the choice and weights for assets that result in high medium- and long-term performance, beating benchmarks in terms of risk and return. GARCH-based multivariate models for asset returns can outperform their IID counterparts, but often only when transaction costs are not accounted for, because of the high turnover induced by the ever-changing dispersion matrix. A setup is considered based on a smooth transition between IID and GARCH-based models, and, based on a large data set of stock returns, demonstrates that the optimum lies between the two extremes. Further enhancements are introduced via use of a LASSO-based shrinkage paradigm.

C1626: **On the frequency of transmission of market volatility to individual stocks: A double asymmetric GARCH–MIDAS approach**
Presenter: **Giampiero Gallo**, NYU in Florence, Italy
Co-authors: Alessandra Amendola, Vincenzo Candila
Volatility in financial markets has both low and high–frequency components which determine its dynamic evolution. Previous modelling efforts in the GARCH context (e.g. the spline–GARCH) were aimed at estimating the low-frequency component as a smooth function of time around which short-term dynamics evolve. Alternatively, recent literature has introduced the possibility of considering data sampled at different frequencies to estimate the influence of macrovariables on volatility. We use a recently developed model, labelled double asymmetric GARCH MIDAS model, where variations in a market volatility variable (in our context, VIX) are observed both at the daily and the monthly level and represent different channels through which market volatility can influence individual stocks. We want to convey the idea that such variations (separately) affect the short– and long–run components, possibly having a separate impact according to their sign. The model is estimated on a panel of stocks chosen by their sizable capitalization across different sectors.

CC647 Room A2 CONTRIBUTIONS IN FORECASTING I Chair: Valderio Anselmo Reisen

C0727: **Robust optimization of forecast combinations**
Presenter: **Stelios Arvanitis**, RC-AUEB, Greece
Co-authors: Thierry Post, Selcuk Karabati
A methodology is developed for constructing robust forecast combinations which improve upon a given benchmark specification for all symmetric and convex loss functions. The optimal forecast combination asymptotically almost surely dominates the benchmark and, in addition, minimizes the expected loss function, under standard regularity conditions. The optimum in a given sample can be found by solving a large convex optimization problem. An application to forecasting of changes of the S&P 500 volatility index shows that robust optimized combinations improve significantly upon the out-of-sample forecasting accuracy of simple averaging and unrestricted optimization.

C1443: **Forecasting realized volatility: Implied volatility, leverage effects & the volatility of realized volatility**
Presenter: **Katerina Tsakou**, Swansea University, United Kingdom
The aim is to assess the forecasting performance of time series models for realized volatility, which take into consideration implied volatility, leverage effects, as well as the volatility of realized volatility. Realized volatility is modeled and forecasted with HAR models for a number of US and European indices. We find that accounting for these stylized facts of volatility leads to a significant improvement of the models' predictive performance. The results suggest that a HAR model which accommodates implied volatility and leverage effects produces the most accurate volatility forecast.

C1606: **Similarity forests for time series classification**
Presenter: **Laura Calzada**, University of Oviedo, Spain
Co-authors: Maria Oskarsdottir, Bart Baesens
The telecommunication industry is a saturated market where a proper implementation of a retention campaign is critical to be competitive due to the fact that retaining a customer is cheaper than achieving a new one. Some research has used binary classification methods to predict churn of customers. Moreover, it has been shown that a customer's social relationships have an influence on the decision of changing of the operator. So, it is crucial to detect customer behavioral patterns and define accurate models to predict potential churners. We propose a novel method to extract the dynamic influence of each customer using social network analysis techniques, predicting accurately both in short and long-term with binary classification methods. Call detail records of telcos customers are used to build a temporal network to extract the churn behavioral patterns. The dynamic influence of each customer is determined by applying centrality metrics and diffusion propagation methods over a sliding window. The time series are classified by a recently proposed binary classification method called similarity forests. In addition, a comparison to other methods like logistic regression and random forests evaluates the accuracy of predicting further in time and the possibility of designing a method that is capable of detecting potential churners in short and long term.

C1436: **Optimal pooling and finite mixture distribution combinations of probabilistic forecasts**
Presenter: **Giulia Mantoan**, University of Warwick, United Kingdom
The combination of two or more density forecasts has a long tradition in the statistics and forecasting literature. However, comparatively little attention in econometrics has been given to the finite mixture distribution as a statistical model for combining density forecasts. Combination procedures based on a mixture density distribution are able to account for parameter uncertainty in addition to combination's weight uncertainty, which are features normally not considered in traditional "two-step" approaches to density forecasts combination. The aim is to compare the "one-step" mixture approach to the more traditional "two-step" approach thereby endowing the decision maker with a tool to elicit the best aggregation method. The superiority of the "one-step" approach is accessed through analytical analyses, several Monte Carlo simulations to account for structural breaks and a macroeconomic application.

CG193 Room N2 CONTRIBUTIONS IN VOLATILITY AND RISK Chair: Giuseppe Storti

C1620: **A DCC-type approach for realized covariance modelling with score-driven dynamics**
Presenter: **Danilo Vassallo**, Scuola Normale Superiore of Pisa, Italy
Co-authors: Giuseppe Buccheri, Fulvio Corsi
A class of dynamic models for realized covariances is introduced where volatilities and correlations are separately estimated. We can thus combine univariate realized volatility models with a recently introduced class of score-driven realized covariance models based on Wishart and F-matrix distributions. The proposed models are computationally simple to estimate in high dimensions and allow complete flexibility in the choice of the univariate specification. Through a Monte-Carlo study, we show that the two-step maximum likelihood procedure provides accurate parameter estimates in small samples. Empirically, we find that the proposed models outperform existing benchmarks, with forecasting gains that increase with the dimension.

C1718: **On hedge funds: New evidence from volatility risk premia embedded in VIX options**
Presenter: **Anmar Al Wakil**, Paris-Dauphine University, France
Co-authors: Serge Darolles
Volatility risk in hedge funds is deciphered from option-based dynamic trading strategies. It is demonstrated that volatility risk premia strategies, as measured by pricing discrepancies between real-world and risk-neutral probability distributions of the volatility of the S&P 500 Index returns embedded in VIX options, are instrumental determinants in hedge fund performance, in both time-series and cross-section. After controlling for Fung-Hsieh factors, a positive one-standard deviation shock to volatility of volatility risk premium is associated with a substantial decline in aggregate hedge fund returns of 25.2% annually. The results particularly evidence hedge funds that significantly load on volatility of volatility (kurtosis of volatility) risk premium subsequently outperform low-beta funds by nearly 11.7% (8.6%) per year. This finding suggests to what extent hedge fund alpha arises actually from selling volatility risk.

C1452: **The shale revolution, geopolitical risk, and oil price volatility**
Presenter: **Fuyu Yang**, University of East Anglia, United Kingdom
Co-authors: Wenxue Wang
The U.S. shale revolution, using new technologies to extract crude oil, has led to new dynamics in the supply side of the global oil market. We ask whether the shale revolution has dampened the role of geopolitical risk in oil price volatility. We extend a reduced form Structural Break Threshold Vector Autoregressive (SBT-VAR) model to a structural SBT-VAR model and identify the structural innovations by allowing for conditional heteroskedasticity. Compared with the conventional reduced form VAR and TVAR models, a SBT-VAR with a constant threshold and a break in April 2014 are supported by the data. We then analyse the conditional (co)variance impulse response with respect to two distinct shock scenarios, one with only a geopolitical risk shock, the other with a simultaneous shale production shock and a geopolitical risk shock. The volatility responses are due to the identified contemporaneous relationships amongst geopolitical risk, shale production and oil prices, and are conditional on volatilities at the points in time. With the extra unit shale production shock, we find that the volatility response of oil prices to a geopolitical risk shock is higher, but the response is less correlated with the geopolitical risk factor.

C1159: **A self-exciting hurdle model for extreme returns in financial markets**
Presenter: **Katarzyna Bien-Barkowska**, Warsaw School of Economics, Poland
Forecasting the occurrence of extreme returns is at the forefront of modern financial econometrics and allows for effective management of financial risk. We propose an application of the discrete variable hurdle model for the occurrence of extreme losses in financial markets which allows us to explain apparent bursts of volatility and clustering of extreme returns. The conditional probability that the loss exceeds a large threshold on a given day is modeled dynamically exhibiting the self-exciting nature, where recently observed days with extreme returns increase the likelihood of experiencing further large losses. We show that the model outperforms the standard peak-over-threshold methods for forecasting value at risk and expected shortfall.

Saturday 15.12.2018 18:10 - 19:25 Parallel Session J – CFE-CMStatistics

EO264 Room A0 APPROACHES TO ANALYZING HIGH DIMENSIONAL DATA Chair: Hernando Ombao

E1245: **A frequency domain approach to stationary subspace analysis of multivariate second-order nonstationary time series**
Presenter: **Raanju Sundararajan**, KAUST, Saudi Arabia

Transforming high dimensional multivariate nonstationary time series into a lower dimensional stationary time series is of great importance in application areas such as neuroscience and economics. Brain signals like electroencephalograms (EEGs) often appear as nonstationary time series and removing the nonstationarity from the observed signal is useful in building classification models for brain-computer interface. Stationary subspace analysis (SSA) finds instantaneous stationary linear transformations of nonstationary processes. We describe an SSA procedure for multivariate second-order nonstationary processes. The key idea is the property of asymptotic uncorrelatedness of the discrete Fourier transform of a second-order stationary time series. A measure of departure from stationarity that captures the sizes of the entries of the DFT covariance matrices is minimized to obtain the transformation matrix. The dimension of the subspace is estimated using a sequential procedure and its asymptotic properties are provided. The non-uniqueness issues in subspace estimation are discussed and a technique to select a subspace from a set of subspaces using canonical angles is provided. We study the performance of the method in detecting the dimension of the true subspace through simulation examples. Finally, we present an application of SSA in constructing a classification model that differentiates healthy subjects from subjects with some neurological disorder.

E1251: **Detecting regime changes in community structure of brain networks using dynamic stochastic block models**
Presenter: **Chee Ming Ting**, King Abdullah University of Science and Technology, Saudi Arabia
Co-authors: Siti Balqis Samdin, Hernando Ombao

Brain networks exhibit the property of modular community structure with highly inter-connected nodes within a same module, but sparsely connected between different modules. Recent neuroimaging studies also suggest dynamic changes in brain connectivity over time. We present a statistical approach based on dynamic stochastic block models (SBM) to characterize changes in community structure of the brain functional networks inferred from neuroimaging time series data. The dynamic SBM is a non-stationary extension combining a static SBM with a Markov process to allow for temporal evolution of the community membership of nodes and the network connectivity. The model is formulated into a state-space form with sequential estimation of the time-varying parameters by Kalman filtering. We further partition the time-evolving community structure into recurring, piece-wise constant regimes or states using an infinite hidden Markov model that can learn an unknown number of states from the data. The method is applied to resting-state and task-based functional magnetic resonance imaging (fMRI) data to detect dynamic reconfiguration of network structure of the brain.

E0428: **Registration for exponential family functional data**
Presenter: **Jeff Goldsmith**, Columbia University, United States
Co-authors: Julia Wrobel, Vadim Zipunnikov, Jennifer Schrack

A novel method is introduced for separating amplitude and phase variability in exponential family functional data. Our method alternates between two steps: the first uses generalized functional principal components analysis (GFPCA) to calculate template functions, and the second estimates smooth warping functions that map observed curves to templates. Existing approaches to registration have primarily focused on continuous functional observations, and the few approaches for discrete functional data require a pre-smoothing step; these methods are frequently computationally intensive. In contrast, we focus on the likelihood of the observed data and avoid the need for preprocessing, and we implement both steps of our algorithm in a computationally efficient way. Our motivation comes from the Baltimore Longitudinal Study on Aging, in which accelerometer data provides valuable insights into the timing of sedentary behavior. We analyze binary functional data with observations each minute over 24 hours for 592 participants, where values represent activity and inactivity. Diurnal patterns of activity are obscured due to misalignment in the original data but are clear after curves are aligned. Simulations designed to mimic the application outperform competing approaches in terms of estimation accuracy and computational efficiency. Code for our method and simulations is publicly available.

EO448 Room Aula 4 NEW DEVELOPMENTS IN STATISTICAL INFERENCE AND COMPUTING Chair: Min-ge Xie

E0622: **Causal inference with measurement error in outcomes**
Presenter: **Grace Yi**, University of Waterloo, Canada

Inverse probability weighting (IPW) estimation has been popularly used to consistently estimate the average treatment effect (ATE). Its validity, however, is challenged by the presence of error-prone variables. In application, measurement error is ubiquitously present in data collection due to various reasons. Naively ignoring measurement error effects usually yields biased inference results. We will discuss the IPW estimation with mismeasured outcome variables. The impact of measurement error for both continuous and discrete outcome variables will be examined. We will describe estimation procedures with the outcome misclassification effects accommodated. Consistency and efficiency will be investigated. Numerical studies will be reported to assess the performance of the proposed methods.

E0734: **Determine the number of states in hidden Markov models via marginal likelihood**
Presenter: **Yang Chen**, University of Michigan, United States
Co-authors: Chu-Lan Kao, Cheng-Der Fuh, Samuel Kou

Hidden Markov models (HMM) have been widely adopted by scientists from various fields to model stochastic systems: the underlying process is a discrete Markov chain and the observations are noisy realizations of the underlying process. Determining the number of hidden states for an HMM is a model selection problem, which has yet to be satisfactorily solved, especially for the popular Gaussian HMM with heterogeneous covariance. We propose a consistent method for determining the number of hidden states of HMM based on the marginal likelihood, which is obtained by integrating out both the parameters and hidden states. Moreover, we show that the model selection problem of HMM includes the order selection problem of finite mixture models as a special case. We give a rigorous proof of the consistency of the proposed marginal likelihood method, which is based on the notion of asymptotic "path-ignorance", and provide simulation studies to compare the proposed method with the currently mostly adopted method, the Bayesian information criterion (BIC), demonstrating the effectiveness of the proposed marginal likelihood method.

E1307: **Bayesian analysis of the co-variance matrix of a multivariate normal distribution with a new class of priors**
Presenter: **Dongchu Sun**, University of Missouri, United States

Bayesian analysis for the covariance matrix of a multivariate normal distribution has received a lot of attention in the last two decades. We propose a new class of priors for the covariance matrix, including both inverse Wishart and reference priors as special cases. The main motivation for the new class is to have available priors – both subjective and objective – that do not "force eigenvalues apart," which is a criticism of inverse Wishart and Jeffreys priors. Extensive comparison of these 'shrinkage priors' with inverse Wishart and Jeffreys priors is undertaken, with the new priors seeming to have considerably better performance. A number of curious facts about the new priors are also observed, such as that the posterior distribution will be proper with just three vector observations from the multivariate normal distribution – regardless of the dimension

of the covariance matrix – and that useful inference about features of the covariance matrix can be possible. Finally, a new MCMC algorithm is developed for this class of priors and is shown to be computationally effective for matrices of up to 100 dimensions.

EO114 Room Aula 5 STATISTICAL METHODS FOR COMPLEX DATA ANALYSIS Chair: Zhuoqiong He

E1366: **Bayesian smoothing spline model and its application in current population survey**
Presenter: **Zhuoqiong He**, University of Missouri, United States
The Current Population Survey (CPS) is conducted to collect the labor force data and to measure the extent of unemployment in the United States of America. The total numbers of employment population and unemployment population are estimated from the CPS sample, and seasonal adjustment of the estimates is needed to observe the changing of economic conditions. We propose a Bayesian smoothing spline (BSS) model to remove the seasonal fluctuations and to capture the fundamental tendency of a labor force total associated with general economic expansions and contractions. This BSS model can be efficiently computed with Markov chain Monte Carlo. The estimation of unemployment based on BSS is illustrated and compared with the seasonally adjusted unemployment estimation published by U.S. bureau of labor statistics.

E1368: **Design and analysis of pragmatic stepped-wedge clustered randomization trials**
Presenter: **Song Zhang**, University of Texas Southwestern Medical Center, United States
The stepped-wedge cluster randomized trial design is particularly suitable and has been frequently adopted by pragmatic trials. Under this design, initially all clusters receive control treatment. Subsequently, at pre-defined time points (steps) clusters are randomized to switch to intervention. Outcomes are measured at every step. All clusters receive the intervention at the end of study. We investigate the design and analysis of data arising from stepped-wedge trials under pragmatic situations such as missing data, uneven steps, various types of correlation structures, random cluster sizes, etc. Simulation studies and application examples to real clinical trials are presented.

E1374: **A Bayesian spatial clustering method and its application in radiology**
Presenter: **Jing Cao**, Southern Methodist University, United States
Co-authors: Song Zhang
Kidney cancer is among the ten most common cancers in human. The dynamic contrast-enhanced MRI (DCE-MRI) takes advantage of the interaction between a contrast agent and adjacent water protons which generates brighter signals in the scan image. A novel Bayesian spatial clustering method based on a mixture of multivariate normal distribution is proposed. A latent conditional regression (CAR) process is employed to account for the spatial correlation among clustering indexes. The proposed method is demonstrated to provide smoother and more accurate clustering of pixels. A simulation study and a real application example are presented.

EO597 Room Aula B ESTIMATION AND OPTIMIZATION IN LARGE-SCALE STATISTICAL SETTINGS Chair: Garvesh Raskutti

E0350: **On Stein's identity and derivate- and Hessian-free stochastic optimization**
Presenter: **Krishnakumar Balasubramanian**, University of California, Davis, United States
Gaussian smoothing based techniques for zeroth-order stochastic optimization are common in the optimization literature. It will be shown that such techniques are essentially instantiations of Stein's identity, popular in the statistics literature. Based on this relationship, the following three results will be discussed. First, under a structural sparsity assumption on the optimization problem, we will illustrate an implicit regularization phenomenon where a derivative-free stochastic gradient algorithm adapts to the sparsity of the problem at hand by just varying the step-size. Next, we will discuss a truncated derivative-free stochastic gradient algorithm, whose rate of convergence depends only poly-logarithmically on the dimensionality under the sparsity assumption. Finally, leveraging the second-order Stein's identity, we will introduce a Hessian-free Newton method with zeroth-order information and discuss its convergence rates.

E0566: **Statistical filtering for optimization under uncertainty**
Presenter: **Vivak Patel**, University of Wisconsin – Madison, United States
Across many disciplines, inference or decision tasks are formulated as optimizing objective functions involving expectations. While there are several paradigms for addressing such problems, such as Bayesian optimization or stochastic gradient methods, these paradigms are either computationally impractical or are too underdeveloped for real applications. Thus, there is still a need for practical optimization methods for these optimization problems. We introduce a novel paradigm that addresses this concern. Our paradigm leverages statistical filters to generate computationally practical subproblems, which can then be solved by mature, deterministic optimization methods. The resulting algorithms perform surprisingly well against state-of-the-art approaches, which we demonstrate on a handful of problems from a number of application areas.

E0724: **Minimax estimation of bandable precision matrices**
Presenter: **Sahand Negahban**, Yale University, United States
Co-authors: Addison Hu
The inverse covariance matrix provides considerable insight for understanding statistical models in the multivariate setting. In particular, when the distribution over variables is assumed to be multivariate normal, the sparsity pattern in the inverse covariance matrix, commonly referred to as the precision matrix, corresponds to the adjacency matrix representation of the Gauss-Markov graph, which encodes conditional independence statements between variables. Minimax results under the spectral norm have previously been established for covariance matrices, both sparse and banded, and for sparse precision matrices. We establish minimax estimation bounds for estimating banded precision matrices under the spectral norm. Our results greatly improve upon the existing bounds; in particular, we find that the minimax rate for estimating banded precision matrices matches that of estimating banded covariance matrices. The key insight in our analysis is that we are able to obtain barely-noisy estimates of $k \times k$ subblocks of the precision matrix by inverting slightly wider blocks of the empirical covariance matrix along the diagonal. Our theoretical results are complemented by experiments demonstrating the sharpness of our bounds.

EO494 Room Aula Magna MODEL SELECTION AND FDR Chair: Sylvain Sardy

E0238: **Beyond FDR: Towards simultaneous selective inference and post-hoc error control**
Presenter: **Aaditya Ramdas**, Carnegie Mellon University, United States
Co-authors: Eugene Katsevich
The false discovery rate (FDR) is a popular error criterion for multiple testing, but it is not without its flaws. Indeed, (a) controlling the mean of the false discovery proportion (FDP) does not preclude large FDP variability, and (b) committing to an error level before observing the data limits its use in exploratory data analysis. We take a step towards addressing both drawbacks by proving uniform FDP bounds for a variety of existing FDR procedures. We open up a middle ground between fully simultaneous inference (guarantees for all possible rejection sets), and fully selective inference (guarantees only for a single rejected set). They allow the scientist to "spot" one or more suitable rejection sets (select post-hoc on the algorithm's trajectory) by picking data-dependent sizes or error-levels, after examining the entire path of estimated FDPs and the uniform upper band on the true FDP. This post-hoc mode of inference addresses both aforementioned drawbacks of FDR. Our bounds apply to online

FDR procedures as well, providing simultaneous high-probability uniform FDP bounds at arbitrary data-dependent query times for arbitrary online procedures. Finally, our analysis unifies existing martingale and empirical process viewpoints on FDR algorithms.

E1496: **Breaking the lasso power-FDR tradeoff diagram by thresholding**
Presenter: **Asaf Weinstein**, Stanford University, United States
Co-authors: Weijie Su, Malgorzata Bogdan, Emmanuel Candes
The lasso is often used by practitioners as a variable selector in large regression problems. Most commonly, the penalty parameter is chosen by cross validation, even though it is well known that this method tends to yield too many false discoveries. Furthermore, recent work has shown that if λ is set so that the false discovery rate is controlled, there is still an inherent cost in power (identification of true nonnulls), adding to the criticism of using the lasso for support estimation. It is also well known among practitioners that this phenomenon can be mitigated by thresholding the lasso estimate (at a value larger than zero). Working with IID Gaussian covariates, we analyze and precisely quantify the advantages that such a procedure can have in terms of the tradeoff between the false discovery proportion and the true positive proportion. Importantly, the penalty parameter λ now plays a crucial role in the ordering of the variables, and, interestingly, we explain why cross-validation is the right way to choose it (at least in the IID Gaussian covariates case).

E1520: **Model selection with lasso-zero tuned by quantile universal thresholding**
Presenter: **Pascaline Descloux**, University of Geneva, Switzerland
Co-authors: Sylvain Sardy
When performing variable selection, controlling the false discovery rate (FDR) while maintaining high power is challenging. A new ℓ_1-based estimator called lasso-zero is introduced for the linear regression problem. It relies on the repeated use of noise dictionaries concatenated to the design matrix for fitting the noise component. The threshold level is tuned by quantile universal thresholding, a general methodology that was introduced to select the regularization parameter of any thresholding estimator. The FDR is provably controlled for orthogonal designs, and empirically for independent Gaussian predictors. In case of correlated variables, simulations show that even though the FDR is no longer controlled, the proposed methodology exhibits a very good tradeoff between low FDR and high true positive rate.

EO556 Room Aula C ROBUST MACHINE LEARNING Chair: Guillaume Lecue

E0541: **Dimension-free PAC-Bayesian bounds for vectors and matrices**
Presenter: **Ilaria Giulini**, Universite Paris Diderot, France
PAC-Bayesian inequalities are used to present new robust estimators for the mean of a random vector and of a random matrix. More precisely, we establish dimension-free bounds and we work under mild polynomial moment assumptions regarding the tail of the sample distribution. Particular attention is devoted to the estimation of the Gram matrix, due to its prominent role in high-dimensional analysis.

E0953: **Median-of-means-type estimators and rates of convergence in normal approximation**
Presenter: **Stanislav Minsker**, University of Southern California, United States
New results are presented for the class of estimators obtained via the generalized median-of-means (MOM) technique. These results stem from connections between performance of MOM-type estimators and the rates of convergence in normal approximation. We provide tight non-asymptotic deviations guarantees in the form of exponential concentration inequalities, as well as asymptotic results in the form of limit theorems. Our techniques will be illustrated with several examples, including the now-classical median-of-means estimator, and robust maximum likelihood estimation.

E1095: **On stochastic approximation under heavier tails**
Presenter: **Philip Thompson**, CREST-ENSAE, France
The solution of regularized stochastic convex optimization problems is considered via the stochastic approximation methodology, i.e., by means of a first order stochastic oracle sampled from the population distribution in an online fashion. We show that by choosing a specific policy for the stepsize sequence and the mini-batch size per iteration (i.e. the sample size used to compute the empirical mean of the gradient), it is possible to obtain (near) optimal iteration and sample complexities under very mild assumptions on the stochastic oracle distribution. For instance, our non-asymptotic rates are valid for oracles with pointwise finite variance and with "multiplicative noise" (the standard deviation of the first order oracle is Lipschitz continuous). This includes, e.g., linear regression problems where the design matrix may have arbitrary unbounded corruptions. Moreover, the proposed iterative estimator possesses a "variance localization" property: the bounds depends only on the variance at solutions. We also show rates of convergence in the setting where the Lipschitz constant is unknown and propose a stochastic approximated line search which adapts the estimator to this lack of information. In this case, some iterative arguments based on empirical process and self-normalization theory are used.

EO530 Room C1 MIXED LINEAR MODELS ANALYSIS: NEW ESTIMATION METHODS AND DIAGNOSTIC TOOLS Chair: Dietrich von Rosen

E1032: **Growth curve model with bilinear random coefficient**
Presenter: **Shinpei Imori**, Hiroshima University, Japan
Co-authors: Dietrich von Rosen, Ryoya Oda
The growth curve model is a classical model useful to analyze repeated measurements data, where response variables are obtained in matrix form. Each row of the response matrix can represent observations on the same time point and each column of the response matrix is assumed to be independently distributed is a conventional framework of the model. However, if each column of the response matrix represents observations on the same space point, this assumption may not be appropriate. However, if we consider an unstructured covariance matrix for the response variables, the number of unknown parameters is greater than the sample size. We solve this problem by introducing a bilinear random coefficient to the (extended) growth curve model, which induces a Kronecker product covariance structure of the response matrix in the growth curve model. An explicit maximum likelihood estimator of the unknown parameters is presented, even when the covariance matrix of the random coefficient is non-negative definite.

E1249: **On prediction in multivariate mixed linear models with structured covariance matrices**
Presenter: **Tatjana von Rosen**, Stockholm University, Sweden
The mixed linear models have become a widely used tool for the analysis of data having complicated structures and exhibiting various dependence patterns, e.g. repeated measures or longitudinal data containing multiple sources of variation. Prediction problems in univariate mixed linear models have got considerable attention due to numerous applications in small area estimation, educational research and animal breeding, among others. Despite complexity of real-life phenomena, the multivariate mixed linear models have received little attention. The focus is on the prediction of linear combinations involving both fixed and random effects in balanced multivariate mixed liner models which can handle both the multivariate response and spatial or/and temporal dependence. More specifically, the equality of linear predictors under two multivariate mixed effects models with different covariance matrices is of interest. In practice, it can be difficult to decide about an appropriate covariance structure of random effect,

so using equivalent linear models can resolve that problem and possibly reduce computations. The task is rather complicated if one aim is to get explicit results, hence we shall focus on a certain class of covariance matrices whose structure is preserved under matrix inversion.

E1306: **Parameter estimation in biclassified blockmodels as mixture of contingency tables via the EM algorithm**
Presenter: **Marianna Bolla**, Institute of Mathematics, Technical University of Budapest, Hungary
Co-authors: Fatma Abdelkhalek, Jozsef Mala
A random contingency table model is introduced, where the entries are independent beta-distributed with parameters depending on their row and column labels. Sufficient statistics are specified, and based on them, an algorithm is given to find the MLE of the parameters, together with convergence proof. The model is extended to the multiclass scenario, where for fixed number of biclusters, the parameters of the beta-distributed entries also depend on their row and column cluster memberships. To find the clusters and estimate the parameters, an EM iteration for mixtures of exponential-family distributions is used. The algorithm is applicable to microarrays, and a genetic example is presented.

EO108 Room D1 TINKERING WITH GINI: ADAPTATIONS OF THE OLD IDEA TO PRESENT-DAY REALITIES Chair: Francesca Greselin

E0190: **Gini shortfall: A new coherent risk measure**
Presenter: **Ricardas Zitikis**, University of Western Ontario, Canada
For quite some time, Value-at-Risk (VaR) was an appealing risk measure, and even the industry and regulatory standard for calculating risk capital in banking and insurance. VaR is still a standard, but its allure for applications has been criticized in many theoretical and empirical works. Expected Shortfall (ES) has been a much welcome breakthrough, in that it always rewards diversification and captures the magnitude of the tail risk. But what about tail variability? The new coherent risk measure, called Gini Shortfall (GS), provides a most welcome missing-piece in the encompassing risk-measurement puzzle. We will introduce and discuss the GS.

E0525: **Cramer type large and moderate deviations for trimmed L-statistics**
Presenter: **Nadezhda Gribkova**, Saint-Petersburg State University, Russia
The class of L-statistics is one of the most commonly used classes in statistical inferences; the famous Gini index also belongs to this class. There is an extensive literature on asymptotic properties of L-statistics, but its part related to large deviations is not so vast. Only a few highly sharp results on large deviations for non-trimmed L-statistics with coefficients generated by a smooth on $(0,1)$ weight function are mentioned. These results, however, do not cover the case of trimmed L-statistics, i.e., the case when the weight function is zero outside of some interval $[\alpha, 1-\beta] \subset (0,1)$. Our recent results on Cramer type large and moderate deviations for trimmed L-statistics will be presented, and our approach for solving this problem will be discussed. This approach is to approximate the trimmed L-statistic by a non-trimmed L-statistic with coefficients generated by a smooth on $(0,1)$ weight function, where the approximating L-statistic is based on order statistics corresponding to a sample of i.i.d. Winsorized random variables.

EO042 Room E1 NONPARAMETRIC METHODS FOR MODERN NETWORK ANALYSIS Chair: Yahui Tian

E0680: **Fusing data depth with complex networks: Community detection with prior information**
Presenter: **Yahui Tian**, Boehringer Ingelheim Investment Co., Ltd., China
Co-authors: Yulia Gel
A new nonparametric supervised algorithm is proposed for detecting multiple communities in complex networks. The key idea behind the new clustering method is the notion of robust and data-driven data depth methodology that still remains new and unexplored in network sciences. The proposed new DDG - method is inherently geometric and allows to simultaneously account for network communities and outliers. We illustrate utility of the new approach using the benchmark political blogs data.

E1268: **Leveraging the power of correlation in a data network: A machine learning approach**
Presenter: **Annalisa Appice**, University of Bari Aldo Moro, Dipartimento di Informatica, Italy
Co-authors: Donato Malerba
Predictive modelling of a data network is made complex due to the presence of correlation. Recent studies have shown that taking label correlations into account may contribute to improving the accuracy of predictive inferences in data network domains. The trend cluster is a space-time pattern defined in machine learning, in order to model the node correlation and the temporal dependence of a data network. Specifically, it describes any cluster of linked nodes which collect measures of a numeric field whose temporal variation, called trend polyline, is similar along a time horizon. Trend cluster discovery is, initially, investigated as an effective means to summarize a geophysical data network. Subsequently, it is combined with Inverse distance weighting and least-square regression, in order to derive a predictive model for the ubiquity interpolation of unobserved data. Finally, it is also investigated as a means to enrich a data network with forecasting ability. In particular, the forecasting ability is used to identify outliers, while the correlation of outliers is analysed, in order to classify changes and reduce the number of false anomalies.

E1687: **Nonparametric methods for change point analysis in multivariate, functional and network data**
Presenter: **Shojaeddin Chenouri**, University of Waterloo, Canada
A nonparametric framework is introduced for change point analysis in variety of data settings such as multivariate, matrix valued, functional and network data. To motivate the methodology, we begin with multivariate case in which we propose a nonparametric change point test for multivariate data using rankings obtained from data depth measures. As the data depth of an observation measures its centrality relative to the sample, changes in data depth may signify a change of scale of the underlying distribution, and the proposed test is particularly responsive to detecting such changes. We provide a full asymptotic theory for the proposed test statistic under the null hypothesis that the observations are stable, and natural conditions under which the test is consistent. The finite sample properties are investigated by means of a Monte Carlo simulation, and these along with the theoretical results confirm that the test is robust to heavy tails, skewness, and high dimensionality. Finally, we extend the methodology to more general data structures by introducing appropriate depth measures.

EO492 Room G1 RECENT DEVELOPMENT IN SEMIPARAMETRIC METHODS FOR SURVIVAL DATA Chair: Liming Xiang

E0770: **Frailty mean residual life regression for survival data from multi-center clinical trials**
Presenter: **Rui Huang**, Nanyang Technological University, Singapore
Co-authors: Liming Xiang, Il Do Ha

A frailty model framework based on mean residual life regression is proposed for analysis of clustered survival data collected from multi-center clinical trials. Our motivation is prompted by facts that 1) most current frailty models used in analysis of such data are either proportional hazards or additive hazards based, and 2) the mean residual life regression offers easily understood and straightforward interpretation for the effects of prognostic factors on the expectation of the remaining lifetime. To overcome estimation challenges, a novel hierarchical quasi-likelihood approach is developed by making use of the idea of hierarchical likelihood in the construction of the quasi-likelihood function, leading to hierarchical estimating equations. Simulation results show favorable performance of the method regardless of frailty distributions. The utility of the proposed methodology is illustrated by its application to the data from a multi-institutional study of breast cancer.

E0861: **Comparison between the marginal hazard models and sub-distribution hazard models with an assumed copula**
Presenter: **Takeshi Emura**, National Central University, Taiwan
Co-authors: Jia-Han Shih, Il Do Ha

For analysis of competing risks data, three different types of hazard functions have been considered in the literature, namely the cause-specific hazard, the sub-distribution hazard, and the marginal hazard function. Accordingly, different types of the Cox model have been proposed to estimate the effect of covariates on each of the three different hazard functions. Many authors studied the difference between the cause-specific hazard and the sub-distribution hazard. However, the study on the marginal hazard function is limited partly due to its model identifiability issue. We apply the assumed copula approach to deal with the model identifiability issue, and compare between the sub-distribution hazard and the marginal hazard function. We establish the mathematical relationship between the two hazard functions by using an assumed copula. We then extend our results to clustered semi-competing risks data. We implement the computing algorithm for marginal Cox regression with clustered competing risks data in the R joint.Cox package. We analyze four datasets for illustration.

E1001: **Generalized partially linear single-index cure mixture models with interval-censored data**
Presenter: **Xiaoyu Liu**, Nanyang Technological University, Singapore
Co-authors: Liming Xiang

The mixture cure model, typically combined the Cox proportional hazards model as the latency component for event time and logistic regression as the incidence component for the probability of cure, is often used to analyse survival data from subjects when a subset of them will never experience the event of interest. However, it is not realistic in some practical cases to assume the cure probability as a known transformation of a linear combination of covariates. We propose a double semiparametric mixture cure model for interval-censored data, allowing nonlinear effects of covariates on the cure probability through a generalized partially single-index model. We develop a Bayesian inference procedure for estimation based on a two-stage data augmentation method for deal with interval censored data, and polynomial splines for approximating nonlinear functions in both components of the proposed model. Simulation results demonstrate the finite sample performance of the proposed Bayesian procedure. To illustrate the proposed method, we apply the proposed procedure to analyse the data from a hypobaric decompression sickness study.

EO106 Room L1 ADVANCES IN MIXTURES WITH COVARIATES Chair: Salvatore Ingrassia

E1215: **Gaussian parsimonious clustering models with covariates**
Presenter: **Keefe Murphy**, University College Dublin, Ireland
Co-authors: Thomas Brendan Murphy

Model-based clustering methods are considered which account for external information available in the presence of covariates by introducing the MoEClust family of models, and a related software implementation. These finite mixture models allow the distribution of the latent cluster membership variable and/or the distribution of the response variables to depend on fixed covariates, under a range of parsimonious eigen-decomposition parameterisations of the component covariance matrices. Thus, the following equivalent aims are addressed: including covariates in Gaussian parsimonious clustering models and incorporating parsimonious covariance structures into the Gaussian mixture of experts framework. The MoEClust models demonstrate significant improvement from both perspectives in applications to data with multivariate responses and covariates of mixed type and provide richer insight into the type of observation which characterises each cluster.

E1234: **Time-varying measurement error in generalized linear models for longitudinal data: A two-step latent Markov approach**
Presenter: **Roberto Di Mari**, Department of Economics and Business, University of Catania, Italy
Co-authors: Antonello Maruotti, Antonio Punzo

A novel approach is proposed for longitudinal data modeling within the Generalized Linear Models (GLM) family, whenever a covariate of interest is affected by measurement error. We jointly model the response (outcome model), the covariate observed with error (measurement model) and the underlying unobserved error-free covariate (true score) along with its dynamics, assumed to follow a first-order latent (hidden) Markov chain. In a full (semi-parametric) maximum likelihood environment, computation is done by means of the EM algorithm. The estimation of the full joint model is hardly feasible as the number of covariates is large, as is typically the case in real-data applications. Thus, we propose a two-step approach to efficiently estimate model parameters. By means of extensive simulation studies, we show that both the one-step and the two-step approaches allow 1) to get correct estimates of the regression coefficients, as well as 2) reliable standard errors. In the real-data application, by modeling the true (unobserved) heart rate and its dynamics, we are able to find a significant effect of heart rate dynamics on the occurrence of a cardiovascular disease in a sample of +80 Chinese elderly.

E1210: **Generalized additive cluster weighted model**
Presenter: **Stefano Barberis**, University of Milano Bicocca, Italy
Co-authors: Salvatore Ingrassia, Giorgio Vittadini

An extension of mixture models with random covariates related to the Cluster Weighted Model (CWM) is presented for model-based clustering applications. The Generalized Additive Cluster Weighted Model (GAM-CWM) is a flexible model, able to capture complex relations between a response variable and a set of covariates in each mixture component. The main difference between models related to the CWM and other mixture models is that in CWM the joint probability $p(x,y)$ of a response variable y and a set of explanatory variables x is modelled in each mixture component rather than the conditional $p(y|x)$. The theory of generalized additive model extends the generalized linear model precisely with the aim of making it more flexible introducing a sum of smooth functions of covariates in the linear predictor. In the same way GAM-CWM extends the generalized linear CWM and the polynomial CWM defining a new powerful and very general class of models where the principles of CWM model and the GAM model are combined together. Maximum likelihood estimates are provided via EM algorithm and model selection is carried out using Bayesian Information Criterion (BIC) and Integrated Completed Likelihood (ICL). With simulated and real data are investigated performances, limits and benefits comparing this model with other mixture models related to it.

EO034 Room M1 NONPARAMETRIC FUNCTIONAL DATA ANALYSIS Chair: Davy Paindaveine

E0920: **A functional data depth based on moments**
Presenter: **Germain Van Bever**, Universite libre de Bruxelles, Belgium
Co-authors: Stanislav Nagy, Pauliina Ilmonen, Sami Helander, Lauri Viitasaari
The aim is to introduce a new depth concept in the functional setup. The integrated depths make up a large class of depth examples in the functional context, that is, say, in a situation where the observations are functions on some interval I. These depth functions typically consist in pointwise integration of (univariate or multivariate) depth values to achieve a global value. Several concepts exist. Consistency of these concepts were also studied. We introduce a new depth, based on moments of the distribution of depth values along I rather than standard integration. We study their universal asymptotic properties and illustrate their usefulness in the classification context. We show that, similar to existing depth concepts, the study of the distribution allows us to take into account variations in location, but also in the shape or roughness of the function.

E0870: **Nonparametric analysis of the shape of random curves**
Presenter: **Stanislav Nagy**, Charles University, Czech Republic
In many situations, the shape of functional observations is an important feature that must be taken into account in statistical analysis. The information about the shape properties can be extracted from the derivatives of the sample trajectories. Though, this approach can be applied only if the curves are regular and smooth, and the derivatives must be estimated. We present a simple alternative to this methodology based on simultaneous evaluation of multivariate projections of the data. This technique does not require smoothness or continuity, yet provides fine recognition of shape traits of the curves. The idea is illustrated on - but not limited to - the functional data depth.

E0469: **Exploratory functional data analysis from depth-based neighborhoods**
Presenter: **Raul Jimenez**, Universidad Carlos III de Madrid, Spain
Co-authors: Antonio Elias
The concept of depth has played an important role solving problems of ordering, outlier detection and clustering. We present an exploratory tool that visually provides more insights about the structure of a functional data set by the study of a simple undirected graph. To do so, we use the concept of depth-neighbourhood for defining a measure of closeness. This allows us to create a network with sample functions as nodes providing a different framework for studying a functional data set through the topology of the graph. Among others features, we show that a disconnected graph reveals the existence of clusters and how the degree of a node highlights inlier functions, outliers and groups boundaries.

EO270 Room O1 DEPENDENCE MODELS AND COPULAS II Chair: Wolfgang Trutschnig

E0466: **Using vine copulas to estimate the structure of directed acyclical graphs**
Presenter: **Eugen Pircalabelu**, Universita catholique de Louvain, Belgium
A new method of estimating and selecting a Bayesian network for continuous data is presented with the goal of stepping outside the class of multivariate normal distributions which are generally used due to their attractive properties. The method combines directed acyclic graphs and their associated probability models with copula C/D vines in order to construct 'copula based DAGs' which allow more flexibility in modeling joint distributions of pairs of nodes in the network. We exploit connections and similarities that exist between these two statistical techniques with the explicit purpose of estimating a directed graphical model, a network, for continuous data that are not necessarily normally distributed. The approach uses a score based learning scheme, where one modifies an initial graph based on improvements in the score, until a local maximum score is reached. A new information criterion is proposed and studied for graph selection tailored to the joint modeling of data based on graphs and copulas. Examples and simulation studies show the flexibility and properties of the method.

E1407: **Metropolis Hastings based estimation of generalized partition of unity copulas**
Presenter: **Andreas Masuhr**, University of Munster, Germany
The recently emerged family of Generalized Partition of Unity Copulas (GPUC) offer a new way for nonparametric modeling of dependencies by using a very general mixture approach. As a special case, GPUC also nest the versatile Bernstein copula, but can also allow for copulas that possess (upper) tail dependence. First, a prior distribution on the parameters of GPUC is established via importance sampling from the space of eligible parameter matrices. Subsequently, two estimation approaches based on the Metropolis-Hastings (MH) algorithm are proposed: a random walk MH and a random blocking random walk MH that makes use of the restrictions on the parameter space. Finally, simulation studies are carried out indicating the superiority of the proposed random blocking algorithm.

E1116: **Spatially homogeneous copulas**
Presenter: **Fabrizio Durante**, University of Salento, Italy
Co-authors: Juan Fernandez Sanchez, Wolfgang Trutschnig
Spatially homogeneous copulas are considered, i.e. copulas whose corresponding measure is invariant under special transformations of the unit square. Their main properties are studied with a view to possible use in stochastic models. In particular, we prove some symmetry properties and demonstrate how spatially homogeneous copulas can be used in order to construct copulas with surprisingly singular properties. Finally, a generalization of spatially homogeneous copulas is also given and a characterization of this new family of copulas in terms of the Markov product of copulas is established.

EO677 Room P1 BRANCHING PROCESSES: THEORETICAL, APPLIED AND COMPUTATIONAL ISSUES II Chair: Ines M. del Puerto

E1221: **Ancestral inference for tree-indexed data**
Presenter: **Anand Vidyashankar**, George Mason University, United States
Tree-indexed data arise in a variety of applications ranging from cell-kinetics to flow of information in social media. In such problems, it is customary to model the underlying tree as either a discrete-time random tree or a continuous-time random tree. The underlying stochastic processes generating the trees and the indexing data are typically correlated making the data analyses challenging. We describe methods to address the issue of ancestral inference; namely, given the data at some time t, how can one construct valid confidence/prediction intervals for parameters of the process at time 0? We develop some new theory to answer such questions and provide few illustrative examples.

E1304: **A two-type controlled branching process as model in cell kinetics**
Presenter: **Miguel Gonzalez Velasco**, University of Extremadura, Spain
Co-authors: Carmen Minuesa Abril, Ines M. del Puerto
Branching processes are relevant models in the development of theoretical approaches to problems in applied fields as, for instance, molecular biology, cell biology, epidemiology, and genetics. We deal with problems arising in cell biology. More specifically, we focus our attention on experimental data generated by time-lapse video recording of cultured oligodendrocyte cells. A reducible age-dependent branching process with emigration to model such population has been previously introduced. We propose a two-type controlled branching process to describe the embedded

discrete branching structure of such an age-dependent branching process. In this context, we tackle the estimation of the offspring distribution of the cell population. We address this problem from a Bayesian framework by making use of disparity measures. Statistical inference results are applied to a real data set.

E1367: **Griffiths-Tavare versus Lambert coalescents: Application in modeling of cancer evolution**
Presenter: **Marek Kimmel**, Rice University, United States

Recent years brought a large amount of work concerning retrospective reconstruction of cancer growth and mutation sometimes called the genetic archaeology of tumors. Most of this work has been based on the mathematical framework of Moran model or Kingman coalescent. However, neither of these approaches assumes an underlying model of proliferation, which would reflect cell divisions, frequently inefficient, of cells in tumors. One way of taking this latter into account is to base the coalescent trees on branching processes, with the simplest nontrivial scenario being the binary fission Markov age-dependent branching process aka birth and death process. This approach has been developed mostly by previously. We derive an explicit expression for the expectation of the site frequency spectrum (SFS) in Lambert model, and develop a simple and efficient simulation scheme based on the iid rv representation. We also examine how the SFS based on birth-and-death process differ from those based on Griffiths-Tavare model. This includes a discussion of the singleton estimation problem as well as the self-renewal fraction versus proliferation rate controversy.

EO406 Room Q1 FUNCTIONAL DATA ANALYSIS Chair: Alicia Nieto-Reyes

E0657: **Forecasting multiple functional time series: A static factor approach**
Presenter: **Gilles Nisol**, ULB, Belgium
Co-authors: Siegfried Hormann, Marc Hallin, Shahin Tavakoli

Theoretical foundations and a practical method to forecast multiple functional time series (FTS) are set. In order to do so, we generalize the static factor model to the case where cross-section units are FTS. We first derive a representation result. We show that if the K first eigenvalues of the covariance operator of the cross-section of the N FTS are unbounded while N grows and if the $K+1$ eigenvalue is bounded, then we can represent the FTS as a sum of a common component driven by K factors and an idiosyncratic component. We then set up an information criterion that chooses jointly the number K of factors and the dimension on which we should project the FTS before estimating the static factor model. We suggest a method of estimation and prediction based on these projected FTS. We assess the performances of the method and information criterion through a simulation exercise. Finally, we consider a real-data application. We show that by applying our method to a cross-section of PM10 concentration curves obtained across several measurement centers in Graz, we have a better prediction accuracy than by limiting the analysis to individual FTS.

E0869: **Simple spatio-temporal models for complex spatial data**
Presenter: **Rosaria Ignaccolo**, University of Turin, Italy
Co-authors: Lara Fontanella, Luigi Ippoliti, Pasquale Valentini

The focus is on the specification of a simple hierarchical generalized spatio-temporal model which warrants consideration when data sets with different types of spatial complexities are available. The model is a three-level hierarchical one, with a component specified by means of drift functions (e.g. we use a set of principal kriging functions or principal splines). Especially under Gaussian assumptions, the model is simple to estimate and particularly useful when reliable estimates of the parameters of a covariance function are difficult to obtain. Results from the analysis of different datasets have shown that our model can provide accurate predictions.

E1199: **Penalized robust estimation for functional regression**
Presenter: **Maria Francesca Carfora**, Istituto per le Applicazioni del Calcolo - CNR, Italy
Co-authors: Anestis Antoniadis, Italia De Feis

Scalar on function regression models, describing the relationship between a scalar response and a set of p functional predictors, are studied considering the problem of selecting the influential regressors in the presence of outliers. Using a classical basis projection approach, the continuous problem is replaced by a linear discrete one, permitting to adapt the classical penalized M estimators to a grouped problem. In particular the loss functions we will adopt include the Tukeys biweight, the Minimax Concave Penalty (MCP), the penalized Least Absolute Deviation (LAD), the nonnegative garrote, the Welsh and the Cauchy. Numerical implementations of the proposed procedures for proximal like algorithms are discussed. The results are illustrated with simulated examples and a real data analysis.

EO667 Room O2 ECOSTA JOURNAL: COMPUTATIONAL STATISTICS Chair: Erricos John Kontoghiorghes

E0739: **Numerical methods for SVD and its generalizations with applications in computational statistics**
Presenter: **Zlatko Drmac**, University of Zagreb, Croatia

The singular value decomposition (SVD) and its generalization, the GSVD (including the QSVD, PSVD and the cosine-sine decomposition CSD of partitioned orthonormal matrices) are the tools of trade in various applications, including computational statistics, least squares modeling, vibration analysis in structural engineering - just to name a few. In essence, the GSVD can be reduced to the SVD of certain products and quotients of matrices. For instance, in the canonical correlation analysis of two sets of variables x, y, with joint distribution and the covariance matrix $C = (C_{xx}, C_{xy}; C_{yx}, C_{yy})$, wanted is the SVD of the product $C_{xx}^{-1/2} C_{xy} C_{yy}^{-1/2}$. However, numerical algorithms are not that simple. We will review the recent advances in this important part of numerical linear algebra and propose improvements, with particular attention to *(i)* numerical robustness, where we show how the new generation of numerical algorithms returns accurate decomposition even in the cases that are considered ill-conditioned in the classical sense; *(ii)* development of reliable mathematical software that performs as predicted by error analysis and perturbation theory. Then, we illustrate the numerical performances on selected applications from computational statistics.

E1654: **Best L4 monotonic regression**
Presenter: **Ioannis Demetriou**, University of Athens, Greece

Let n measurements of a real valued monotonic function be given, where the measurements are so rough that have lost monotonicity. The problem of making the least sum of 4th powers change to the data so that the smoothed values have nonnegative first differences is considered. A special algorithm is proposed for this highly structured convex programming calculation. Some numerical results illustrate the method and compare it to the corresponding L1 and L2 calculations.

E0520: **Estimating the VC dimension with applications to model selection**
Presenter: **Bertrand Clarke**, University of Nebraska at Lincoln, United States
Co-authors: Merlin Mpoudeu

An objective function is derived that can be optimized to give an estimator of the Vapnik-Chervonenkis dimension for model selection in regression problems. We verify our estimator is consistent. Then, we verify it performs well compared to several other model selection techniques. We do this for simulated data, two benchmark data sets, and data from a designed agronomic experiment.

EO660 Room P2 RECENT ADVANCES IN BAYESIAN MODELING AND COMPUTATION Chair: Christopher Franck

E0332: **Scalable bayesian non-linear SVMs for big data problems**
Presenter: **Sounak Chakraborty**, University of Missouri, Columbia, United States

Bayesian non-linear SVM models are developed for Big Data platforms. In Big Data platforms, nonlinear SVMs are not very popular due to the difficulties in calculating and using the Gram/Kernel matrix. We employ a MCMC and Quasi-MCMC based solution to extract low dimensional random features and use them for approximating the Kernel matrix very efficiently and then use it in the model for faster and more accurate calculations. Our Bayesian SVM model is primarily for solving classification problems (binary and multiclass support vector machines). The feature selection is integrated in the framework Gaussian spike and slab priors. We propose a computationally scalable Gibbs sampling algorithm, which has linear computational complexity for covariate selections. In addition to that, we also consider Bayesian semi-supervised learning and propose a novel Bayesian approach for variable selection with scalable Gibbs algorithm. Our proposed novel Gibbs sampler called Skinny Gibbs which is much more scalable to high dimensional problems, both in memory and in computational efficiency. It can also avoid large matrix computations needed in standard Gibbs sampling algorithms. In terms of computational complexity for our Skinny Gibbs, it grows only linearly in the number of predictors. Efficiency of our method for supervised and semi-supervised SVM models are demonstrated based on several simulation studies and data analysis.

E0894: **Objective Bayesian analysis for Gaussian hierarchical models with intrinsic conditional autoregressive priors**
Presenter: **Christopher Franck**, Virginia Tech, United States
Co-authors: Matthew Keefe, Erica Porter, Marco Ferreira

Bayesian hierarchical models are commonly used for modeling spatially correlated areal data. Vague proper prior distributions have frequently been used for this type of model, which requires the careful selection of suitable hyperparameters. We propose a reference prior for hierarchical models with intrinsic conditional autoregressive spatial random effects. We present results from a simulation study that compares frequentist properties of Bayesian procedures that use several competing priors, including the derived reference prior. We demonstrate that using the reference prior results in favorable coverage, interval length, and mean squared error. Thus, the reference prior is a convenient automatic approach for the analysis of spatially correlated areal data that exhibits favorable inferential properties. We illustrate our methodology with an application to 2012 housing foreclosure rates in the 88 counties of Ohio. Finally, we share open-source computational resources available to everyday practitioners for fitting hierarchical models with intrinsic conditional autoregressive spatial random effects.

E1226: **Bayesian nonparametric differential analysis for dependent multigroup data with application to DNA methylation analyses**
Presenter: **Subharup Guha**, University of Florida, United States
Co-authors: Chiyu Gu, Veerabhadran Baladandayuthapani

Cancer' omics datasets involve widely varying sizes and scales, measurement variables, and correlation structures. An overarching scientific goal in cancer research is the development of general statistical techniques that can cleanly sift the signal from the noise in identifying genomic signatures of the disease across a set of experimental or biological conditions. We propose BayesDiff, a nonparametric Bayesian approach based on a novel class of first order mixture models, called the sticky Poisson-Dirichlet process or multicuisine restaurant franchise. The BayesDiff methodology flexibly utilizes information from all the measurements and adaptively accommodates any serial dependence in the data, accounting for the inter-probe distances, to perform simultaneous inferences on the variables. The technique is applied to analyze the motivating DNA methylation gastrointestinal cancer dataset, which displays both serial correlations and complex interaction patterns. In simulation studies, we demonstrate the effectiveness of the BayesDiff procedure relative to existing techniques for differential DNA methylation. Returning to the motivating dataset, we detect the genomic signature for four types of upper gastrointestinal cancer. The analysis results support and complement known features of DNA methylation as well as gene association with gastrointestinal cancer.

EO112 Room Q2 BAYESIAN SEMI- AND NONPARAMETRIC MODELLING II Chair: Raffaele Argiento

E0783: **Efficient Gibbs sampling methods for hierarchical processes**
Presenter: **Tommaso Rigon**, Bocconi University, Italy
Co-authors: Antonio Lijoi, Igor Pruenster

Within a Bayesian nonparametric framework, there is an increasing interest in flexibly learning how the distribution of a response variable changes across groups of observations. Popular models in such a setting are the hierarchical Dirichlet process and the wider class of hierarchical normalized random measures. The latter provides additional modeling flexibility, for instance because it allows for a deeper control of the underlying clustering mechanism. This obviously comes at a higher computational cost: posterior inference, although theoretically possible, might be cumbersome in practice. We aim to fill this gap by proposing an approximation for a general class of hierarchical processes, which leads to a straightforward Gibbs sampling algorithm. To this purpose, we employed a deterministic truncation of the involved random probability measures, obtaining a finite dimensional approximation of the original prior law. We provide both empirical and theoretical support for such a truncation. Our proposal is assessed through simulation studies and finally employed in an illustrative analysis.

E0355: **Detecting and leveraging structural information with Bayesian forests**
Presenter: **Antonio Linero**, Florida State University, United States

Bayesian methods based on ensembles of decision trees, such as Bayesian additive regression trees (BART) have proven to be extremely useful in a wide variety of statistical problems. We show how a-priori known structural information, such as graphical or group structures of the predictors, can be used to improve the efficiency and stability of the ensemble. This structural information is encoded through priors on the splitting proportions of BART ensembles, and can be used to encode, for example, sparsity within or between groups of predictors. We also consider the problem of detecting interactions among predictors in BART-type models. Using a clustering of trees in the ensemble, we allow the model to smoothly vary between a sparse additive model (SPAM) model and a dense model in which interactions between variables are not penalized. We illustrate the methodology proposed on a variety of simulated and real datasets.

E0853: **A Bayesian non-parametric causal inference model for synthesizing randomized clinical trials and real-world evidence**
Presenter: **Gary Rosner**, Johns Hopkins University, United States
Co-authors: Chenguang Wang

With the wide availability of various real-world data (RWD), there is an increasing interest in synthesizing information from both randomized clinical trials and RWD for health-care decision making. The task of addressing study-specific heterogeneities is one of the most difficult challenges in synthesizing data from disparate sources. Bayesian hierarchical models with non-parametric extensions provide a powerful and convenient platform that formalizes the information borrowing strength across the sources. We propose a propensity score-based Bayesian non-parametric Dirichlet process mixture model that summarizes subject-level information from randomized and registry studies to draw inference on the causal treatment effect. Simulation studies are conducted to evaluate the model performance under different scenarios. We demonstrate the proposed method using data from a clinical study.

EG031 Room H1 CONTRIBUTIONS IN DIRECTIONAL DATA Chair: Arthur Pewsey

E1420: **New contributions to Mobius transformation induced distributions on the disc**
Presenter: **Priyanka Nagar**, University of Pretoria, South Africa
Co-authors: Andriette Bekker, Mohammad Arashi

There is a need for developing flexible distributions on the hyper-disc, which has support of the interior of the hyper-sphere, as it allows for modelling the combination of angular and linear observations. A new family of distributions is proposed which has support on the unit disc in two dimensions that includes the bivariate spherically symmetric beta distribution. By applying a conformal mapping to this distribution the new Mobius distribution class emanates. Modality, symmetry, marginal distributions and maximum likelihood estimation are considered. The flexible behaviour of the proposed models on the hyper-disc will be graphically demonstrated and applied to model fitting.

E1415: **Estimating the wrapped stable distribution via indirect inference**
Presenter: **Marco Bee**, University of Trento, Italy

Directional data are frequently encountered in applications and require a special treatment. One way of constructing probability distributions for directional data exploits the idea of wrapping on the unit circle a distribution defined on the real line. We study estimation of the wrapped stable distribution, and propose a novel approach based on constrained indirect inference. Since the wrapped stable density does not exist in closed-form, simulation-based methods are an appealing alternative to maximum likelihood. The problem is tackled by means of a strategy already used for indirect inference estimation of the linear stable distribution. In particular, we use the wrapped version of the same auxiliary model, namely the skewed-t distribution. We study numerically the impact of the inputs, and especially of the weighting matrix, in finite samples. Simulation experiments suggest that indirect inference is more efficient than numerical maximum likelihood, from both the statistical and the computational point of view.

E1216: **New algorithms for wrapped normal models estimation**
Presenter: **Anahita Nodehi**, Tarbiat Modares University, Iran
Co-authors: Claudio Agostinelli, Mousa Golalizadeh

There are a lot of discussions in every statistical context about how to estimate the parameters after choosing a model. One of the crucial problem which deal with wrapped normal distribution is estimating the parameters, especially in multivariate cases. This is due to the form of the density function which is constituted by large sums, and cannot be simplified as close form. The likelihood-based inference for such distribution can be very complicated and computationally intensive. Also, periodic feature of data makes all methods in hands infeasible. The statistics to deal with such data is called directional statistics. Since the shortest distance between two points are not straight line as Euclidean space, it is worth to extend existing methods for such data. Two fast and reliable methods based on Expectation-Maximization (EM) and Classification Expectation-Maximization (CEM) algorithm are suggested. We show the performance of proposal methods in simulation study and real application in compare to existing iterative method.

EG207 Room N1 CONTRIBUTIONS IN EXTREME VALUES Chair: Yuri Goegebeur

E1527: **Mixed-frequency extreme value regression: Estimating the effect of MCS on extreme rainfall in the Midwest**
Presenter: **Luca Trapin**, Universita Cattolica del Sacro Cuore Milano, Italy
Co-authors: Debbie Dupuis

More frequent and longer-lasting mesoscale convective systems (MCS) are the principal driver of observed increases in springtime extreme rainfall in the Central United States. We develop new models that integrate and exploit hourly MCS information in analyses of extremes of maximum hourly rainfall over a much longer time period, e.g. one month, and gain some insight into the increases to these extremes and how MCS may have driven the changes. This requires extreme value regression models handling observations sampled at different frequencies. We borrow some elements from the MIxed-DAta Sampling (MIDAS) regression literature and propose a flexible, data-driven aggregation scheme to face this challenge. We study the monthly maximum hourly precipitation in five US midwest cities from 1979 to 2014. We model these maxima with a Generalized Extreme Value (GEV) distribution, and let the location parameter of this model vary as a function of the monthly number of MCS occurring in each of the 24 hours covering a day. Our mixed-frequency GEV model confirms that the occurrence of an MCS is a good predictor of the extreme rainfall, also reveals that MCS occurring in different parts of the day contribute differently to explain the increased rainfall intensity.

E1556: **Flexible extreme value modelling in insurance and finance**
Presenter: **Gaonyalelwe Maribe**, University of Pretoria, South Africa

Extreme value theory (EVT) is used to model rare events (extreme observations). One particular approach in EVT is the Peaks over threshold (POT) method, which employs the asymptotically motivated generalized Pareto distribution (or simple Pareto) to model exceedances (or relative excesses) above a sufficiently high threshold. The choice of this threshold however remains an open question. In recent years several attempts have been made to extend tail modelling towards the modal part of the data. Dynamic mixtures of two components with a weight function smoothly connecting the bulk and the tail of the distribution has been proposed. Most recently a review of this topic has been made, along with a statistical model which is in compliance with extreme value theory and allows for a smooth transition between the modal and tail part. We discuss special cases of these approaches and revisit and extend the second order refined POT approach to all max-domains of attraction using flexible semiparametric modelling of the second order component. Practical cases in insurance and finance are given were such models can be of importance.

E1343: **Use of censored distribution in the intervals estimator of the extremal index**
Presenter: **Jan Holesovsky**, Brno University of Technology, Czech Republic
Co-authors: Michal Fusek

From the theory it follows that the local dependence in a stationary series causes clustering of extreme values. Hence, the inference for extremes typically requires proper identification of clusters of high threshold exceedances and estimation of the extremal index which is the primary measure of the local dependence. An intervals estimator of the extremal index based on the distribution of interexceedances times has been previously introduced. Direct application of the limiting distribution to interexceedances times of a stationary series may cause the intervals estimator to be biased toward independence. Several modifications have been proposed including the K-gaps likelihood estimator, where K determines the intra- and intercluster spacings. The aim is to introduce a new estimator of the extremal index based on censored distributions that can be viewed as an alternative to the K-gaps estimator without using fixed replacements of the intracluster spacings. Properties of the estimator are studied using simulations. The main benefit lies in reducing the bias of the estimates, especially when large clusters are present in the series.

CO116 Room B2 THE ECONOMETRICS OF CRYPTOCURRENCIES Chair: Leopoldo Catania

C0325: **Conditional tail-risk in cryptocurrency markets**
Presenter: **Nicola Borri**, LUISS University, Italy
The CoVaR risk-measure is used to estimate the conditional tail-risk in the markets for bitcoin, ether, ripple and litecoin and find that these cryptocurrencies are highly exposed to tail-risk within cryptomarkets while they are not exposed to tail-risk with respect to other global assets, like the U.S. equity market or gold. Although cryptocurrencies are highly correlated one with the other, both unconditionally and conditionally, we find that idiosyncratic risk can be significantly reduced and that portfolios of cryptocurrencies offer better risk-adjusted and conditional returns than the individual cryptocurrencies. These results indicate that portfolios of cryptocurrencies could offer attractive returns and hedging properties when included in investors' portfolios.

C1111: **Bitcoin price dynamics and market attention**
Presenter: **Gianna Figa Talamanca**, University of Perugia, Italy
Recent developments are addressed about Bitcoin price modeling and related applications. Precisely, we consider a bivariate model to describe the behavior of Bitcoin price and of the investors' attention on the overall network. The attention index affects Bitcoin price through a suitable dependence of the drift and diffusion coefficients and a possible correlation between the sources of randomness represented by the driving Brownian motions. The model is fitted on historical data of Bitcoin prices, by considering the total trading volume and the Google search volume index as proxies for the attention measure. Moreover, a closed formula is computed for European style derivatives on Bitcoin. Finally, we discuss two possible extensions of the model. Precisely, we investigate the relation between the correlation parameter and possible bubble effects in the asset price; further, we consider a multivariate framework to represent the special feature of Bitcoin being traded on several exchanges and we discuss conditions to rule out arbitrage opportunities in this setting.

C1478: **Market risk of cryptocurrencies**
Presenter: **Annalisa Molino**, University of Rome Tor Vergata, Italy
Due to the increasing popularity of cryptocurrencies, understanding the risk features of this market is important for both investors and regulators. An analysis of the risk of holding hypothetical portfolios made of cryptocurrencies is carried out by using Monte Carlo simulations. The key decision for a Monte Carlo simulation is the choice of the probability distribution that better approximates the return distribution. The latter has to be modeled with a particular focus on the tails, from which the extreme quantiles are extracted. The univariate distribution of returns is modeled by a combination of extreme value theory and kernel estimation, and copulas are used to model their dependence. The findings point to the fact that cryptocurrencies returns have peculiar characteristics that make them a very risky investment: the portfolios experience indeed extraordinary large gains and losses. Due to the zero or low correlations with fiat currencies, allocating a percentage of investment in cryptocurrencies reduces the risk of holding a hypothetical portfolio of fiat currencies for one month. Besides the empirical study, the contribution is twofold. First, it suggests a semiparametric way to model the distribution of cryptocurrencies. Second, it suggests a flexible way to estimate the Value-at-Risk and the Expected Shortfall of cryptocurrencies.

CO058 Room D2 NETWORK ECONOMETRICS Chair: Roberto Casarin

C0904: **Predictability, spillover, and disagreement in signed financial networks**
Presenter: **Lorenzo Frattarolo**, Universita Ca Foscari, Italy
Co-authors: Monica Billio, Roberto Casarin, Michele Costola
The link between return predictability and disagreement is investigated by studying consensus dynamics on VAR(1)-based network on stock returns. We include the sign information of the VAR(1) in the analysis of disagreement on statistical financial networks by proposing two new financial stability measures: time to consensus and disagreement persistence.

C1004: **Bayesian dynamic tensor regression**
Presenter: **Monica Billio**, University of Venice, Italy
Co-authors: Roberto Casarin, Sylvia Kaufmann, Matteo Iacopini
Multidimensional arrays (i.e. tensors) of data are becoming increasingly available and call for suitable econometric tools. We propose a new dynamic linear regression model for tensor-valued response variables and covariates that encompasses some well-known multivariate models such as SUR, VAR, VECM, panel VAR and matrix regression models as special cases. For dealing with the over-parametrization and over-fitting issues due to the curse of dimensionality, we exploit a suitable parametrization based on the parallel factor (PARAFAC) decomposition which enables to achieve both parameter parsimony and to incorporate sparsity effects. The aim is twofold: first, we provide an extension of multivariate econometric models to account for both tensor-variate response and covariates; second, we show the effectiveness of proposed methodology in defining an autoregressive process for time-varying real economic networks. Inference is carried out in the Bayesian framework combined with Monte Carlo Markov Chain (MCMC). We show the efficiency of the MCMC procedure on simulated datasets, with different size of the response and independent variables, proving computational efficiency even with high-dimensions of the parameter space. Finally, we apply the model for studying the temporal evolution of real economic networks.

C1219: **Idiosyncratic volatility puzzle: The role of assets' interconnections**
Presenter: **Roberto Panzica**, Goethe University House of finance, Italy
The aim is to investigate the determinants of the idiosyncratic volatility puzzle by allowing linkages across asset returns. The first contribution is to show that portfolios sorted by increasing indegree computed on the network based on Granger causality test have lower expected returns, not related to idiosyncratic volatility. Secondly, empirical evidence indicates that stocks with higher idiosyncratic volatility have the lower exposition on the indegree risk factor.

CO390 Room E2 ECONOMETRICS FOR POLICY ANALYSIS Chair: Christos Savva

C0202: **Monetary policy and bank lending behavior**
Presenter: **Christos Savva**, Cyprus University of Technology, Cyprus
Co-authors: Demetris Koursaros, Nektarios Michail
The aim is to test the conjecture that easy money policies of central banks, setting low rates for long, could have an impact on bank lending behavior. If the conjecture holds then policy rate shocks should have persistent effects on bank behavior either through the bank lending or the risk-taking channel. Using ten euro-area countries under a shock persistence methodology, we find only country-specific, idiosyncratic effects of the policy rate on bank lending growth and no effect on credit risk. Consequently, the findings do not support the view that the prolonged duration of relatively low rates has been the culprit for the excess risk-taking behavior, and instead support the view proposed by Milton Friedman that changes in the policy rate can only have a transitory impact on the economy.

C0559: **On the impact of quantitative easing**
Presenter: **Nektarios Michail**, Cyprus University of Technology, Cyprus

Transmission channels through which asset purchases are supposed to affect the economy are examined using US data in a Bayesian VAR setup. Once we distinguish between GDP components, in order to account for the increase in government spending during the period, the results suggest that quantitative easing is likely transmitted to the economy only through the portfolio rebalancing channel. No support is found for the uncertainty and expectations channels. Asset purchases lower the cost of public debt, which assists in increasing government spending, to some extent. The conclusions are the same regardless of using monthly or quarterly data. Overall, asset purchases may have eased funding conditions albeit they had little impact on the real economy. It appears more likely that QE was just beneficial in lowering the governments cost of debt, without whose increase it would be unlikely that the economy would have been affected.

C0203: **Sales and promotions and the great recession deflation**
Presenter: **Demetris Koursaros**, Cyprus University of Technology, Cyprus
Co-authors: Christos Savva, Niki Papadopoulou

The effect of sales and promotions on the pricing decisions of firms is investigated. A theoretical model is provided where firms face menu costs when adjusting their price and apply sales offers that decrease temporarily the listed price to attract higher demand, especially because households exert effort to locate the price deals. Thus, each period the final price is determined by the price set by the firm which is common knowledge to all agents and a sales deal that is a draw from a distribution with endogenous time-varying support. In a recession, even though prices in the economy look sticky, firms increase the frequency and the range of sales on their products substantially. This implies that traditional inflation measures are overstated in recessions, because they ignore the surge in sales and promotions and the consumers' tendency to hunt those limited time offers more actively. This framework can explain the mild deflation experienced during the Great Recession. Moreover, it is demonstrated that using traditional inflation measures can prolong recessions.

CO056 Room F2 ECONOMETRICS OF ART MARKETS Chair: Douglas Hodgson

C0257: **Participation in the Venice biennale and the implications for artists careers and trajectories: Evidence from Australia**
Presenter: **Bronwyn Coate**, RMIT University, Australia

Research being conducted in conjunction with the Australia council of the arts is drawn on to explore how participation in the Venice Biennale impacts Australian artists' careers and professional trajectories. We hypothesise that as a signal of artist quality and acceptance by the visual arts sector, participation should aid artists ' career development reflected in the prices of works by represented artists. Currently there is little empirical evidence to support this view and, given the current orientation of Australian arts funding directed to this endeavour, the aim is to provide new evidence on the effects of participation to benefit artists and the reputation of Australian art internationally. For the study we draw on auction sales data for the set of Australian artists who have represented Australia at the Biennale since 1978 ($n = 39$). We model the effect that participation has upon prices of works by the artists to establish the presence of a Venice Biennale price premium associated with sales occurring around the time of participation. Interestingly evidence on the longer term impact of participation is mixed; suggesting that participation in the Venice Bienalle this does not in itself guarantee artists commercial success.

C0260: **Artistic movement membership and the career profiles of Canadian painters**
Presenter: **Douglas Hodgson**, UQAM, Canada

Psychologists and economists have studied many aspects of the effects on human creativity, especially that of artists, of the social setting in which creative activity takes place. In the last hundred and fifty years or so, the field of advanced creation in visual art has been heavily characterized by the existence of artistic movements, small groupings of artists having aesthetic or programmatic similarities and using the group to further their collective programme, and, one would suppose, their individual careers and creative trajectories. Certainly this is true of Canadian painting, and such movements as the Group of Seven or the Automatistes are at least as well-known to the general public as the individual artists belonging to them. We econometrically investigate the effect on career dynamics of artists as represented by the life-cycle pattern of prices obtained by their works at auction, in estimating a hedonic regression, pooled over a large sample of Canadian painters, in which variables representing the effect of a number of specific movements on the career price profiles of the members of the movements are included. These pooled movement effects are then compared with individual profiles obtained from individual-level models to gauge the degree to which these latter are influenced by movement membership.

C1058: **From afternoon to evening: Price dynamics and bidding behaviour in evening auctions for fine art**
Presenter: **Christiane Hellmanzik**, Technical University of Dortmund, Germany
Co-authors: Roland Fuess, Moritz Burkhardt

The global auction market continues to deliver extraordinarily high prices for Post-War and Contemporary Art which might be indicative of sentiment-driven behaviour rather than rational investment motives. In order to test this hypothesis, we exploit variation in the timing of auctions to investigate whether an evening auction effect exists. Such an effect captures the tendency of auction participants to condition their price expectations during day auction on the relative price level of preceding evening auctions. Although only weak evidence has been observed of record prices positively affecting subsequent auction results, bidders do perceive a high share of bought-in items as negative pricing signal.

CO080 Room H2 COINTEGRATION: STABILITY, LINEARITY AND MONITORING Chair: Martin Wagner

C1329: **Testing linear cointegration against smooth transition cointegration: Theory and an application to long-run money demand**
Presenter: **Martin Wagner**, Technical University Dortmund, Germany
Co-authors: Oliver Stypka

Simple tests are developed for the null hypothesis of linear cointegration against the alternative of smooth transition cointegration. The test statistics are based on extensions of the fully modified and integrated modified OLS estimators to render them applicable to Taylor approximations of smooth transition functions. We consider both integrated variables as well as time as transition variables. For the integrated modified OLS based tests we consider in addition to standard asymptotic inference also fixed-b inference. The properties of our tests are assessed by means of a simulation study that also includes a previous test as benchmark. Finally, we apply our tests to investigate linearity respectively stability of long-run money demand for a number of individual countries as well as the Euro area. We find strong evidence against linearity and stability of long-run money demand.

C1473: **Cointegrating polynomial regression with an integrated regressor with drift: Fully modified OLS estimation and inference**
Presenter: **Karsten Reichold**, Technical University Dortmund, Germany
Co-authors: Martin Wagner

Fully modified OLS cointegrating polynomial regression analysis is reconsidered by focusing on the case where the integrated regressor has a drift, with in general unknown drift parameter. In case the deterministic component and the powers of the integrated regressor share at least one identical power of time, the ensuing asymptotic multi-collinearity needs to be addressed. This is done, as usual in the unit root and cointegration literature, by an appropriate linear transformation of the stochastic regressors. The corresponding inverse transformation then leads to the singular

limiting distribution of the FM-OLS estimator corresponding to the untransformed variables. It is important to note that in case of unknown drift the FM-OLS limiting distribution of the intercept is contaminated by a second order bias term stemming from the fact that the drift parameter is also estimated at rate square root of sample size.

C1491: **Bubble detection in error correction models**
Presenter: **Leopold Soegner**, Institute for Advanced Studies, Austria
Co-authors: Martin Wagner
Consistent monitoring procedures are developed with the goal of detecting bubbles in a Johansen-type error correction model. In particular, we consider breaks where the cointegration rank remains constant as well as breaks changing the cointegration rank. We develop Lagrange multiplier tests allowing to monitor these kinds of breaks. The monitoring procedure is used to detect possible bubbles in the triangular arbitrage parity.

CO100 Room N2 RECENT ADVANCE IN COMPLEX TIME SERIES ANALYSIS Chair: Jinyuan Chang

C0323: **Lasso-driven inference in time and space**
Presenter: **Chen Huang**, University of St. Gallen, Switzerland
Co-authors: Victor Chernozhukov, Wolfgang Haerdle, Weining Wang
The estimation and inference in a system of high-dimensional regression equations is considered allowing for temporal and cross-sectional dependency in covariates and error processes, covering rather general forms of weak dependence. A sequence of large-scale regressions with lasso is applied to reduce the dimensionality, and an overall penalty level is carefully chosen by a block multiplier bootstrap procedure to account for multiplicity of the equations and dependencies in the data. Correspondingly, oracle properties with a jointly selected tuning parameter are derived. We further provide high-quality de-biased simultaneous inference on the many target parameters of the system. We provide bootstrap consistency results of the test procedure, which are based on a general Bahadur representation for the Z-estimators with dependent data. Simulations demonstrate good performance of the proposed inference procedure. Finally, we apply the method to quantify spillover effects of textual sentiment indices in a financial market and to test the connectedness among sectors.

C0554: **Efficient estimation by fully modified GLS with an application to the environmental Kuznets curve**
Presenter: **Yicong Lin**, Maastricht University, Netherlands
Co-authors: Hanno Reuvers
The aim is to develop the asymptotic theory for a Fully Modified Generalized Least Squares (FMGLS) estimator for multivariate cointegrating polynomial regressions. These regressions allow both deterministic and stochastic trends and their integer powers to enter the cointegrating relation and therefore form natural extensions of the linear cointegration framework. The FMGLS estimator relies on the inverse autocovariance matrix of the multidimensional errors and requires bias correction terms for endogeneity and serial correlation. These quantities are unknown in practice. To obtain a feasible FMGLS estimator, we propose a consistent estimator of the inverse matrix. The bias correction terms for endogeneity and serial correlation are conveniently obtained as byproducts from our estimation framework. The resulting feasible FMGLS estimator allows for standard asymptotic inference. Extending earlier work, we elaborate on the conditions that make this FMGLS estimator asymptotically equivalent to its ordinary least squares counterpart. Both finite sample and asymptotic efficiency gains can be substantial if the least squares estimators have different limiting distributions. A comprehensive simulation study supports our theoretical outcomes. As a practical illustration, we test for the Environmental Kuznets Curve (EKC) hypothesis in a panel of European countries.

C1228: **Estimation of subgraph densities in noisy networks**
Presenter: **Jinyuan Chang**, Southwestern University of Finance and Economics, China
Co-authors: Eric Kolaczyk, Qiwei Yao
While it is common practice in applied network analysis to report various standard network summary statistics, these numbers are rarely accompanied by some quantification of uncertainty. Yet any error inherent in the measurements underlying the construction of the network, or in the network construction procedure itself, necessarily must propagate to any summary statistics reported. We first study the problem of estimating the density of edges in a noisy network. Under a simple model of network error, we show that consistent estimation of such densities is impossible when the rates of error are unknown and only a single network is observed. We then develop method-of-moment estimators of network edge density and error rates for the case where a minimal number of network replicates are available. These estimators are shown to be asymptotically normal. We also provide the confidence intervals for quantifying the uncertainty in these estimates based on either the asymptotic normality or a bootstrap procedure. We further investigate the estimation for higher-order subgraph counts such as those for 2-star edges and triangles. Bootstrap confidence intervals for those high-order counts are constructed based on a new algorithm for constructing a graph with the pre-determined counts for edges, two-star edges and triangles.

CG061 Room M2 CONTRIBUTIONS IN INTERNATIONAL FINANCE Chair: Matteo Cacciatore

C0716: **Financial cycles across G7 countries: A view from wavelet analysis**
Presenter: **Michael Scharnagl**, Deutsche Bundesbank, Germany
Co-authors: Martin Mandler
The cross-country dimension of financial cycles is analyzed by studying cyclical co-movements in financial variables and house prices across the G7 economies. We use wavelet-based statistics to assess at which frequencies cyclical fluctuations and their cross-country co-movements are important and how these change over time. We show cycles in interest rates and equity prices to be at least as synchronised as cycles in real GDP while cycles in credit and house prices are less synchronised. FR, UK and US have closely linked cycles in all financial variables and house prices. Furthermore, credit and house price cycles are linked to cycles in real GDP in many countries. Frequency ranges and country coverage of co-movements highlight the importance of country-specific developments.

C1488: **Intra euro area capital flows and the current account balance**
Presenter: **Andreas Savvides**, Cyprus University of Technology, Cyprus
Co-authors: Nektarios Michail
The purpose is to examine the hypothesis that one of the fundamental factors behind the Euro Area (EA) crisis is the reversal of intra-EA capital flows between the core and periphery economies. We compile a data set on aggregate and disaggregate gross and net intra-EA flows between Germany and Greece, Ireland, Italy, Portugal and Spain (EA-5) during 1999Q1-2016Q1. We find evidence of a reversal of intra-EA gross capital inflows to the EA-5 after the crisis that can be explained by surges/stops in gross capital inflows and general economic and financial conditions in the EA-5. We estimate a panel VAR model of the joint determination between the current account balance of the EA-5 and foreign capital flows distinguishing between intra-EA capital flows and from the rest of the world. The impulse response functions reveal an important role for capital flows from Germany to the EA-5 in financing the current account balance of the EA-5.

C1448: **A global look at stock market comovements**

Presenter: **Kei Ichiro Inaba**, Bank of Japan, Japan

International stock market comovements are analysed for 37 advanced and emerging countries in 1996-2015. The degrees of the comovements were substantial: the sample-country and sample-period average is 56 per cent. These degrees had upward and downward trends in 23 and 7 of the sample countries, respectively. They were greater in advanced countries than in emerging ones. The comovements changed over time in a similar fashion for different country-groups - a representation of a global financial cycle -, but increased more rapidly in emerging countries than in advanced ones. The driving forces behind differences in the commovements across countries and over time were country-specific heterogeneities and time-varying factors. These factors relate to the scale of national economy, the openness of international trade, and policies of monetary authorities: the level of short-term interest rates, the openness of the capital account, and the variability of foreign exchange rates. The comovement tended to be smaller in a country with a less open capital account and a less flexible foreign exchange market. A countrys short-term interest-rate differentials with respect to the United States were a negative determinant of the countrys stock market comovement when its capital account is controlled- corroboration for a monetary policy dilemma.

Sunday 16.12.2018 10:05 - 12:10 Parallel Session L – CFE-CMStatistics

EO675 Room A0 DATA INTEGRATION AND SAE FOR EQUITABLE AND SUSTAINABLE DEVELOPMENT Chair: Caterina Giusti

E1178: **Small area estimation of latent economic wellbeing**
Presenter: **Angelo Moretti**, University of Manchester, United Kingdom
Co-authors: Natalie Shlomo, Joseph Sakshaug
Factor analysis (FA) models are used in data dimensionality reduction problems where the variability among observed variables can be described through a smaller number of unobserved latent variables. This approach is often used to estimate the multidimensionality of wellbeing. We employ FA models and use multivariate EBLUP (MEBLUP) to predict a vector of means of factor scores representing economic wellbeing for small areas. We compare this approach to the standard approach whereby we use SAE (univariate and multivariate) to estimate a dashboard of EBLUPs on original variables and then averaged. Our simulation study shows that the use of factor scores provides estimates with lower variability than weighted and simple averages of standardised MEBLUPs and univariate EBLUPs. Moreover, we find that when the correlation in the observed data is taken into account before small area estimates are computed multivariate modelling does not provide large improvements in the precision of the estimates over the univariate modelling. We close with an application using the EU survey on income and living conditions data with a particular focus on economic wellbeing.

E1190: **Small area estimation in the framework of multivariate models for sustainable development**
Presenter: **Emilia Rocco**, University of Florence, Italy
Co-authors: Emilia Rocco, Maria Francesca Marino, Alessandra Petrucci
The analysis of complex phenomena, such as the equitable and sustainable development, often requires the estimation of correlated descriptive measures. Multivariate models are specifically designed to take into account the correlation of several variables and, typically, fit to these kind of situations. We introduce a multivariate mixed model for the case in which small area estimates of different, possibly heterogeneous, survey variables are required. In this framework, estimates of model parameters in the different equations may borrow strength one from the other and, thus, efficiency may be improved. In particular, this is achieved by introducing in the model specification a set of correlated latent effects that allow us to capture the dependence among the different outcomes of interest. Further, the estimated correlation between the latent effects provides an indirect measure of the dependence between the outcomes themselves and, thus, offers a deeper understanding of the phenomenon under investigation. Estimation of model parameters is discussed within a likelihood-based approach. Predictions of small area parameters are derived and a parametric bootstrap method is proposed for estimating the mean squared error of such predictions. Results are supported by an intensive simulation study in which different scenarios are investigated.

E1230: **Quality evaluation of statistical processes based on administrative data**
Presenter: **Roberta Varriale**, Centro Dagum c/o Dip. Economia e Management, Universita di Pisa, Italy
Co-authors: Fabiana Rocci, Orietta Luzi
Over the last decade, National Statistical Institutes (NSIs) have progressively moved from single- to multi-source statistics. By combining different data sources (direct survey, administrative and big data) NSIs can increase the detail of information, save data production costs and reduce burden on respondents. The Italian NSI (Istat) has strongly increased the use of administrative archives as primary source for statistical production purposes. To this aim, a system of statistical registers based on the integrated use of administrative sources is under development, and many statistical processes have being accordingly re-designed. Such a change calls for a tailoring of the current approaches for quality measurement and assessment. While in Istat a total quality framework based on the Total Survey Error (TSE) is well developed for surveys, a quality framework supporting the design of the new required statistical processes, based on the use of several types of sources, their evaluation and monitoring is still missing. To this extent, the adaptation of the TSE lately proposed in literature for statistical processes using administrative data sources has been taken as reference. We illustrate as the proposed quality framework has been tested on a new process - the statistical register Frame-SBS - that supports the estimation of structural statistics on businesses. As a major result, a proposal for an additional quality assessment phase is described.

E1353: **Regional disparities in well-being dimensions: The case of Italy**
Presenter: **Laura Neri**, Universita di Siena, Italy
Co-authors: Achille Lemmi, Marco Lonzi
Goal 10 of the SDGs explicitly states that economic growth is not sufficient to reduce poverty if it is not inclusive and if it does involve the three dimensions of sustainable development economic, social and environmental. According to this view, we identify three multidimentional indicators of well-being and analyse them at regional level in Italy. The indicators involved in the analysis, as identified and defined previously, are the environment indicator (FS5), the financial indicator (FS4), and the work and education indicator (FS6). The idea of considering them as key indicators is that despite the fact that in Italy statistical indicators does not show an increase in inequality, a widespread sense of impoverishment and weakening of future prospects is perceived by people in low income and in the disappeared middle class. For the most disadvantaged, it is very difficult to break away from a vicious circle of educational underachievement, low skills, poor employment prospects and lacking means to adapt to a changing economy and fast technological change. Moreover, they also suffer disproportionally by pollution and environmental degradation. On the other hand, inclusive growth is the process of creating shared prosperity in many well-being dimensions, many of those are heavily conditioned by peoples opportunities which are themselves heavily affected by peoples socioeconomic status.

E1676: **Local index of house prices in Italy**
Presenter: **Davide Fiaschi**, University of Pisa, Italy
Co-authors: Luigi Biggeri, Monica Pratesi, Tiziana Laureti
An index of house prices is developed at Italian sub-municipality level for the period 2002-2016 starting from a dataset on the transaction prices made available by OMI, and complemented by information on the composition of the local stock of houses. We then use this local index to study the spatial dispersion of house prices and its dynamics over time. We characterize the speed of diffusion of local shocks as well as the importance of aggregate shocks for the observed dynamics of house prices. We conclude by comparing this index with local measures of nominal wages, thus providing an insight on spatial dispersion of Italian real wages.

E1451: **Discussions on 'Advances in data integration and SAE for equitable and sustainable development'**
Presenter: **Jacques Silber**, Bar-Ilan University, Israel
Several papers in the session entitled Advances in data integration and SAE for equitable and sustainable development are discussed. Particular attention will be given to different approaches to the use of small areas data to estimate multi-dimensional well-being and sustainable development. The techniques used will be compared, and the empirical findings of these different studies discussed.

EO276 Room Aula 5 DATA PRIVACY AND STATISTICAL DISCLOSURE CONTROL Chair: Matthew Reimherr

E0319: **Differentially private uniformly most powerful tests for binomial data**
Presenter: **Jordan Awan**, Penn State University, United States
Co-authors: Aleksandra Slavkovic
Uniformly most powerful (UMP) tests are derived for simple and one-sided hypotheses for a population proportion within the framework of Differential Privacy (DP), optimizing finite sample performance. We show that in general, DP hypothesis tests for exchangeable data can always be expressed as a function of the empirical distribution. Using this structure, we prove a 'Neyman-Pearson lemma' for binomial data under DP, where the DP-UMP only depends on the sample sum. The tests can also be stated as a post-processing of a random variable, whose distribution we coin 'Truncated-Uniform-Laplace' (Tulap), a generalization of the Staircase and discrete Laplace distributions. Furthermore, we obtain exact p-values, which are easily computed in terms of the Tulap random variable. We show that our results also apply to distribution-free hypothesis tests for continuous data. Our simulation results demonstrate that our tests have exact type I error, and are more powerful than current techniques.

E0797: **Improved differentially private regression and classification with Gaussian processes**
Presenter: **Michael Smith**, University of Sheffield, United Kingdom
Co-authors: Mauricio Alvarez
The cloaking method described previously applies differential privacy (DP) to the outputs of Gaussian process (GP) regression, achieving successful predictions for low-dimensional datasets in regions of high data density. We cover several shortcomings of the cloaking method, starting with the problem of predictions in the surrounding outlier regions of the dataset. We experiment with the use of inducing inputs to provide a sparse approximation and show that these can provide robust differential privacy in sparse areas and at higher dimensions. We then show how one can use the framework of coregionalised multiple output GPs to provide group privacy and how one can perform GP classification by applying the cloaking method to the optimisation step in the Laplace approximation. We finally look at the issue of DP hyperparameter selection. Overall this provides a useful toolkit of methods for applying DP to GP models.

E1078: **Statistical disclosure control for functional PCA**
Presenter: **Ana Kenney**, Pennsylvania State University, United States
Co-authors: Jordan Awan, Matthew Reimherr, Aleksandra Slavkovic
Differential Privacy (DP) is a common and rigorous approach to quantify the disclosure risk of statistical procedures performed on sensitive data. Much work has been done on count and multivariate data, however the functional setting remains relatively unexplored even though large amounts of identifying information may be present. In functional data analysis (FDA), principal components are widely used for interpretation and as a dimension reduction technique for further study. It is therefore important to design a method that can still retain these properties while guaranteeing some level of DP. However, simply adding noise can result in very poor output due to the structure of principal components. In past literature, the exponential mechanism was utilized for PCA on multivariate data, and here we extend this approach to the functional setting. We demonstrate through a simulation study that the output of our method results in components comparable to the non-private ones for even moderate sample size.

E1094: **Sharing social network data: Differentially private estimation of exponential-family random graph models**
Presenter: **Vishesh Karwa**, Temple University, United States
Differential privacy has emerged as a powerful tool to reason rigorously about privacy and confidentiality issues. In its purest form, differential privacy limits direct access to raw data, allowing interaction only through a noisy interface. This requires new approaches to statistical inference. We will introduce the definition of differential privacy, followed by some of its key properties. We will present a framework for performing statistical inference under the constraint of differential privacy and its connections to measurement error and missing data models. The primary focus will be on sharing social network data for estimation of exponential random graph models. A case study using a version of the Enron e-mail corpus data-set demonstrates the application and usefulness of the proposed techniques in solving the challenging problem of maintaining privacy and supporting open access to network data to ensure reproducibility of existing studies and discovering new scientific insights. We use a simple yet effective randomized response mechanism to generate synthetic networks under edge differential privacy, and then use likelihood based inference for missing data and Markov chain Monte Carlo techniques to fit exponential-family random graph models to the generated synthetic networks.

E1181: **Differentially private significance tests for regression coefficients**
Presenter: **Andres Barrientos**, Duke University, United States
Co-authors: Jerome Reiter, Ashwin Machanavajjhala, Yan Chen
Many data producers seek to provide users access to confidential data without unduly compromising data subjects' privacy and confidentiality. One general strategy is to require users to do analyses without seeing the confidential data; for example, analysts only get access to synthetic data or query systems that provide disclosure-protected outputs of statistical models. With synthetic data or redacted outputs, the analyst never really knows how much to trust the resulting findings. In particular, if the user did the same analysis on the confidential data, would regression coefficients of interest be statistically significant or not? We present algorithms for assessing this question that satisfy differential privacy. We describe conditions under which the algorithms should give accurate answers about statistical significance. We illustrate the properties of the proposed methods using artificial and genuine data.

EO074 Room A1 ADVANCES IN STATISTICAL IMAGING Chair: Michele Guindani

E0444: **Spatial temporal analysis of multi-subject fMRI data**
Presenter: **Tingting Zhang**, University of Virginia, United States
Functional magnetic resonance imaging (fMRI) data analysis faces several challenges, including extensive computation and difficulty in obtaining statistically efficient estimates of the brain responses. We propose a new statistical model and computational algorithm to address these challenges. Specifically, we develop a new multi-subject, low-rank model within the general linear model framework for stimulus-evoked fMRI data. The new model assumes that the brain responses of different brain regions and subjects fall into a low-rank structure and can be represented by a few principal functional shapes. As such, the new model enables borrowing information across subjects and regions and increasing the ensuing estimation efficiency of brain responses, while accommodating the variation of brain activities across subjects, stimulus types, and regions. We propose two different optimization functions and a new fast-to-compute algorithm to address two research questions of broad interest in psychology studies: evaluating brain responses to different stimuli and identifying brain regions with different responses. Through both simulation and real data analysis, we show that the new method can outperform the existing methods by providing more efficient estimates of brain responses to designed stimuli.

E0503: **Robust and Gaussian spatial functional regression models for analysis of event-related potentials**
Presenter: **Jeff Morris**, MD Anderson Cancer Center, United States
Co-authors: Hongxiao Zhu, Philip Rausch
Event-related potentials (ERPs) are times series with both spatial correlation across electrodes and nested correlations within subjects. Commonly

used analytical methods focus on pre-determined extracted components and ignore the correlation among electrodes or subjects, which can miss important insights, and tend to be sensitive to outlying subjects, time points or electrodes. We introduce a Bayesian spatial functional regression framework that models the entire ERPs as spatially correlated functional responses and stimulus types as covariates, relying on mixed models to characterize stimuli effects while accounting for the multilevel correlation structure, including both Gaussian and more robust models using heavier-tailed likelihoods. The spatial correlation is captured through basis-space Materns that are separable or nonseparable over time. We induce both adaptive regularization over time and spatial smoothness across electrodes via a correlated normal-exponential-gamma prior. Our proposed analysis produces global tests for stimuli effects across entire time (or time-frequency) and electrode domains, plus multiplicity-adjusted pointwise inference based on EER or FDR to flag spatiotemporal (or spatio-temporal-frequency) regions that characterize stimuli differences, and can also produce inference for any prespecified waveform components. Our analysis of the smoking cessation ERP data set reveals numerous effects across different types of visual stimuli.

E0633: **Functional features in brain imaging**
Presenter: **Donatello Telesca**, UCLA, United States
Brain imaging techniques produce data which can be fruitfully interpreted as the realization of stochastic processes over functions of one or several evaluation domains. We propose a modeling approach for the identification of functional features, which combine to define the complete pattern observed for several statistical units. A probabilistic interpretation of the Karhunen-Loeve construction is merged with a prior on finite binary feature allocation matrices. The approach examined in the context of several case studies involving both benchmark datasets used in functional data analysis and brain imaging studies of developmental neurocognition.

E0736: **Bayesian image-on-image latent factor models for predicting task fMRI using task-free MRI**
Presenter: **Timothy Johnson**, University of Michigan, United States
Co-authors: Cui Guo, Jian Kang
Our brains show different activity during task performance across many behavioral domains. Prevailing thought was that individual differences in brain response were attributed to two factors: differences in gross brain morphology and differences in task strategy and/or cognitive processes. However, recent research suggests that these differences can attributed to task-free MRI. That is, individual differences in task-evoked brain activity are inherent features of individual brains such that can be predicted from task-free MRI. Their model is simply a linear regression of the z-score maps on the task-free MRI features. They fit one regression model for each of 50 parcels of the cortex for each individual. We set out to build a more sophisticated statistical model and compare results. We propose an image-on-image Bayesian latent factor regression model. We model the task-evoked maps via basis functions. The low-dimensional representation of the basis parameters is obtained using a sparse latent factor model. Then we use a scalar-on-image regression model to link the latent factors with the task-free maps which are selected using a Bayesian variable selection procedure. Head-to-head comparison shows that our modelling strategy is statistically more efficient than the simple model originally proposed.

E1026: **Statistical methods for modeling heritability of EEG connectivity**
Presenter: **Hernando Ombao**, King Abdullah University of Science and Technology (KAUST), Saudi Arabia
A number of recent studies have found evidence that characteristics of functional brain connectivity are significantly associated with various genetic markers, however, the majority of work in this area has been restricted to resting state fMRI data. We develop novel measures of connectivity that capture complex dependence structures and present new models that quantify heritability in EEG connectivity. We present the results from a novel study of EEG spectral-based connectivity measures during a working memory task from 350 healthy university students. Using recently developed statistical methods for testing associations between high-dimensional feature sets (which improves upon existing statistical methods through the use of non-Euclidean metrics), we identify specific sets of channels for which the coherence measures in the delta, theta, alpha, and beta frequency bands are significantly associated with a set of genetic markers previously implicated as risk factors for Alzheimer's disease. Additionally, we compare these heritability estimates with genome-wide heritability and with estimates from a set of neurotransmitter genes related to dopamine regulation. These results suggest that some genetic factors linked to Alzheimer's disease may also play a role in working memory performance in healthy individuals.

EO617 Room B1 GRAPHICAL MODELS IN THE LIFE SCIENCES Chair: Sofia Massa

E0944: **Extensions of graphical models with applications in genetics and genomics**
Presenter: **Pariya Behrouzi**, Wageningen University and Research, Netherlands
Co-authors: Ernst Wit
Several problems related to modeling complex systems are addressed. Fields such as systems genetics, systems biology, epidemiology, and bioinformatics often involve large-scale models in which thousands of components are linked in complex ways. What is perhaps most distinctive about the graphical model approach is its suitability in formulating probabilistic models of complex phenomena in applied fields, while maintaining control over the computational cost associated with these models. In real world, not all datasets are continuous. Discrete data or mixed discrete-and-continuous datasets routinely arise in above-mentioned fields. We introduce a method for reconstructing a conditional independence network from non-Gaussian data, in particular for ordinal and for mixed ordinal-and-continuous data. Such data are common in systems genetics, where the main focus is to understand the flow of biological information that underlies complex traits. We focus on the trait survival: we aim to find loci –locations on a genome– that do not segregate independently conditional on other loci. The network estimation relies on penalized Gaussian copula graphical models; this accounts for a large number of markers p and a small number of individuals n.

E0951: **Bayesian inference of high-dimensional graphical models: Application to brain connectivity**
Presenter: **Reza Mohammadi**, University of Amsterdam, Netherlands
In graphical models, Bayesian frameworks provide a straightforward tool, explicitly incorporating underlying graph uncertainty. In principle, the Bayesian approaches are based on averaging the posterior distributions of the quantity of interest, weighted by their posterior graph probabilities. However, Bayesian inference has not been used in practice for high-dimensional graphical models, because computing the posterior graph probabilities is hard and the number of possible graph models is very large. We discuss the computational problems related to Bayesian structure learning and we offer several solutions to cope with the high-dimensionality problems. We apply our method to high-dimensional fMRI data from brain connectivity studies to show its empirical usefulness. In addition, we have implemented our method in the R packages BDgraph and ssgraph which are available online.

E0979: **Causal inference with directed acyclic graphs: A case study in psychosis**
Presenter: **Giusi Moffa**, Institute of clinical epidemiology and biostatistics, University of Basel; and UCL Division of Psychiatry, Switzerland
Co-authors: Jack Kuipers
Directed acyclic graphs (DAGs) are common tools to describe causal mechanisms across different fields, ranging from social science to biology. Traditionally they have been used in forward causation to estimate the effects of causes given a postulated causal structure informed through domain experts. Thanks to computational progress in structure learning for Bayesian networks, DAGs have now also gained popularity in reverse causation.

Inferring the DAG structure from observational data allows us to gain insights about putative causal mechanisms, though only under very strict assumptions. Given the networks we learn from the data we can then derive putative intervention effects. However, to ensure robust inference it is essential to account for the uncertainty in the estimation of the DAG structure. This is now possible thanks to substantial advance in sampling of Bayesian networks from their posterior distribution given the data. As a result, we can follow a fully Bayesian approach to derive a posterior distribution of putative causal effects. We focus specifically on binary variables and present a case study in Psychosis. The method applies both to cross-sectional data as well as to longitudinal data, when we consider dynamic Bayesian networks.

E1101: **Bayesian analysis of multiple related molecular networks**
Presenter: **Gwenael Leday**, University of Cambridge, United Kingdom
Co-authors: Ilaria Speranza, Leonardo Bottolo, Sylvia Richardson
The problem of inferring and comparing multiple graphical structures from high-dimensional molecular data is considered. We propose a hierarchical Bayesian model that allows the borrowing of strength across groups of samples and the joint estimation of multiple (inverse) covariance matrices. Closed-form Bayes factors are then used to identify, say, common or group-specific structures via multiple testing. The proposed approach has the advantage of allowing directionality and the testing of biologically relevant hypotheses, such as edge losses and gains in a two-group comparison. It is also computationally very efficient, addressing problems with thousands of variables in a few seconds. We illustrate the proposed method on simulated data and various real data examples.

E1621: **Inferring networks from next generation sequencing data**
Presenter: **Thi Kim Hue Nguyen**, University of Padova, Italy
Co-authors: Monica Chiogna
MicroRNAs (miRNAs) have been reported to play a pivotal role in regulating key biological processes, for example, post-transcriptional modifications and translation processes. Some studies revealed that some disease-related miRNAs can indirectly regulate the function of other miRNAs associated with the same phenotype. Hence, studying the interaction pattern of miRNAs in some conditions might help understand complex phenotype conditions. Inferring the interaction pattern is a challenging task, as data measuring miRNA expression are usually high dimensional, discrete, possibly showing a large number of zeros and measured on a small number of units. From a technical point of view, the interactions among miRNA are well represented by a graph, where miRNAs and their connections are, respectively, nodes and edges. We propose a new algorithm for learning the structure of undirected graphs for count data, called PC-LPGM, and we prove its theoretical consistence in the limit of infinite observations. The proposed algorithm shows promising results when compared to some competitors using simulated data. Moreover, it provides biologically interpretable results when applied to real data downloaded from The Cancer Genome Atlas portal.

EO560 Room E1 STATISTICAL METHODS FOR NETWORKS AND INTEGRATIVE STUDIES Chair: Min Jin Ha

E0809: **An efficient sampling algorithm for network motif detection**
Presenter: **Yinghan Chen**, University of Nevada, Reno, United States
Co-authors: Yuguo Chen
Network motifs are substructures that appear significantly more often in a given network than in random networks. Motif detection is crucial for discovering new characteristics in biological, developmental, and social networks. We propose a sequential importance sampling strategy to estimate subgraph frequencies and detect network motifs. The method is developed by sampling subgraphs sequentially node by node using a carefully chosen proposal distribution. Viewing the subgraphs as rooted trees, we propose a recursive formula that approximates the number of subgraphs containing a particular node or set of nodes. The proposal used to sample nodes is proportional to this estimated number of subgraphs. The method generates subgraphs from a distribution close to uniform, and performs better than competing methods.

E0990: **Fused lasso regression for identifying differential correlations in brain connectome graphs**
Presenter: **Donghyeon Yu**, Inha University, Korea, South
A procedure is proposed to find differential edges between two graphs from high-dimensional data. We estimate two matrices of partial correlations and their differences by solving a penalized regression problem. We assume sparsity only on differences between two graphs, not graphs themselves. Thus, we impose an ℓ_2 penalty on partial correlations and an ℓ_1 penalty on their differences in the penalized regression problem. We apply the proposed procedure in finding differential functional connectivity between healthy individuals and Alzheimer's disease patients.

E1135: **Hierarchical structured component analysis for integrative analysis of multi-omics data**
Presenter: **Taesung Park**, Seoul National University, Korea, South
Co-authors: Yongkang Kim
Identification of multi-markers is one of most challenging issues in personalized medicine era. Nowadays, many different types of omics data are generated from the same subject. Although many studies have been developed to identify appropriate markers for each omics data, not many methods are available to identify integrated markers for various omics data. We propose a hierarchical structured component analysis of integrative multi-omics data. As an illustration, we consider miRNA-mRNA integration analysis. Many recent studies have shown that miRNAs are related to the pathogenesis of cancer and that miRNAs would be triggers of cancer initiation. It is well known that miRNAs affect phenotype only indirectly by regulating mRNA expression or protein translation. Although many researches have tried to use inhibition information of miRNAs to mRNAs, they could not use the information how much mRNA expression is regulated by miRNA to identify specific disease. Thus, we suggest an integration model which accounts for this biological relationship in the structured component and provides the integrated markers efficient. Through an application to pancreatic cancer data, our proposed model is shown to identify well the integrated markers of miRNA and mRNA for early diagnosis with better biological interpretation.

E0841: **Personalized integrated network modeling**
Presenter: **Min Jin Ha**, UT MD Anderson Cancer Center, United States
Co-authors: Sayantan Banerjee, Rehan Akbani, Han Liang, Gordon Mills, Kim-Anh Do, Veerabhadran Baladandayuthapani
Personalized (patient-specific) approaches have recently emerged with a precision medicine paradigm that acknowledges the fact that molecular pathway structures and activity might be considerably different within and across tumors. The functional cancer genome and proteome provide rich sources of information to identify patient-specific variations in signaling pathways and activities within and across tumors; however, current analytic methods lack the ability to exploit the diverse and multi-layered architecture of these complex biological networks. We assessed pan-cancer pathway activities across 32 tumor types from The Cancer Proteome Atlas by developing a personalized cancer-specific integrated network estimation (PRECISE) model. PRECISE is a general framework for integrating existing interaction databases, data-driven de novo causal structures, and upstream molecular profiling data to estimate cancer-specific integrated networks, infer patient-specific networks and elicit interpretable pathway-level signatures. PRECISE-based pathway signatures, can delineate pan-cancer commonalities and differences in proteomic network biology within and across tumors, demonstrates robust tumor stratification that is both biologically and clinically informative and superior prognostic power compared to existing approaches.

E1421: **iDINGO: Integrative differential network analysis in genomics**
Presenter: **Caleb Class**, UT MD Anderson Cancer Center, United States
Co-authors: Min Jin Ha, Veerabhadran Baladandayuthapani, Kim-Anh Do
Differential network analysis is an important way to understand the network rewiring involved in disease progression and development. Building differential networks from multiple 'omics data provides insight into the holistic differences of the interactive system under different patient-specific groups. DINGO was developed to infer group-specific dependencies and build differential networks. However, DINGO and other existing tools are limited to analyze data arising from a single platform, and modeling each of the multiple 'omics data independently does not account for the hierarchical structure of the data. We developed the iDINGO R package to estimate group-specific dependencies and make inferences on the integrative differential networks, considering the biological hierarchy among the platforms. A Shiny application has also been developed to facilitate easier analysis and visualization of results, including integrative differential networks and hub gene identification across platforms.

EO120 Room F1 CLUSTERING OF MULTIVARIATE DEPENDENT DATA Chair: F Marta L Di Lascio

E0620: **Hidden semi-Markov models with multivariate leptokurtic-normal components: Application to daily returns series**
Presenter: **Luca Bagnato**, Catholic University of the Sacred Heart, Italy
Co-authors: Antonio Punzo, Antonello Maruotti
The recently proposed multivariate leptokurtic-normal (MLN) distribution is a heavy-tailed generalization of the multivariate normal distribution with an additional parameter governing/denoting excess kurtosis. Advantageously with respect to other multivariate heavy-tailed elliptical distributions, the MLN is directly parametrized according to the moments of interest, i.e. the mean vector, the covariance matrix, and the excess kurtosis. With the aim of modelling the distributional and dynamic properties of daily returns, we consider the MLN as emission distribution to build hidden Markov and semi-Markov models. We outline an EM algorithm for maximum likelihood estimation which exploits recursions developed within the hidden (semi-)Markov literature. As an illustration, we provide an example based on the analysis of a bivariate time series of stock market returns.

E0942: **Model-based clustering of high dimensional data using copulas**
Presenter: **Marta Nai Ruscone**, LIUC, Italy
Finite mixtures are applied to perform model-based clustering of multivariate data. Existing models do not offer great flexibility for modelling the dependence of the data since they rely on potential undesirable correlation restrictions and strict assumptions on the marginal distribution. We proposed recently a model-based clustering method via R-vine copula that allows overcoming the previous restrictions by building flexible dependence models for an arbitrary number of variables using bivariate building blocks. This method shows a disappointing behavior in high-dimensional spaces since it leads to over-parametrized models. We propose a more parsimonious version of model-based clustering method via R-vine copula to alleviate the computational burden and the risk of overfitting. The model is based on the selection of the hyper-parameters of sparse model classes using truncated and thresholded R-vine copulas. We use simulated and real datasets to illustrate the proposed procedure.

E0245: **Non parametric frailty Cox model for clustering time-to-event data**
Presenter: **Anna Maria Paganoni**, MOX-Politecnico di Milano, Italy
Co-authors: Francesca Gasperoni, Francesca Ieva, Chris Jackson, Linda Sharples
An innovative model for hierarchical time-to-event data (i.e., healthcare data in which patients are grouped by healthcare providers) is described. The most popular model for this kind of data is the Cox proportional hazard model, with parametric frailties shared among patients belonging to the same group. We relax the parametric assumption on the frailty term by using a nonparametric discrete distribution with an unknown finite number of points in its support. Our aim is two-fold: on one hand, we want to propose a more flexible model for grouped survival data; on the other hand, we want to detect clusters of providers and characterize them through an a posteriori analysis supported by group specific covariates. A tailored expectation-maximization algorithm is introduced to estimate the number of clusters, the frailty discrete distribution, the proportion associated to each cluster and the classical parameters of a Cox model. To conclude, we show an application to a clinical administrative database, in which some information of patients suffering from heart failure is collected. We are able to detect a latent clustering structure among hospitals and this result has a clear impact both on patients' side and on hospital managements' side.

E1085: **Clustering of spatially dependent functional data**
Presenter: **Vincent Vandewalle**, Inria, France
Co-authors: Cristian Preda, Sophie Dabo
Two approaches for clustering spatial functional data are presented. The first one is the model-based clustering that uses the concept of density for functional random variables and logistic weights on the prior cluster probabilities depending on spatial coordinates. The second one is the hierarchical clustering based on univariate statistics for functional data such as the functional mode or the functional mean, and includes spatial weights in the distances computation. These two approaches take into account the spatial features of the functional data: two observations that are spatially close share a common distribution of the associated random variables. The two methodologies are illustrated by an application to air quality data.

E1138: **Comparing EM to a greedy search algorithm to optimize ICL for mixture models**
Presenter: **Arthur White**, Trinity College Dublin, Ireland
Co-authors: Jason Wyse, Gilles Celeux
The integrated complete-data likelihood (ICL) is a popular criterion in model-based clustering for choosing the number of clusters of a finite mixture model. Typically, the ICL is computed using a BIC-like approximation, which depends on maximum likelihood estimates that are found using the expectation-maximisation (EM) algorithm. Recently, an alternative method for clustering with the ICL has been introduced, that calculates the exact ICL in closed form within a Bayesian framework. A greedy search (GS) algorithm is then used to allocate observations to clusters in order to maximise the ICL directly and hence obtain an optimal clustering solution. This approach has the added benefit of simultaneously searching the model space. To better understand the properties of the GS method, we conducted an extensive simulation study comparing its performance to the standard EM approach, in terms of number of clusters selected, cluster accuracy, and computational cost. The performance of the methods on real data is also discussed.

EO574 Room G1 FLEXIBLE SURVIVAL METHODS Chair: Anneleen Verhasselt

E0518: **Goodness-of-fit tests in proportional hazards models with random effects**
Presenter: **Ingrid Van Keilegom**, KU Leuven, Belgium
Co-authors: Wenceslao Gonzalez-Manteiga, Lola Martinez-Miranda

The aim is to test the functional form of the covariate effects in a Cox proportional hazards model with random effects, like for instance a shared frailty model. We assume that the responses are clustered and incomplete due to right censoring. The estimation of the model under the null (parametric covariate effect) and the alternative (non-parametric effect) is performed using the full marginal likelihood. Under the alternative, the non-parametric covariate effects are estimated using orthogonal expansions. The test statistic is the likelihood ratio statistic, and its distribution is approximated using a bootstrap method. The performance of the proposed testing procedure is studied through simulations. The method is also applied on real data coming from a study on the chronic granulotomous disease.

E1379: **Joint model for bivariate zero inflated recurrent event data with terminal event**
Presenter: **Yang-Jin Kim**, Sookmyung Women University, Korea, South

Multivariate recurrent event arises when a subject has experienced several types of event repeatedly over time. These observations can be stopped by a terminal event which may be related with the recurrent event. Furthermore, there exist a substantial portion of subjects without any recurrent events during a fairly long follow-up time which results in a zero-inflated nature of data. For simultaneously considering both zero inflation and terminal event in a context of multivariate recurrent event data, a joint model is applied to model several dependencies through frailty effects; Correlation between recurrent events, one between recurrent event and terminal event and one between cure fraction and terminal event. Diverse simulation studies are performed to evaluate the suggested models. Infection data from AML (acute myeloid leukemia) patients are analyzed as an application.

E0808: **On the validity of time-dependent AUC estimation in the presence of cure fraction**
Presenter: **Anouar El Ghouch**, The University catholique de Louvain, Belgium
Co-authors: Abderrahim Oulhaj, Kassu Mehari Beyene

During the last decades, several approaches have been proposed to estimate the time-dependent area under the ROC curve (AUC) of risk tools derived from survival data. The validity of these estimators relies on some regularity assumptions among which a survival function being proper. In practice, this assumption is not always satisfied because a fraction of the population may not be susceptible to experience the event of interest even for long follow-up. Studying the sensitivity of the proposed estimators to the violation of this assumption is of substantial interest. We investigate the performance of the Li's estimator, a recently proposed estimator of the time-dependent AUC, when the population exhibits a cure fraction. Motivated from the current practice of deriving risk tools in cardiovascular disease, we also assess the loss, in terms of predictive performance, when deriving risk tools from survival models that do not acknowledge the presence of cure. The simulation results show that the Li's estimator is still valid even under the presence of cure. They also show that risk tools derived from survival models that ignore the presence of cure have smaller AUC compared to those derived from survival models that acknowledge the presence of cure.

E1091: **Parametric estimation of the association parameters in hierarchical survival data by nested Archimedean copula functions**
Presenter: **Mirza Nazmul Hasan**, Hasselt University, Belgium
Co-authors: Roel Braekers

There has been a growing interest in modeling hierarchical clustered multivariate survival data, which are possibly censored and/or missing. This type of data arise when a sample consists of clusters and each cluster has several, correlated sub-clusters contains various, dependent survival times, such that two layers of dependence occurs into the data-set. In the analysis of such survival times, two approaches are commonly used when we want to take the association between the survival times within a cluster and/or sub-cluster into account. A first approach is through frailty models while a second approach is by using copula models. A frailty model is a conditional model which assumes that different individuals within the same cluster are independent, conditionally on a common frailty term. In contrast, a copula model assumes that the joint survival function can be described by a copula function evaluated in the marginal survival functions of different individuals within a cluster. We use nested Archimedean copula functions to describe the dependency between different event times and investigate a one stage parametric estimation procedure for the association parameters of the models for hierarchical survival data, where both the clusters and sub-clusters are allowed to be moderate to large and varying in size. We perform a simulation study to check the finite sample properties of the estimators and also illustrate the method on a real life data-set.

E1698: **Modeling the future development of IBNR and RBNS claims in the presence of covariates**
Presenter: **Katrien Antonio**, University of Amsterdam and KU Leuven, Belgium
Co-authors: Jonas Crevecoeur, Roel Verbelen

Holding sufficient capital is essential for an insurance company to ensure its solvability. Predicting the amount of capital needed to fulfill future liabilities in an accurate way is an important actuarial task. Insurers record detailed information related to claims and policies for pricing insurance contracts. However, this same information is largely neglected when estimating the reserve. We present a flexible framework for including these claim specific covariates. Our framework focuses on three building blocks in the development process: the time to settlement, the number of payments and the size of each payment. We present a well-chosen generalized linear model (GLM) for each of these stochastic building blocks. Standard model selection techniques for GLMs allow us to determine the appropriate covariates in these models. We demonstrate how these covariates determine the granularity of our reserving model. On the one extreme, including many covariates, leads to large differences in the development process of individual claims. On the other extreme, including no covariates corresponds to specifying a model for data aggregated in a triangle. The set of selected covariates then naturally determines the position the actuary should take in between those two extremes.

EO146 Room H1 ADVANCES IN ORDINAL DATA ANALYSIS Chair: Cristina Mollica

E0326: **A goodness-of-fit test for the ordered stereotype model**
Presenter: **Daniel Fernandez**, Victoria University of Wellington, New Zealand

A new goodness-of-fit test is presented for an ordered stereotype model used for an ordinal response variable. The proposed test is based on the well-known Hosmer-Lemeshow test and its version of the proportional odds regression model. The latter test statistic is calculated from a grouping scheme assuming that the levels of the ordinal response are equally spaced which might be not true. One of the main advantages of the ordered stereotype model is that it allows us to determine a new uneven spacing of the ordinal response categories, dictated by the data. The proposed test takes the use of this new adjusted spacing to partition data. A simulation study shows good performance of the proposed test under a variety of scenarios. Finally, the results of the application are presented.

E0512: **Simultaneous clustering and dimensional reduction of mixed-type data**
Presenter: **Monia Ranalli**, University of Rome Tor Vergata, Italy
Co-authors: Roberto Rocci
In real applications, it is very common to have the true clustering structure masked by the presence of noise variables and/or dimensions. A mixture model is proposed for simultaneous clustering and dimensionality reduction of mixed-type data: the continuous and the ordinal variables are assumed to follow a Gaussian mixture model, where, as regards the ordinal variables, it is only partially observed. To recognize discriminative and noise dimensions, the variables are considered to be linear combinations of two independent sets of latent factors where only one contains the information about the cluster structure while the other one contains noise dimensions. In order to overcome computational issues, the parameter estimation is carried out through an EM-like algorithm maximizing a composite log-likelihood based on low-dimensional margins.

E0892: **Revealing subgroup structure in ranked data using a Bayesian WAND**
Presenter: **Daniel Henderson**, Newcastle University, United Kingdom
Ranked data arise in many areas of application ranging from the ranking of up-regulated genes for cancer to the ranking of academic statistics journals. Complications can arise when rankers do not report a full ranking of all entities; for example, they might only report their top-M ranked entities after seeing some or all entities. It can also be useful to know whether rankers are equally informative, and whether some entities are effectively judged to be exchangeable. We propose a flexible Bayesian nonparametric model for dealing with heterogeneous structure and ranker reliability in ranked data. The model is a Weighted Adapted Nested Dirichlet (WAND) process mixture of Plackett-Luce models and inference proceeds through a simple and efficient Gibbs sampling scheme for posterior sampling. The richness of information in the posterior distribution allows us to infer many details of the structure both between ranker groups and between entity groups (within ranker groups). The methodology is illustrated using several real data examples.

E1186: **A Bayesian Mallows approach to non-transitive pair comparison data: Application to sounds perception**
Presenter: **Marta Crispino**, Inria Grenoble, France
Co-authors: Valeria Vitelli, Arnoldo Frigessi, Elja Arjas, Natasha Barrett
The focus is on learning how listeners perceive sounds as having human origins. An experiment was performed with a series of electronically synthesized sounds, and listeners were asked to compare them in pairs. We propose a Bayesian probabilistic method to learn individual preferences from non-transitive pairwise comparison data, as happens when one (or more) individual preferences in the data contradicts what is implied by the others. We build a Bayesian Mallows model in order to handle non-transitive data, with a latent layer of uncertainty which captures the generation of preference misreporting. We then develop a mixture extension of the Mallows model, able to learn individual preferences in a heterogeneous population. The results of our analysis of the musicology experiment are of interest to electroacoustic composers and sound designers, and to the audio industry in general, whose aim is to understand how computer generated sounds can be produced in order to sound more human.

E1239: **Bayesian Mallows model for clicking data**
Presenter: **Qinghua Liu**, University of Oslo, Norway
Learning individual preferences from clicking data is an important step in order to make personal recommendations. One of the most popular approaches to personal recommendation is Collaborative Filtering (CF), which is based on a low rank matrix factorization technique. One important challenge of CF is the lack of reliable uncertainty quantification. We developed a Bayesian Mallows Model (BMM) to make inference from clicking data to make personal recommendations for each user. We treated clicking data as pairwise comparisons, and the method includes clustering of users and relies on Bayesian data augmentation. Recommendations are made based on posterior probabilities. We compare the accuracy of BMM with that of CF's.

EO236 Room I1 NEW ADVANCES ON STATISTICAL MODELING OF COMPLEX DATA I Chair: Mauricio Castro

E0322: **Multivariate-t nonlinear mixed models for censored multi-outcome longitudinal data**
Presenter: **Wan-Lun Wang**, Feng Chia University, Taiwan
Co-authors: Tsung-I Lin
In multivariate longitudinal studies, multi-outcome repeated measures on each subject over time may contain outliers, and the responses are often subject to an upper or lower limit of detection depending on the quantification assays. We consider an extension of the multivariate nonlinear mixed effects model by adopting a joint multivariate-t distribution for random effects and within-subject errors and taking the censoring information of multiple responses into account. The proposed model, called the multivariate-t nonlinear mixed model with censored responses (MtNLMMC), allows for analyzing multi-outcome longitudinal data exhibiting nonlinear growth patterns with censorship and fat-tailed behavior. Utilizing the Taylor-series linearization method, a pseudo-data version of expectation conditional maximization either (ECME) algorithm is developed for iteratively carrying out maximum likelihood estimation. We demonstrate our methods with HIV/AIDS data examples and simulation studies. Experimental results signify that the MtNLMMC performs favorably compared to its normal analogue and some existing approaches.

E0449: **Robust estimation in plant breeding: Evaluation using simulation and empirical data**
Presenter: **Vanda Lourenco**, Faculty of Sciences and Technology - New University of Lisbon, Portugal
Co-authors: Hans-Peter Piepho, Joseph O. Ogutu
Genomic prediction (GP) is used to determine the best genotypes for selection in plant breeding. Accurate estimation of predictive accuracy (PA) that measures the effectiveness of GP is thus of paramount importance for GP. Regression models are the models of choice for analyzing field data in plant breeding. However, when their underlying assumptions are violated, models that use the classical likelihood typically perform poorly, often resulting in biased parameter estimates. In plant genetics such biases usually result in inaccurate estimates of heritability (H) and predictive accuracy, and hence compromise the predictive performance of GP. Since phenotypic data are susceptible to contamination, improving the methods for estimating heritability and predictive accuracy can enhance the performance of GP. Robust statistical methods provide an intuitively appealing and a theoretically well justified framework for overcoming some of the drawbacks of classical regression. We introduce and evaluate the performance of a robust approach to two recently proposed methods from the literature for estimating heritability and predictive accuracy of GP against the classical approach through simulation under several plausible scenarios of data contamination. An example application to a rye dataset is presented and used to empirically assess the adequacy and usefulness of the robust approach.

E0950: **A Bayesian approach to differential recruitment with respondent-driven sampling data**
Presenter: **Isabelle Beaudry**, Pontificia Universidad Catolica de Chile, Chile
Co-authors: Krista Gile
Respondent-driven sampling (RDS) is a sampling mechanism that has proven very effective to sample hard-to-reach human populations connected through social networks. A small number of individuals typically known to the researcher are initially sampled and asked to recruit a small fixed number of their contacts who are also members of the target population. Each subsequent sampling waves are produced by peer recruitment until the desired sample size is achieved. However, the researcher's lack of control over the sampling process has posed several challenges to producing valid statistical inference from RDS data. For instance, participants are generally assumed to recruit completely at random among their contacts

despite the growing empirical evidence that suggests otherwise and the substantial sensitivity of most RDS estimators to this assumption. The main contributions are to parameterize an alternative recruitment behavior and propose a Bayesian estimator to correct for nonrandom recruitment.

E0986: **Bayesian nonparametric inference for the coefficient of overlap**
Presenter: **Vanda Inacio**, University of Edinburgh, United Kingdom
Co-authors: Javier Garrido Guillen, Maria Xose Rodriguez-Alvarez
Accurate diagnosis of disease is of fundamental importance in medical research and clinical practice. The major goal of a diagnostic test is to distinguish between diseased and nondiseased individuals and before a test is widely used in practice, its discriminatory ability must be rigorously assessed through statistical analysis. The overlap coefficient, which is defined as the proportion of overlap area between two density functions, has gained unarguably popularity as a summary measure of diagnostic accuracy. We propose a Bayesian nonparametric modelling framework, based on a combination of Dirichlet process mixtures and the Bayesian bootstrap, for the overlap coefficient. The performance of our methods is assessed through multiple simulation studies and an application to real data is provided.

E1102: **Bayesian hierarchical modeling of growth curve derivatives via sequences of quotient differences**
Presenter: **Garritt Page**, Brigham Young University, United States
Growth curve studies are typically conducted to evaluate differences among group or treatment-specific curves. Most analysis focus solely on the growth curves, but it has been argued that growth curve derivatives are able to highlight differences among groups that may be masked when considering the raw curves only. Motivated by the desire to estimate derivative curves hierarchically, we introduce a new sequence of quotient differences (empirical derivatives) which, among other things, are well behaved near the boundaries compared to other sequences in the literature. Using on the sequence of quotient differences, we develop a Bayesian method to estimate curve derivatives in a multi-level setting (a common scenario in growth studies) and show how the method can be used to estimate individual and group derivative curves and make comparisons. We apply the new methodology to data collected from a study conducted to explore the impact that radiation-based therapies have on growth in female children diagnosed with acute lymphoblastic leukemia.

EO362 Room L1 RECENT DEVELOPMENTS IN MULTIVARIATE DATA ANALYSIS Chair: Anne Ruiz-Gazen

E0427: **Statistical properties of second-order tensor decompositions**
Presenter: **Joni Virta**, Aalto University, Finland
Co-authors: Niko Lietzen, Klaus Nordhausen
Two classical tensor decompositions are considered from a statistical viewpoint: the Tucker decomposition and the higher order singular value decomposition (HOSVD). Both decompositions are shown to be consistent estimators of the parameters of a certain noisy latent variable model. The decompositions asymptotic properties allow comparisons between them. Also inference for the true latent dimension is discussed. The theory is illustrated with examples.

E0436: **Testing for principal component directions under weak identifiability**
Presenter: **Davy Paindaveine**, Universite libre de Bruxelles, Belgium
Co-authors: Julien Remy, Thomas Verdebout
The problem of testing is considered which is on the basis of a p-variate Gaussian random sample, the null hypothesis $H_0 : \theta_1 = \theta_1^0$ against the alternative $H_1 : \theta_1 \neq \theta_1^0$, where θ_1 is the "first" eigenvector of the underlying covariance matrix and θ_1^0 is a fixed unit p-vector. In the classical setup where eigenvalues $\lambda_1 > \lambda_2 \geq \ldots \geq \lambda_p$ are fixed, the likelihood ratio test (LRT) and the Le Cam optimal test for this problem are asymptotically equivalent under the null, hence also under sequences of contiguous alternatives. We show that this equivalence does not survive asymptotic scenarios where $\lambda_{n1} - \lambda_{n2} = o(r_n)$ with $r_n = O(1/\sqrt{n})$. For such scenarios, the Le Cam optimal test still asymptotically meets the nominal level constraint, whereas the LRT becomes extremely liberal. Consequently, the former test should be favored over the latter one whenever the two largest sample eigenvalues are close to each other. By relying on the Le Cam theory of asymptotic experiments, we study in the aforementioned asymptotic scenarios the non-null and optimality properties of the Le Cam optimal test and show that the null robustness of this test is not obtained at the expense of efficiency. Our asymptotic investigation is extensive in the sense that it allows r_n to converge to zero at an arbitrary rate.

E0492: **On estimating the number of signals in multivariate time series**
Presenter: **Klaus Nordhausen**, Vienna University of Technology, Austria
Co-authors: Joni Virta
A p-variate second order source separation (SOS) model is considered under the assumption that only $q < p$ source components exhibit serial autocorrelations and the remaining $p - q$ components can be considered noise. The goal is then to estimate the number of signals q and the signals itself. To solve the problem bootstrap and asymptotic hypothesis tests for the signal dimension as well as the ladle estimator are introduced. The tests and estimators are all based on the two SOS methods AMUSE and SOBI.

E0972: **Clustering and visualizing large cattle-trading networks using self-organizing maps**
Presenter: **Madalina Olteanu**, Pantheon-Sorbonne University, France
Co-authors: Kevin Pame, Gael Beaunee, Caroline Bidot, Elisabeta Vergu
Networks have drawn quite a burst of attention in the last years, and two of the related questionings are understanding the underlying structure(s) of the network and visualizing simplified version(s) of it. When one aims at bringing into light how the groups of entities in a graph are organized and how they interact, clustering and visualization were proven to be very useful. More particularly, the use of a recent relational version of the self-organizing maps (SOM) algorithm provides a unified tool for both purposes, while allowing for a wide range of alternatives in terms of assessing the similarity between vertices and/or edges. The aim is to adapt a bagged version of relational SOM for time-varying networks, and to explore the French cattle-trading network. The network is represented as a dynamical graph with a daily resolution level, where the vertices are the farms (and the commercial operators), and the edges are represented by the animal exchanges. The graph is directed (from sellers to buyers), weighted (by the number of exchanged animals), and time varying (a transaction occurs at a given time-instant). Furthermore, additional information on the vertices, such as the geographical situation, the type of farm, and so on, is used either during the clustering procedure or for validating the results.

E0455: **The R package sparsePCA for block approaches and group-sparse PCA**
Presenter: **Marie Chavent**, University of Bordeaux, France
Co-authors: Guy Chavent
Most of the algorithms developed in the recent years for sparse PCA aim at determining one single sparse principal component, and rely on the deflation process inherited from the unconstrained PCA when it comes to compute more than one sparse principal component. However, the use of the PCA deflation scheme in the sparse context where loadings and components are not necessarily orthogonal can lead to difficulties and joint optimization with respect to all loadings is expected to be more effective for variance maximization than sequential optimization. We will present a generalisation of the block sparse-ℓ_1 algorithm to the case where sparsity is required to hold on group of variables rather than on the individual variables. We will also present the R package sparsePCA (github.com/chavent/sparsePCA) implementing block and deflation approaches for sparse

and group-space PCA. We will then compare numerically the performance of block and deflation approaches for group-sparse PCA on simulated synthetic data and illustrate the influence of the group information on the retrieval of the sparsity pattern.

EO154 Room M1 RECENT ADVANCES IN FUNCTIONAL AND MULTIVARIATE DATA ANALYSIS Chair: Yuko Araki

E0293: **On the number of principal components in high dimensions**
Presenter: **Sungkyu Jung**, Seoul National University, Korea, South
Modern big data challenges suggest investigation of growing dimension, with limited sample size. While the high dimension, low sample size asymptotics has been a powerful tool in understanding the success and failure of some linear multivariate methods, current tools also exhibit limitations. We will discuss some of the limitations and potential solutions. In particular, the problem of how many components to retain in the application of principal component analysis when the dimension is much higher than the number of observations will be discussed in detail. The proposed estimation strategy for the number of components is to sequentially test skewness of the squared lengths of residual scores that are obtained by removing leading principal components. The residual lengths are asymptotically left-skewed if all principal components with diverging variances are removed, and right-skewed if not. Some asymptotic properties of the proposed estimator will be discussed. Specifically, the estimator is shown to be consistent. The proposed estimator performs well in high-dimensional simulation studies, and provides reasonable estimates in a number of real data examples.

E0648: **Sparse principal component regression via singular value decomposition**
Presenter: **Shuichi Kawano**, The University of Electro-Communications, Japan
Principal component regression (PCR) is a two-stage procedure. The first stage is principal component analysis (PCA). The second stage is regression in which the obtained principal components are regarded as new explanatory variables. Since PCA is based on the explanatory variables, the principal components have no information on the response variable. To address this problem, we propose a one-stage procedure for PCR based on singular value decomposition. The loss function consists of a combination of the regression loss and PCA loss with sparse regularization. The proposed method enables us to obtain principal component loadings that are related to both explanatory variables and a response variable. We conduct numerical studies to examine the effectiveness of the proposed method.

E0938: **Inference on active domains of functional data via functional linear regression**
Presenter: **Masaaki Imaizumi**, Institute of Statistical Mathematics, Japan
Co-authors: Kengo Kato
An inference method on active domains of functional data is developed via a functional linear regression and a principal component analysis (PCA) based estimator. In a linear regression model with a functional covariate and a scalar response variable, an active domain of a functional covariate is defined as a subset of a domain on which a functional data has a positive effect on outputs. Based on a functional linear regression model, an active domain is regarded as a level set of a slope function of the regression model. We propose an estimator for an active set by combining the PCA-based estimator and a kernel convolution approach. Also, we provide a multiplier bootstrap method for confidence analysis for an active set based on the high-dimensional Gaussian approximation technique. Our confidence analysis is shown to be valid asymptotically with ordinal conditions for a PCA-based estimator. We also propose a practical selection method for hyperparameters such as a cut-off level for basis functions and kernel width. The experimental analysis supports the validity of our method.

E1099: **Functional classification with direct and indirect effects for high dimensional data**
Presenter: **Yuko Araki**, Shizuoka University, Japan
Recent years have seen that functional data analysis are capable of extracting intrinsic features from recently arising complicated and high dimensional data, such as three dimensional brain sMRI, hundreds of records of human gait, or traffic flow data for example. We introduce statistical methods for solving a classification problem which contains complex associations among several variables including high dimensional intermediate variables. The proposed method is based on composite basis function, which is an extended version of basis expansions with the help of sparse PCA. Further, L_1-type penalty constraints are imposed in estimation. This two-step regularization method accomplishes both covariates selection and estimation of unknown model parameters simultaneously. The crucial issue is how to select the regularization parameters used in model estimation. We propose a model selection method based information criterion. The proposed models are evaluated through Monte Carlo simulations to examine the efficiency of our modeling strategies.

EO206 Room N1 EXTREME VALUES Chair: Katharina Hees

E1148: **Estimation of the spectral measure from convex combinations of jointly regularly varying random variables**
Presenter: **Marco Oesting**, University of Siegen, Germany
Co-authors: Olivier Wintenberger
The extremal dependence structure of a regularly varying random vector X is fully described by its limiting spectral measure. We investigate how to recover characteristics of the measure, such as extremal coefficients, from the extremal behaviour of convex combinations of components of X. Our considerations result in a class of new estimators of moments of the corresponding combinations for the spectral vector. We show asymptotic normality and discuss the optimization of the asymptotic variance.

E0699: **Ordinal patterns in clusters of extremes of regularly varying time series**
Presenter: **Alexander Schnurr**, University Siegen, Germany
Co-authors: Marco Oesting
The purpose is to investigate temporal clusters of extremes defined as subsequent exceedances of high thresholds in a stationary time series. Two meaningful features of these clusters are the probability distribution of the cluster size and the ordinal patterns within a cluster. The latter have been introduced in order to handle data sets with several thousand data points appearing in medicine, biology, finance and computer science. Since these patterns take only the ordinal structure of consecutive data points into account, the method is robust under monotone transformations and measurement errors. We verify the existence of the corresponding limit distributions in the framework of regularly varying time series, develop non-parametric estimators and show and their asymptotic normality under appropriate mixing conditions.

E1644: **Ordinal pattern dependence in contrast to other concepts of dependence**
Presenter: **Ines Muenker**, University of Siegen, Germany
Co-authors: Alexander Schnurr
Recently, the concept of ordinal pattern dependence was introduced. It is an intuitive measure for the degree of co-monotonic behavior between two time series. The question has been raised what the connection between this concept and other dependence measures (like correlation, Kendall's tau and Spearman's rho) is. We give a partial answer to this question by an empirical study of various types of data sets. The key result is that the new method is suitable to measure dependence and it is much more robust against shocks and/or structural breaks than other dependence concepts. Finally, we shed some light on the applications of ordinal pattern dependence in the context of extremal events.

E0692: **Extreme value theory for bursty time series**
Presenter: **Katharina Hees**, TU Dortmund University, Germany
In many complex systems, inter-arrival times between events such as solar flares, trades, neuron voltages or earthquakes follow a heavy-tailed distribution. The set of event times is fractal-like, being dense in some time windows and empty in others, a phenomenon which has been dubbed "bursty". The return times of the extremes, or more precisely of the exceedances above a high threshold, are then no longer exponential distributed and this results in a serial clustering of the extreme events. Such a behavior was also observed for midlatitude cyclones. The aim is to model extreme events of such a bursty time series with heavy tailed inter-arrival times. For high thresholds and infinite mean waiting times, we show that the times between threshold crossings are Mittag-Leffler distributed, and thus form a fractional Poisson-process, which generalizes the standard Poisson-process. We provide graphical means of estimating model parameters and assessing model fit. Along the way, we apply our inference method to a real-world time series, and show how the memory of the Mittag-Leffler distribution affects the predictive distribution for the time until the next extreme event.

E1493: **Threshold selection in univariate extreme value analysis**
Presenter: **Laura Fee Schneider**, University of Goettingen, Germany
Co-authors: Andrea Krajina, Tatyana Krivobokova
Threshold selection plays a key role for various aspects of statistical inference of rare events. Most classical approaches tackling this problem for heavy-tailed distributions crucially depend on tuning parameters or critical values to be chosen by the practitioner. To simplify the use of automated data-driven threshold selection methods we introduce two new procedures not requiring the choice of any parameters. The first method measures the deviation of the log-spacings from the exponential distribution and works well for estimating high quantiles. The second approach estimates the asymptotic MSE of the Hill estimator unbiasedly if $\rho = -1$, and we illustrate that the approach still performs well for other values of the second order parameter.

EO142 Room O2 ECOSTA JOURNAL: COPULAS Chair: Takeshi Emura

E1069: **Copula additive regression models with endogenous binary treatment and count response**
Presenter: **Giampiero Marra**, University College London, United Kingdom
Co-authors: Rosalba Radice, David Zimmer
Copula regression models are discussed for a count response and an endogenous binary treatment, where the marginals and copula function can be chosen from a rich set of distributions and all the models parameters can be flexibly specified as functions of additive predictors. Estimation is achieved using a simultaneous penalised likelihood approach with automatic multiple smoothing parameter selection. Inferential results are also briefly discussed. The modelling framework is implemented in the R package GJRM (Generalised Joint Regression Models). The approach is illustrated on a case study which investigates the effect of insurance status (a binary measure) on doctor visits (a count measure).

C0420: **A nested copula duration model for competing risks with multiple spells**
Presenter: **Ralf Wilke**, Copenhagen Business School, Denmark
Co-authors: Enno Mammen, Simon Lo
A copula graphic estimator for the competing risks duration model with multiple spells is presented. It is a general nested copula model that allows for different degrees of dependence between risks and spells and therefore breaks up an implicit restriction of popular duration models such as multivariate mixed proportional hazards. It is shown that the dependence structure between spells is identified and can be estimated, in contrast to the dependence structure between competing risks. Thus, by allowing these two components to differ, the model is not identified. This is an important finding related to the general identifiability of competing risks models. Various features of the model are investigated by simulations and its practicality is illustrated by an application to unemployment duration data.

E0544: **On the goodness of standard copulas**
Presenter: **Dragan Radulovic**, Florida Atlantic University, United States
A series of simulated and numerical examples demonstrating that, more often than not, standard model copulas do not capture the underlying dependency structure are shown. We believe that copula models, unlike other statistical tools, are too readily accepted by practitioners. Rigorous, goodness of fit tests are commonly replaced by off hand statements like: it works well. To this end, we offer a theoretical result, an umbrella type theorem tailored for creating numerous Goodness of Fit tests for copulas.

E0638: **Search of a vine structure in vine copula model based on sampling order proximity**
Presenter: **Dorota Kurowicka**, Delft University of Technology, Netherlands
Vine copulas become recently very popular in modelling continuous distributions with complicated dependence structures. In this model, a joint density of random vector $(X_1,...,X_d)$ is specified by the product of marginal distributions and $(d-1)d/2$ (un)conditional bivariate copulas. There are exponentially many decompositions of a density into these bivariate building blocks, and theoretically all these vine structures are equivalent. In practice, however, when copulas are fitted to data sequentially level by level of the vine structure and conditional copulas are assumed not to depend directly on the conditioning variables some vine structures constitute a better model of the data than the others. A heuristic search of the best vine structure has been briefly introduced recently. It has been observed that for two vine structures, common sampling orders consistent with these vine structures, give an indication of how similar (how many repeated bivariate copulas they have in the decomposition) these vines are. We present a thorough evaluation of the heuristic based on sampling order proximity. We present an algorithm to find all vine structures with the given number of sampling orders in common and show results of an extensive simulation study of a heuristic search based on sampling order proximity.

E0246: **Diagonal distributions**
Presenter: **Miguel de Carvalho**, School of Mathematics, The University of Edinburgh, United Kingdom
Co-authors: Manuele Leonelli, Rodrigo Rubio
Diagonal distributions are introduced as an extension of marginal distributions. The main diagonal will be examined in detail, which consists of a mean-constrained univariate distribution on $[0,1]$, that summarizes key features on the dependence structure, and whose variance connects to Spearman's rho. We will comment on D-dimensional extensions, and on the fact that in the case of independence there is a connection with the so-called Irwin–Hall distribution. Mean-constrained histograms and smoothing methods are developed so to learn about diagonal distributions. Real and simulated data will be used in order to illustrate the key concepts and methods.

EO518 Room Q2 BAYESIAN SEMI- AND NON-PARAMETRIC MODELLING Chair: Silvia Liverani

E0308: **A noisy MCMC sampler for latent position network models**
Presenter: **Riccardo Rastelli**, University College Dublin, Ireland
Co-authors: Florian Maire, Nial Friel

Latent position models are widely used for the statistical analysis of networks in a variety of research fields. In fact, these models possess a number of desirable theoretical properties, and are particularly easy to interpret. However, algorithms that can fit these models generally require a computational cost which grows with the square of the number of nodes in the graph. This makes the analysis of large social networks impractical. We will show a new algorithm characterized by a linear computational complexity, which may be used to fit latent position models on networks of several tens of thousands nodes. The approach relies on an approximation of the likelihood function, where the amount of noise introduced can be arbitrarily reduced at the expense of computational efficiency. We will illustrate some theoretical results that show how the likelihood error propagates to the invariant distribution of the Markov chain Monte Carlo sampler. Finally, we will show some applications of the method to simulated networks and to a large network of co-authorships, demonstrating that one can achieve a substantial reduction in computing time and still obtain a reasonably good estimation of the latent structure.

E0416: **Quantification of the uncertainty of a partition coming from the Dirichlet process mixture model**
Presenter: **Aurore Lavigne**, University of Lille, France
Co-authors: Silvia Liverani

Results on the quantification of the uncertainty link to partition obtained from a Dirichlet process mixture model (DPMM) are presented. This model is popular for model-based clustering under the Bayesian framework, and is used in numerous fields (machine learning, epidemiology, genetic). In the DPMM, the Dirichlet process is assigned as prior of the mixture distribution, that allows to not specify the expected number of mixture components. Moreover, numerous inference methods are now well established. However, the extraction of a unique partition from the partitions sampled in their posterior distributions is a sensitive task. Numerous methods are proposed, but in practice, they lead to partitions which may turn out to be very different, making the interpretation difficult. We propose a method to quantify the uncertainty of a partition regarded as "optimal". The approach is based on an analogy with finite mixture models. We break down the predictive distribution into a mixture of densities and the weights are proportional to the cluster size of the "optimal" partition. We show that the densities are not parametric and that they do not depend only on observations coming from the class they represent. Finally, we propose a diagram in order to show for each observation, its probability of being clustered in each class of the "optimal" partition.

E0509: **Colombian women's life choices: A Bayesian nonparametric multivariate regression approach**
Presenter: **Isadora Antoniano-Villalobos**, Bocconi University, Italy
Co-authors: Andrea Cremaschi, Raffaella Piccarreta, Sara Wade

Women in the Latin America and Caribbean countries face difficulties related to the patriarchal traits of their society. In Colombia, the well-known conflict afflicting the country since 1948, has increased the risk of vulnerable groups. It is important to determine if recent efforts to improve the welfare of women have had a positive effect extending beyond the capital, Bogota. In an initial effort to shed life on this matter, we analyze cross-sectional data arising from the Demographic and Health Survey Program which collects and disseminates data on random samples of households selected from a national sampling frame. The aim is to study the relationship between baseline socio-demographic factors and variables associated to fertility, partnership patterns and work activity. We propose a flexible Bayesian nonparametric multivariate regression model, which can capture nonlinear regression functions and the presence of non-normal errors, such as heavy tails or multi-modality. The model has interpretable covariate-dependent weights constructed through normalization, allowing for combinations of both categorical and continuous covariates, as well as censoring in one or more of the responses. Computational difficulties for inference are overcome through an adaptive truncation algorithm combining adaptive Metropolis-Hastings and sequential Monte Carlo to create a sequence of automatically truncated posterior mixtures.

E0748: **The Bayes Lepski's method and credible bands through volume of tubular neighborhoods**
Presenter: **William Weimin Yoo**, Queen Mary University of London, United Kingdom
Co-authors: Aad van der Vaart

For a general class of priors based on random series basis expansion, we develop the Bayes Lepski's method to estimate unknown regression function. In this approach, the series truncation point is determined based on a stopping rule that balances the posterior mean bias and the posterior standard deviation. Equipped with this mechanism, we present a method to construct adaptive Bayesian credible bands, where this statistical task is reformulated into a problem in geometry, and the band's radius is computed based on finding the volume of certain tubular neighborhood embedded on a unit sphere. We consider two special cases involving B-splines and wavelets, and discuss some interesting consequences such as the uncertainty principle and self-similarity. Lastly, we show how to program the Bayes Lepski's stopping rule on a computer, and numerical simulations in conjunction with our theoretical investigations concur that this is a promising Bayesian uncertainty quantification procedure.

E1089: **Semi-supervised multi-view Bayesian nonparametric clustering for integrative genomics**
Presenter: **Paul Kirk**, University of Cambridge, United Kingdom

Although the challenges presented by high dimensional data in the context of regression are well-known and the subject of much current research, comparatively little work has been done on this in the context of clustering. In this setting, the key challenge is that often only a small subset of the covariates provides a relevant stratification of the population. Identifying relevant strata can be particularly challenging when dealing with high-dimensional datasets, in which there may be many covariates that provide no information whatsoever about population structure, or - perhaps worse - in which there may be (potentially large) covariate subsets that define irrelevant stratifications. For example, when dealing with genetic data, there may be some genetic variants that allow us to group patients in terms of disease risk, but others that would provide completely irrelevant stratifications (e.g. which would group patients together on the basis of eye or hair colour). Bayesian profile regression is a semi-supervised model-based clustering approach that makes use of a response in order to guide the clustering toward relevant stratifications. Here we consider how this approach can be extended to the "multiview" setting, in which different groups of covariates ("views") define different stratifications. We present some results in the context of cancer subtyping to illustrate how the approach can be used to perform integrative clustering of multiple 'omics datasets.

EC641 Room Aula 4 CONTRIBUTIONS IN COMPUTATIONAL STATISTICS Chair: Yiming Ying

E0198: **Parallel Gibbs variable selection for high-dimensional generalized linear models**
Presenter: **Guangbao Guo**, Shandong University of Technology, China

A novel parallel Gibbs variable selection procedure is proposed for high-dimensional generalized linear models. In the procedure, the data are randomly split into some subsets according to the given rules. We propose a series of weights to obtain optimized stationary distributions. Through the Gibbs method, we can quickly select effective parallel group variables. In aspect of theoretical properties, we obtain convergence of the method.

E1550: **A graph approach to find the best grouping for each possible number of clusters**
Presenter: **Cristian Gatu**, Alexandru Ioan Cuza University of Iasi, Romania
Co-authors: Cornel Barna, Ana Colubi, Erricos John Kontoghiorghes

The unsupervised non-hierarchical clustering of a set of n data points is a NP-problem. A theoretical approach to solve this problem is proposed. Specifically, a graph structure which can be employed to enumerate and evaluate all possibilities to cluster a number of observations is introduced. Any complete traversal of the graph generates all possible clustering solutions. The structure of the graph is exploited in order to design an efficient branch-and-bound algorithm that finds the optimal clustering solution without traversing the whole graph. A heuristic version of the branch-and-bound algorithm that reduces the execution time at the expense of the solution quality is also presented. In addition, a p-combination method that considers at each level of the graph not all, but only the best p groupings is also investigated. The Ward method is a special case for $p = 1$. This allows for trading between exploration and computational efficiency. Experimental results are presented and analyzed. The new theoretical solution gives an insight to the clustering problem that could be the foundation for further developments on clustering related problems.

E1382: **Estimating a Poisson autoregressive model with the backfitting algorithm**
Presenter: **Paolo Victor Redondo**, University of the Philippines Diliman, Philippines
Co-authors: Erniel Barrios, Joseph Ryan Lansangan

A Poisson autoregressive model (PAR) that accounts for discreteness and autocorrelation of count time series data is typically estimated within the context of state-space modelling with maximum likelihood estimation (MLE). The complexity of dependencies exhibited by count time series data however, complicates MLE. PAR is viewed as an additive model and is estimated using a hybrid of conditional least squares and MLE in the backfitting framework. Simulation studies show that estimation of PAR model viewed as an additive model is always better than PAR model in the state-space context whenever the non-normality of covariates for the latter is evident. In cases where the MLE of the PAR model in the state-space context exists, the estimates are comparable with the proposed method. The proposed method is then used in modelling incidence of tuberculosis, elucidating the role of various stakeholders in curbing the prevalence rate of the disease.

E1459: **Constrained matrix completion algorithm considering individual differences**
Presenter: **Yuki Morioka**, Doshisha University, Japan
Co-authors: Kensuke Tanioka, Hiroshi Yadohisa

The matrix completion problem has attracted considerable attention for recommendation systems, largely through the famous Netflix competition. Recently, many matrix completion methods have been proposed and evaluated in terms of estimation accuracy or calculation speed. Various matrix completion methods for recommendation systems are applied to data where the row, column, and value indicate the user, item, and evaluation, respectively. However, the existing methods have a limitation in that the estimation accuracy is low for data where individual user evaluations tend to be lower or higher because these methods do not consider individual differences among users. Moreover, if the data consists of different scales, such as a nominal scale or ordered scale, it is difficult to evaluate the information and estimate values correctly. Therefore, we propose a novel matrix completion method that considers users individual differences and mixed scales of user's external information as dummy variables. The proposed method estimates each parameter using biased inductive matrix completion. We evaluate the estimation accuracy by conducting numerical experiments including a simulation study and real data analysis.

E1508: **An attention algorithm for solving large scale structured l_0-norm penalized estimation problems**
Presenter: **Tso-Jung Yen**, Academia Sinica, Taiwan
Co-authors: Yu-Min Yen

Technology advances have enabled researchers to collect large amounts of data with lots of covariates. Because of the high volume (large n) and high variety (large p) properties, model estimation with such kind of big data has posed great challenges for statisticians. We focus on the algorithmic aspect of these challenges. We propose a numerical procedure for solving large scale regression estimation problems involving a structured l_0-norm penalty function. This numerical procedure blends the ideas of randomization, proximal operators and blockwise coordinate descent algorithms. In particular, it adopts an attention-based sampling distribution for picking up regression coefficients for updates based on a closed form representation of the proximal operator of the structured l_0-norm penalty function. Simulation study shows the proposed numerical procedure is competitive to the benchmark algorithm for sparse estimation in terms of runtime and statistical accuracy when both the sample size and the number of covariates become large.

EC633 Room O1 CONTRIBUTIONS IN TIME SERIES II Chair: Stephen Pollock

E1565: **Dimension reduction for time series in a BSS context using R**
Presenter: **Markus Matilainen**, University of Turku/Turku PET Centre, Finland
Co-authors: Klaus Nordhausen, Jari Miettinen, Joni Virta, Sara Taskinen

Multivariate time series observations are increasingly common in multiple fields of science but the complex dependencies of such data often translate into intractable models with large number of parameters. An alternative is given by first reducing the dimension of the series and then modelling the resulting uncorrelated signals univariately, avoiding the need for any covariance parameters. A popular and effective framework for this is blind source separation. We present dimension reduction tools for time series available in the R package tsBSS. These include methods for estimating the signal dimension of second-order stationary time series, dimension reduction techniques for stochastic volatility models and supervised dimension reduction tools for time series regression. Examples are provided to illustrate the functionality of the package.

E1668: **A contribution to forecast time series with structural break**
Presenter: **M Rosario Ramos**, FCiencias.ID, Portugal
Co-authors: Clara Cordeiro

The presence of structural instability affects the estimation, inference and prediction. In this case, for a time series $Y = y_1, \cdots, y_n$, a structural change exists at a unknown time t if $y_1, \cdots, y_t$ differ from $Y^* = y_{t+1}, \cdots, y_n$, namely in the trend. This has an impact on the study of economic, environmental, climatic and other variables, since the forecast can be strongly biased. The aim is to contribute for the improvement of forecasts in this scenario. Given a time series Y, the first step is the estimation and removing of the seasonality based on the seasonal-trend decomposition by Loess (STL). The selection of the best STL fit was performed by the algorithm stl.fit, which runs all the possible combinations of the smoothing parameters and in the end, find the optimal parameters combination which minimized an accuracy measure. Secondly, the detention of a structural break in the seasonally adjusted time series is performed by the R package strucchange. Thirdly, based on the time series Y^*, fit the best model and obtain forecasts. A comparison between the forecasts of Y and Y^* is presented and the performance is evaluated using real data sets.

C1652: **A spatio-temporal GARCH model**
Presenter: **Hans Arnfinn Karlsen**, University of Bergen, Norway
Co-authors: Sondre Holleland

A spatio-temporal GARCH model is considered on an infinite spatial temporal grid. Due to the time dimension of the model it possible to extend the time series GARCH model to a spatial setting. By using this approach we obtain conditions for the existence of a stationary ergodic solution of

the spatio-temporal GARCH model. We also see that the squared variables of the solution satisfy a spatial temporal ARMA model. For estimation of the unknown parameter vector a conditional likelihood method, as used in the time series case, suffers from a boundary that increases with sample size. However, the problem is solved by maximizing a modified likelihood function. This gives both consistent estimates and asymptotic normality. Extension to a spatial temporal ARMA model with GARCH residuals is also discussed.

E1434: **Comparative mean value with an application to personal financial data analysis**
Presenter: **Andrej Svetlosak**, University of Edinburgh, United Kingdom
Co-authors: Miguel de Carvalho, Raffaella Calabrese
A statistical method is proposed for comparing a reference subject to a group of peers, where peers are selected based on similarity of values of a comparison variable. The proposed methodology is motivated by the need of comparing personal finances of a reference subject to subjects with similar characteristics. Comparative mean values are introduced so to identify if behaviour of an individual varies from the average of their peers. The similarity can be controlled by a parameter. We analyse the properties of the construct in respect to this parameter. First, we introduce our construct, the comparative mean value, in a time –invariant setting. A time –varying version is developed by taking advantage of non –parametric regression methods. The construct is tested in a simulation study. Results support an overall good performance of the methods. We then apply the proposed method to data provided by a major financial service provider in the UK, Money Dashboard. The dataset consists of 10,689 customers of a financial services provider and contains information on customer accounts from 49 financial institutions in the UK. We showcase how the time –varying comparative mean expenditure can be used to identify the spending behaviour of an individual in comparison to their peers.

E1542: **Alternative methods of seasonal adjustment**
Presenter: **Stephen Pollock**, University of Leicester, United Kingdom
The conventional methods of seasonal adjustment rely on filters realized in the time domain that nullify the sinusoidal elements of a data sequence that are to be found at the seasonal frequencies at its harmonics. Such filters combine a low pass filter that attenuates the high frequency elements of the data with a so-called comb filter that eliminates the seasonal frequency and its harmonics. Typically, the filters are derived from the estimates of a structural time-series model or of a reduced-form ARMA model. Often such filters fail fully to nullify data components that are adjacent to the seasonal frequencies, which serve to modulate the patterns of seasonal variation. They may leave a residue of the seasonal fluctuations in the filtered data. The aim is to analyze the effects of the common methods of the seasonal adjustment and to propose alternative methods that operate in the frequency domain and that enable a careful choice to be made of the seasonal elements that should be eliminated from the data.

EG385 Room Aula A CONTRIBUTIONS IN STOCHASTIC PROCESSES Chair: Hiroki Masuda

E1438: **Noise estimation for ergodic Levy driven stochastic differential equation model**
Presenter: **Yuma Uehara**, The Institute of Statistical Mathematics, Japan
Co-authors: Hiroki Masuda
To describe non-Gaussian activity in high frequency data obtained from financial, biological, and technological phenomenon, Levy driven stochastic differential equation model is plausible and used in various fields. However, the closed form of its genuine likelihood is not generally given, and thus the information of its driving noise (Levy process) is difficult to be estimated from observed data. To solve such a problem, we propose a new method based on Euler residuals constructed by Gaussian quasi-likelihood estimator: we approximate unit time increments of the driving noise by summing up the corresponding Euler residuals, and making use of them, we can conduct parametric estimation methods of the driving noise with bias correction. We will also present numerical experiments to see the performance of our methods.

E1450: **Consistent model selection for ergodic SDEs**
Presenter: **Shoichi Eguchi**, Osaka University, Japan
Co-authors: Yuma Uehara
There are several studies of model selection for stochastic differential equations (SDEs), for example, the contrast-based information criterion for ergodic diffusion processes and the Schwarz type information criterion for locally asymptotically quadratic models. However, most of the existing theoretical literature has been developed for nested models. We will give the mathematical validity of Bayesian model comparison for possibly misspecified ergodic SDE models and propose the quasi-Bayesian information criterion (QBIC).

E1568: **A small sample analysis of discretely observed diffusion processes**
Presenter: **Giuseppina Albano**, University of Salerno, Italy
Co-authors: Michele La Rocca, Cira Perna
Diffusion processes are commonly used to model stochastic phenomena, such as dynamics of financial securities and short-term loan rates. Several methods for the inference have been proposed, essentially based on MLE or its generalizations. Numerical approximations to the unknown likelihood function also lead to efficient estimators. We consider two well-known processes, Vasicek and CIR models. Sample properties of MLE estimators for the involved parameters of such processes are known when the sample size tends to infinity. Moreover, bootstrap procedures to reduce the bias of the drift estimates can be successfully applied. Other methods also lead to estimators that seem to work well in an asymptotic regime. Anyway, in many applications, data are yearly or quarterly observed, so in the estimation of the involved parameters the asymptotic condition of the sample size means to observe the phenomenon for a long period and likely the time series present structural breaks. This is the case in which Vasicek and CIR models are used in insurance for the valuation of life insurance contracts or also to model short-term interest rates. We focus on small sample properties of some alternative estimators. We consider time series with a length between 10 and 100, typically values observed in these contexts. We perform a simulation study in order to investigate which properties of the parameter estimator still remain valid.

E1705: **Stochastic differential equation modeling of social influence in networks**
Presenter: **Nynke Niezink**, Carnegie Mellon University, United States
People, organizations and countries are examples of social actors, operating within networks of interdependencies. The attributes of social actors, such as physical, psychological or performance measures, can be affected by the actors to whom they are connected, a process generally known as social contagion or social influence. We present a new methodology for the estimation of social influence effects on static networks, using stochastic differential equation modeling. To estimate the model, we propose a computationally efficient likelihood evaluation method that avoids inverting very large matrices.

E1432: **Nonparametric inference on Levy-driven Ornstein-Uhlenbeck processes**
Presenter: **Daisuke Kurisu**, Tokyo Institute of Technology, Japan
Nonparametric inference is studied for a stationary Levy-driven Ornstein-Uhlenbeck (OU) process with a compound Poisson subordinator. We propose a new spectral estimator for the Levy measure of the Levy-driven OU process under macroscopic observations. We derive multivariate central limit theorems for the estimator over a finite number of design points. We also derive high-dimensional central limit theorems for the estimator in the case that the number of design points increases as the sample size increases. Building upon these asymptotic results, we develop methods to construct confidence bands for the Levy measure and propose a practical method for bandwidth selection.

EG417 Room C1 CONTRIBUTIONS IN SPATIAL STATISTICS Chair: Marie-Colette Van Lieshout

E0823: **A time varying parameter model to estimate the short-term effects of air pollution on human health**
Presenter: **Pasquale Valentini**, University G. d Annunzio of Chieti-Pescara, Italy
Co-authors: Clara Grazian, Luigi Ippoliti, Lara Fontanella

A hierarchical spatio temporal regression model is introduced to study the spatial and temporal association existing between health data and air pollution. The model is developed for handling measurements belonging to the exponential family of distributions and allows the spatial and temporal components to be modelled conditionally independently via random variables for the (canonical) transformation of the measurements mean function. Pollution exposure are linked with the health outcomes through a regression model which allows for time variation in parameters (e.g. Markov switching, structural break models, threshold models, etc.).

E1585: **Likelihood approximation and prediction for large datasets of spatial data using hierarchical matrices**
Presenter: **Anastasiia Gorshechnikova**, University of Padova, Italy
Co-authors: Carlo Gaetan

Large datasets with n irregularly sited locations are difficult to handle for several applications of Gaussian random fields such as maximum likelihood estimation (MLE) and kriging prediction since they require a computational complexity of order $O(n^3)$. For relatively large n, the exact computation becomes unfeasible and alternative methods are necessary. Several approaches have been proposed to tackle this problem. Most of them assume a specific form for the spatial covariance function and use different methods to approximate the resulting covariance matrix. A methodology was developed using hierarchical matrices that resulted in a log-linear computational cost due to the partitioning of the matrix into dense and low-rank blocks according to specific given conditions. The approximation of the covariance matrix in this format allowed for fast computation of the matrix-vector products and matrix factorisations followed by the efficient MLE and kriging prediction. This method was then applied to a real dataset on the atmospheric carbon dioxide mole fraction of the earth and the prediction accuracy and computational time were compared with other methods. The first experiments show that the developed approach is the most efficient in terms of the root mean-squared prediction error and computational time.

E1492: **Effective probability distributions for spatially dependent processes**
Presenter: **Anastassia Baxevani**, University of Cyprus, Cyprus
Co-authors: Dionissios Hristopulos

Spatially distributed physical processes can be modeled as random fields. The complex spatial dependence is then incorporated in the joint probability density function. Knowledge of the joint probability density allows predicting the field values at points where measurements are missing. The probability distribution of spatially dependent processes often exhibits significant deviations from Gaussian behavior. However, only a few non-Gaussian joint probability density models admit explicit expressions. In addition, spatial random field models based on Gaussian or non-Gaussian joint densities incur formidable computational costs for big datasets. We propose an "effective distribution" approach which replaces the joint probability density with a product of univariate conditional probability density functions modified by a local interaction term. The effective densities involve localized parameters that link the densities at different locations. The prediction of the field at unmeasured locations is formulated in terms of the respective effective distribution and local constraints. We also propose a sequential simulation approach for generating multiple field realizations based on the effective distribution approach. The effective probability density model can capture non-Gaussian dependence, and it can be applied to large spatial datasets, since it does not require the storage and inversion of large covariance matrices.

E1059: **Assessing segregation in complex networks through a multifocal approach**
Presenter: **Julien Randon-Furling**, Universite Paris 1 Pantheon Sorbonne, France
Co-authors: Madalina Olteanu

The issue of revealing and quantifying multiscalar patterns of segregation in complex social networks is addressed. Instead of clustering or detecting communities, our new method provides a multifocal image of the network while highlighting its most segregated zones, its hotspots in terms of segregation. Inspired by a previous work with spatial data, we consider a connected graph with weighted edges, weights representing distances (spatial, social,). Each vertex in the graph carries the value taken by a random variable. The empirical probability distribution in the whole network is known. To each vertex, we sequentially aggregate its neighbours according to a shortest path rule and/or a random walk. For each aggregated group, the probability distribution within the group is compared to that of the entire network. Eventually, for each of these trajectories of aggregates, the distance converge to zero, but the way this is achieved encompasses all information on the relative singularity of the starting vertex within the network. Furthermore, by comparing the actual trajectories with those obtained from random permutations of the vertex values, one may characterize the global structure of the network and its global level of segregation.

E1490: **Testing parametric regression models when the errors are spatially correlated**
Presenter: **Andrea Meilan-Vila**, Universidade da Coruna, Spain
Co-authors: Jean Opsomer, Mario Francisco-Fernandez, Rosa Crujeiras

A common technique in a statistical data analysis is to determine the appropriateness of a parametric model to represent a dataset. As part of this determination, it is advisable to formally test the model, by treating the parametric model as the null hypothesis against an alternative model, and evaluating the probability of obtaining the observed data under the null hypothesis. The choice of the alternative hypothesis model is crucial in this determination. Nonparametric models may be a choice, since they are quite flexible. A spatial stochastic process, which consists of a collection of random variables indexed on a domain of R^d, is considered. In this framework, the observed data tend to exhibit an important feature, close observations tend to be more similar than those that are far apart. Therefore, such observations cannot be treated as independent and the dependence structure should be taken into account and properly introduced into the model. In a spatial context, a weighted L_2-test comparing nonparametric and parametric spatial regression fits is presented and theoretically studied. The nonparametric multivariate local linear regression estimator is used in this procedure. Additionally, the finite sample performance of the test is addressed by simulation, introducing a bootstrap calibration procedure.

EG033 Room D1 CONTRIBUTIONS IN NONPARAMETRIC STATISTICS Chair: Germain Van Bever

E1384: **Nonparametric changepoint analysis of multiple time series**
Presenter: **Elfred John Abacan**, College of Arts and Sciences, University of the Philippines Visayas, Philippines
Co-authors: Erniel Barrios, Joseph Ryan Lansangan
Analysis on changes in the level of time series helps in characterizing common components of multiple time series that helps identify the shared behavior of the data generating process with some known events that causes perturbations in the behavior of the time series. Change in variance of the error structure leads to model misspecification and more serious violation of other assumptions that facilitates model estimation. Volatility models are used to incorporate variance structure into the mean model, but this often suffers from overparameterization especially in multiple series data. A model for structural change in the variance component is estimated through the backfitting algorithm, this is used then as a reference for a test based on sieve bootstrap to detect changes in the variance of multiple time series. Simulation study shows the test performs well in terms of power and size.

E1658: **Nonparametric testing of conditional independence using asymmetric kernels**
Presenter: **Eduardo Fonseca Mendes**, Fundacao Getulio Vargas, Brazil
Co-authors: Marcelo Fernandes
Statistical tools for testing conditional independence between X and Y given Z are developed. In particular, we test whether the conditional density of X given Y and Z is equal to the conditional density of X given Z only. We gauge the closeness between these conditional densities using a generalized entropic measure. To avoid degeneracy issues, we transform the variables of interest (X,Y,Z) to bound them in the unit interval, and then estimate their conditional densities using beta kernels. The latter are convenient because they are free of boundary bias. We show that our test statistics are asymptotically normal under the null hypothesis as well as under local alternatives. We assess the finite-sample properties of our entropic-based tests of conditional independence through Monte Carlo simulations.

E1629: **Multiscale inference and long-run variance estimation in nonparametric regression with time series errors**
Presenter: **Marina Khismatullina**, University of Bonn, Germany
Co-authors: Michael Vogt
New multiscale methods to test qualitative hypotheses about the trend function in the nonparametric regression model with time series errors are developed. Practitioners are often interested in whether the trend has certain shape properties. For example, they would like to know whether it is constant or increasing/decreasing in certain time regions. Our multiscale methods allow us to test for such shape properties of the trend. In order to perform these tests, we require an estimator of the long-run error variance. We propose a new difference-based estimator of it for the case that the errors follow a general AR(p) process. In the technical part, we derive asymptotic theory for the proposed multiscale tests and the estimator of the long-run error variance. The theory is complemented by a simulation study and an empirical application to climate data.

E1416: **Bandwidth selection in nonparametric regression with very large sample size**
Presenter: **Daniel Barreiro Ures**, Universidade da Coruna, Spain
Co-authors: Mario Francisco-Fernandez, Ricardo Cao
In the context of nonparametric regression estimation, the behaviour of kernel methods such as the Nadaraya-Watson or local linear estimators is heavily influenced by the value of the bandwidth parameter, which determines the trade-off between bias and variance. This clearly implies that the selection of an optimal bandwidth, in the sense of minimizing some risk function (MSE, MISE, etc.), is a crucial issue. However, the task of estimating an optimal bandwidth using the whole sample can be very expensive in terms of computing time in the context of big data, due to the computational complexity of some of the most used algorithms for bandwidth selection (leave-one-out cross validation, for example, has $O(n^2)$ complexity). To overcome this problem, we propose two methods that estimate the optimal bandwidth for several subsamples of our large dataset and then extrapolate the result to the original sample size making use of the asymptotic expression of the MISE bandwidth. Preliminary simulation studies show that the proposed methods lead to a drastic reduction in computing time, while the statistical precision is only slightly decreased.

E1665: **Volatility estimation in nonlinear heteroscedastic functional regression model with martingale difference errors**
Presenter: **Mohamed Chaouch**, United Arab Emirates University, United Arab Emirates
The aim is to study the asymptotic properties of the conditional variance estimator in a nonlinear heteroscedastic functional regression model with martingale difference errors. A kernel-type estimator of the conditional variance is introduced when the response is a real-valued random variable and the covariate takes values in an infinite dimensional space endowed with a semi-metric. Under stationarity and ergodicity assumptions, a uniform almost sure consistency rate as well as the asymptotic distribution of the estimator, are established. To build confidence intervals for the conditional variance, two approaches are discussed. The first one is based on the normal approximation approach and the second applies empirical likelihood method. We stress on the fact that errors are assumed to form a martingale difference and may depend on the covariate. Moreover, our results hold under a general dependency structure (ergodicity) and without assuming any mixing conditions which allow to cover a larger class of dependent processes. Two simulation studies are carried out to show the performance of the proposed estimator and to compare the two methods in building confidence intervals. An application to volatility prediction of the daily peak electricity demand in France, using the previous intraday load curve, is also provided.

EP689 Room Ground Level Hall POSTER SESSION II Chair: Panagiotis Paoullis

E1299: **A comparative study on the measures for the prediction accuracy of a survival model**
Presenter: **Asanao Shimokawa**, Tokyo University of Science, Japan
Co-authors: Etsuo Miyaoka
Evaluating the prediction accuracy of a survival analysis model is one of the important problems in practical applications. Also, these evaluations are used for assessment purpose of an interesting risk score ability in the model. Several different measures have been proposed for this purpose: discrimination based approaches like Harrell's C-index and modification versions of it, sensitivity and specificity based approaches or ROC and AUC based approaches in a similar sense, and explained variation or loss function based approaches. We compare the performances of these evaluation measures of the Cox proportional hazard model through simulation studies under several situations such as incorrect the covariates or interactions that should be included in the model. In addition to this, we show the results of applying these measures to the model obtained from an actual data.

E1393: **Analysis of trends in congenital anomalies prevalence: Method comparison simulation study**
Presenter: **Jan Klaschka**, Institute of Computer Science of the Czech Academy of Sciences, Czech Republic
Co-authors: Marek Maly, Antonin Sipek
Data of several national health registries covering years 1994 - 2015 are analyzed in order to detect trends in the prevalence of different kinds of congenital anomalies (birth defects). The standard analysis tool in such a situation is the Poisson regression, but non-statisticians frequently use the simple linear regression instead. A provocative question then arises, whether - in our specific conditions - the former approach is better enough

than the latter one. In order to answer, a simulation study was designed, comparing the two methods both under the null hypothesis of no trend, and in presence of trends of different shapes and intensities. The settings examined mimicked the reality by using the actual series length and yearly numbers of newborns, and realistic range of event frequencies. The results speak uniformly in favour of the Poisson regression: The simplistic approach (linear regression) was outperformed in the terms of test power over a broad spectrum of parameter values used.

E1403: **Detection and comparison of high suicidal execution area in Japan by areal statistics of committed suicide**
Presenter: **Takafumi Kubota**, Tama University, Japan
Some high suicidal execution areas in Japan are detected. The study also compared these areas between years, sexes, and age groups to find out demographic characteristics of high suicidal risk. The study covered two kinds of regional type; prefecture and municipality in Japan.

E1412: **Numerical analysis of intracellular amino acid profiles of breast cancer cells**
Presenter: **Jaesik Jeong**, Chonnam National University, Korea, South
Various efforts to understand the relationship between biological information and disease have been done using many different types of high-throughput data such as genomics and metabolomics. However, information obtained from previous studies was not satisfactory, implying that new direction of studies is in need. Thus, we have tried profiling intracellular free amino acids in normal and cancerous cells to extract some information about such relationship by way of the change in IFAA levels in response to the treatment of three kinase inhibitors. We define two measures such as relative susceptibility (RS) and relative efficacy (RE) to numerically quantify susceptibility of cell line to treatment and efficacy of treatment on cell line, respectively. We applied principal component analysis (PCA) to the intracellular free amino acids (IFAAs) of isogenic breast cells with oncogenic mutation in K-Ras or PI3K genes to investigate the change in IFAA levels in response to the treatment of three kinase inhibitors. Two-dimensional plot, which was graphically represented by using the first two principal components (PCs), enabled us to evaluate the treatment efficacy in cancerous cells in terms of the quantitative distance of two IFAA profiles from cancerous and normal cells with the same treatment condition.

E1425: **The hybrid Phillips curve formula for DSGE models with a finite number of firms**
Presenter: **Pawel Zajac**, AGH University of Science and Technology, Poland
In recent years DSGE models, which assume the existence of a finite number of agents on the market, are becoming more and more popular. In particular this modeling approach allows the possibility of default in the equilibrium state. In result, these models are used as a tool to analyze the relationship between macroeconomic variables and business demography in the economy. A common practice while constructing DSGE models is to use an equation describing inflation. A hybrid variant of the new Keynesian Phillips curve is often applied to realize this purpose. The hybrid Phillips curve for DSGE Models with a finite number of firms is being developed.

E1437: **Evaluating water meters performance: An industry case study**
Presenter: **Ana Borges**, Porto Polytechnic - School of Management and Technology of Felgueiras, Portugal
Co-authors: Clara Cordeiro, Regina Casimiro
Infraquinta is the water utility that manages the water and the wastewater services of a well-known tourist place in Algarve, Portugal. Infraquinta seeks to improve mechanisms for predictive planning based on data analysis. In particular on evaluating water meters performance by using historical data. Hence, the main purpose of this analysis is to find out the water meter performance breakpoint and when to replace it. Firstly, the decomposition of monthly time series into seasonality, trend and irregular components is performed. The nonparametric Seasonal-Trend decomposition by Loess can identify a seasonal component that changes over time, a non-linear trend and it can be robust in the presence of outliers. Secondly, the detection of structural breaks in the trend component is performed since it is the one related to the meter performance. This approach tests for structural changes in linear regression models, estimating the number of segments and the breakpoints by minimizing the Bayesian information criterion and the residual sum of squares. Outcomes shown that this methodology can detect the water meter breaking points and that if incorporated in a system that automatically analyses the data and updates the information, could be the solution sought by Infraquinta.

E1485: **Classification of molecular characteristics to identify disease targets by score function of violations**
Presenter: **Shalem Leemaqz**, University of Adelaide, Australia
Co-authors: Irene Hudson, Andrew Abell
Expanding disease modifying targets to pharmacological manipulation is vital to reducing morbidity and mortality, and is critical to drug discovery. Modelling disease targets would allow for prediction and prioritisation based on their molecular characteristics and druggability. Classification rules by support vector machine (SVM), Recursive partitioning (RP) and Random forest (RF) based on 8 molecular parameters were performed to classify disease targets with high (≥ 4) or low (< 4) violator scores. Predictors includes the 4 traditional parameters of Lipinskis rule of five (Ro5), plus 4 extra parameters (polar surface area PSA, number of rotatable bonds and rings, N and O atoms, and distribution coefficient log D). A total of 1279 small molecules from the DrugBank database (Knox 2011), combining detailed drug (i.e. chemical, pharmacological and pharmaceutical) data with drug disease target information, were analysed and these were shown to be aligned with 172 targets. For the validation set SVM gave an AUC of 93.2% (95% CI 85.8%-99.9%). The RF classification gave similar but slightly lower AUC (91.5%; 95%CI 83.4%-99.5%), followed by RP with AUC 87.5% (95%CI 78%-97%). Results illustrated that SVM used in combination with simple molecular descriptors of disease targets can provide a reliable assessment of violation scores, and hence druggability.

E1504: **Business perceptions of corruption in Europe: A multilevel analysis**
Presenter: **Nikolai Witulski**, Instituto Universitario de Lisboa ISCTE, Portugal
Co-authors: Jose Dias
A two-level latent-variable model is applied to Eurobarometer data regarding employees corruption perception in the 28 countries of the European Union (EU28). For the first time the explicit impact of company and country characteristics on the perception of business corruption is studied in EU28. Both characteristics are added as first and second levels of the hierarchy, respectively. In particular, the first-order latent factor structure covers four dimensions of employees corruption perception: general perception, current practices, widespread & influence, and punishment. The goodness of fit of all models, including the two conditional models that result from adding company characteristics and country-level indicators, can be confirmed (SRMR, CFI, TLI, and RMSEA). Company characteristics (e.g., number of employees, sector, and turnover) are statistically significant and explain the corruption perception of employees in the EU28. However, when the country-level indicators are added (e.g., Hofstedes cultural dimensions, free press, civil liberties, women representation in parliament, and the three pillars of sustainability), company characteristics become less important (only three variables remain statistically significant: sector, number of employees, and turnover). Five dimensions of Hofstede framework, two pillars of sustainable development, representation of women in parliament, free press, and civil liberties are statistically significant country-level indicators.

E1497: **A piecewise linear approximation approach for the simultaneous optimisation of price and loan-to-value**
Presenter: **Marius Smuts**, North West University, South Africa
Co-authors: Fanie Terblanche
In a competitive financial industry, one of the challenges for secured retail lending products is to determine the optimal prices (i.e. interest rates) that maximise both the loan take-up probability and expected revenue. However, the technologically driven and competitive nature of the financial

industry require the computation of the optimal prices for secured lending products be performed as efficiently, accurately and timeously as possible, while still adhering to certain risk distribution constraints. We discuss a response model which relates take-up probabilities with price and loan-to-value (loan amount expressed as a percentage of the value of the underlying asset), hence taking into account a customer's willingness to pay for a loan. A piecewise linear approximation approach is then followed to simultaneously determine the optimal price and loan-to-value percentage for a potential customer. Furthermore, in an attempt to reduce computing times, linear interpolation is applied to these optimally calculated values.

E1619: **Dependence structure of realized volatility and illiquidity**
Presenter: **Henryk Gurgul**, AGH University of Science and Technology, Poland
The results of an investigation into the relationship between illiquidity and realized volatility of the time series of stocks listed in the Warsaw Stock Exchange (WSE) and Vienna Stock Exchange (VSE) are presented. The first measure of illiquidity is the well-known Amihud ratio AMI, and the second is a transformation of the Liquidity Index called ILLIX. The results of the detection of structural breaks, their removal, calculation of long memory, and finally the dependence structure of the illiquidity and realized volatility by copulas are also demonstrated. Both types of series exhibit structural breaks and long-memory properties. Despite the similarity of the illiquidity measures, their associations with realized volatility are different. The dependence structures described by copulas for pairs AMI- realized volatility show a dependence in the upper tails; i.e., the high values of illiquidity are related to high volatility. However, in the case of the ILLIX-realized volatility pairs, the dependence was detected in the lower tail; i.e., the low ILLIX is accompanied by low realized volatility.

C1696: **Application of a principal component regression for the electricity markets data**
Presenter: **Leila Louhichi**, Faculty of economic science and management of Sousse, Tunisia
Co-authors: Zied Kacem, salwa ben ammou
Controlling energy consumption, particularly electricity consumption, is one of the pillars mentioned by the public authorities for implementing the energy transition. As a result, producing and consuming electricity in the residential sector is among the strategic objectives in Tunisia as a result of constraints related to the exploitation of existing energy sources. We will determine the factors that explain the electricity consumption in the residential sector. To do this, and to reduce the number of explanatory variables and solve the multi-colinearity problem of the variables used, we opted for principal component regression as a regression model. This principal component regression model (PCR) will be validated by an application on a set of regional data previously performed on variables initially correlated. Application of a principal component regression for the electricity markets data.

E1610: **How many people participated in candlelight protests: Estimating dynamic crowd size**
Presenter: **Seonghun Cho**, Seoul National University, Korea, South
Co-authors: Johan Lim, Woncheol Jang
The recent controversy on the size of crowd at candlelight protests raises a serious question on how to estimate the crowd size. There was a significant discrepancy between police and protest organizers on the size of crowds at each protest. While police often reports the crowd at its peak, organizers want to count anyone who participated in the event from its start to finish. We propose a new counting method using an inverse probability weight estimator. Assuming that the arrival and departure times of crowd can be estimated using additional survey, we estimate the probability of a subject being in the events at a given time. We also propose a bootstrap procedure to obtain the variance of the estimator. We demonstrate the performance of the proposed method with simulation studies and the data from Korea's March for science, a global event across the world on Earth day April 22, 2017.

E1758: **Estimation of selection-bias adjusted FDR for the overall assessment in follow-up studies**
Presenter: **Donghwan Lee**, Ewha Womans University, Korea, South
Co-authors: Boram Jeong, Yudi Pawitan, Woojoo Lee
To verify the generalizability and reproducibility of scientific findings, follow-up studies are often conducted where only a few significant discoveries in the previous experiment are re-evaluated. We consider multiple testing problems with the selection-bias and propose a way to estimate the false discovery rate (FDR) for adjusting such bias. We show how to use the selection-bias adjust FDR estimation to assess the overall significance of the discoveries based on the integrated data in follow-up studies as well as the preliminary study. Simulation studies and two metabolomics datasets are considered to illustrate the application of the proposed method in high-throughput data analysis.

E1760: **Multi-component ridge regression for heterogeneous correlation structure of covariates**
Presenter: **Junghwan Kim**, Institute of Water Resources System, Inha University, Korea, South
Co-authors: Woojoo Lee
Ridge regression is a classical regression method applicable to the data showing high multicollinearity. In practice, one-dimensional tuning parameter is often used for ridge regression. However, in high-dimensional data, blockwise heterogeneous correlation structures among predictors are often observed so that one tuning parameter may not be sufficient to control different degree of multicollinearity simultaneously. We propose a multi-component ridge regression for doing two folds: (1) it automatically finds the correlation blocks of predictors and (2) it allows multi-dimensional tuning parameters. We compare the predictive performance of the proposed multi-component ridge regression with traditional ridge regression through a numerical study and real data analysis.

E1761: **On heritability estimation and genome-wide association study for Yorkshire pig family data**
Presenter: **Hyunman Sim**, Department of Statistics, Inha University, Incheon, Republic of Korea, Korea, South
Co-authors: Woojoo Lee, Kyunghyun Nam, Donghyun Shin
Yorkshire pig family data collected in Korea contain various traits associated with meat, pedigree information and single nucleotide polymorphism (SNP) data. First, they were analyzed for heritability estimation and genome-wide association studies using genome-wide complex trait analysis (GCTA). However, since GCTA was originally developed for unrelated individuals, the validity of GCTA for our pig family data should be checked carefully. We compare (1) two heritability estimates and (2) two genome-wide association study results from pedigree information and SNP data, and study key causes to make differences between two approaches. Based on these, we discuss whether GCTA using SNP data provide reliable results for our Yorkshire pig family data.

CO232 Room Aula Magna ADVANCES IN LATENT VARIABLE MODELLING Chair: Robert Kohn

C0691: **A flexible Bayesian model for treatment effects on panel outcomes**
Presenter: **Helga Wagner**, Johannes Kepler University, Austria
Identification and estimation of treatment effects is an important issue in many application fields, e.g. to evaluate of the effectiveness of social programs, government policies or medical interventions. In contrast to randomized studies, unobserved confounding due to endogeneity of treatment selection has to be taken into account for data from observational studies. In the Bayesian approach this is accomplished by specifying a joint model of treatment selection and the potential outcomes. For the estimation of dynamic effects of a binary treatment on a continuous outcome observed over subsequent time periods two models, the switching regression model and the shared factor model, have been suggested so far. We show that both impose restrictions on the joint correlation structure of treatment selection and the two outcomes sequences that can result in biased treatment effects estimates. To achieve more flexibility we propose a new model that allows us to separate longitudinal association of the outcomes from association due to endogeneity of treatment selection. We employ this model to analyse the effects of a long maternity leave on earnings of Austrian mothers, where we exploit a change in the parental leave policy in Austria that extended maternal benefits from 18 months since birth of the child to 30 months. This analysis is based on data from the Austrian Social Security Register which contains individual employment histories since for all Austrian employees.

C1340: **Bayesian dynamic regularised forecast combinations**
Presenter: **Marcel Scharth**, The University of Sydney Business School, Australia
Co-authors: Andrey Vasnev, Haonan Zhang
How can we leverage the diversity of professional forecasts and plausible models available for many forecasting problems? Multiple empirical studies have documented that while pooling forecasts improves accuracy in many applications, it is often challenging to design combination schemes that outperform a simple average of the available forecasts–an empirical pattern known as the forecasting combination puzzle. We address this problem by developing a Bayesian approach that automatically computes regularised forecast combinations based on promising subsets of forecasters or models. We formulate this approach as a state space model that additionally accounts for the potential dynamic features of the data, such as autocorrelated errors, heteroscedasticity, attrition of forecasters, and time-varying performance. We propose an estimation algorithm based on Sequential Monte Carlo (SMC). Empirical results illustrate that our approach can improve accuracy relative to simple average combinations and other recently proposed methods based on subset selection and regularisation.

C1580: **Well-tempered Hamiltonian Monte Carlo on active-space**
Presenter: **Ritabrata Dutta**, Warwick University, United Kingdom
Co-authors: Antonietta Mira
When the gradient of the log-target distribution is available, Hamiltonian Monte Carlo (HMC) has been proved to be an efficient simulation algorithm. However, HMC performs poorly when the target is high-dimensional and it has multiple isolated modes. To alleviate these problems we propose to perform HMC on a locally and continuously tempered target distribution. This tempering is based on an efficient approach to simulate molecular dynamics in high-dimensional space, known as well-tempered meta-dynamics. The tempering we suggest is performed locally and only along the directions of the maximum changes in the target which we identify as the active space of the target. The active space is the span of the eigenfunctions corresponding to the dominant eigenvalues of the expected Hessian matrix of the log-target. To capture the state dependent non-linearity of the target, we iteratively estimate the active space from the most recent batch of samples obtained from the target. Finally, we suggest a re-weighting scheme to provide importance weights for the samples drawn from the continuously-tempered distribution. We illustrate the performance of this scheme for target distributions with complex geometry and multiple modes on high-dimensional spaces in comparison with traditional HMC with No-U-Turn-Sampler (NUTS).

C1577: **A latent threshold approach to large-scale mixture innovation models**
Presenter: **Gregor Kastner**, WU Vienna University of Economics and Business, Austria
Co-authors: Florian Huber, Martin Feldkircher
A straightforward algorithm is proposed to carry out inference in large time-varying parameter vector autoregressions (TVP-VARs) with mixture innovation components for each coefficient in the system. We significantly decrease the computational burden by approximating the latent indicators that drive the time-variation in the coefficients with a latent threshold process that depends on the absolute size of the shocks. The merits of our approach are illustrated with two applications. First, we forecast the US term structure of interest rates and demonstrate forecast gains of the proposed mixture innovation model relative to other benchmark models. Second, we apply our approach to US macroeconomic data and find significant evidence for time-varying effects of a monetary policy tightening.

C1728: **Efficiently combining pseudo marginal and particle Gibbs sampling**
Presenter: **Robert Kohn**, University of New South Wales, Australia
Co-authors: Christopher K Carter, David Gunawan
The focus is on the problem of statistical inference for both the unobserved latent states and the parameters in a class of non-linear and non-Gaussian state space models. We develop an efficient particle Markov chain Monte Carlo (PMCMC) sampling scheme that converges to the posterior distribution of the unobserved latent states and the parameters and extends the PMCMC methods developed previously.

CO580 Room Aula C LEARNING COMPLEX DATASETS IN ECONOMETRICS Chair: Yuan Ke

C0747: **Leverage subsampling for vector autoregression**
Presenter: **Shuyang Bai**, University of Georgia, United States
Co-authors: Ping Ma, WenXuan Zhong, Rui Xie, Zengyan Wang
When performing statistical analysis of streaming multivariate time series data, the computational capacity often cannot afford to allow all the available data to estimate the model parameters. The sample size needs to be reduced and a subsample must be selected. However, a naive subsample selection approach may lose statistical efficiency. In the context of vector autoregression, we propose a subsample selection method based on leverage score, which is shown to achieve a high efficiency compared with some naive approaches. Furthermore, the method can be adapted to an online decentralized computing setup.

C1010: **Forecast US bond risk premia with partially observed factors**
Presenter: **Yuan Ke**, Penn State University, United States
Co-authors: Jianqing Fan, Yuan Liao
The forecast of US bond risk premia with factor models is studied when the latent factors can be partially explained by observed covariates. With those covariates, both the factors and loadings are identifiable up to a rotation matrix even only with a finite dimension. To incorporate the explanatory power of these covariates, we propose a smoothed principal component analysis (PCA): (i) regress the data onto the observed covariates, and (ii) take the principal components of the fitted data to estimate the loadings and factors. We show that both the estimated factors

and loadings can be estimated with improved rates of convergence compared to the benchmark method. The degree of improvement depends on the strength of the signals, representing the explanatory power of the covariates on the factors. We can also accurately estimate the percentage of unexplained components in factors. The proposed estimator is robust to possibly heavy-tailed distributions, which are encountered in many high-dimensional applications for factor analysis.

C0790: **A general theory for detecting changes-in-mean and changes-in-slope**
Presenter: **Chao Zheng**, Lancaster University, United Kingdom
The aim is to study the finite sample behaviour of an approach to detecting changepoints that is based on maximising a penalised likelihood. These give general results as to when such a procedure can consistently estimate the number of changes and accurately estimate their position. The results we obtained are applied to the problem of detecting changes-in-mean and changes-in-slope. In the latter case we obtain tighter results on the value of penalty that can be used as compared to existing theory. Moreover, the techniques can be easily adapted to other scenarios as long as some basic properties for detecting a single changepoint are satisfied. We demonstrate the usefulness of our approach through numerical experiments on both synthetic data and real data examples.

C1212: **Nonparametric production function estimation with shape constraints**
Presenter: **Yining Chen**, London School of Economics and Political Science, United Kingdom
A new approach is developed to estimate a production function based on the economic axioms of the Regular Ultra Passum (RUP) law and convex non-homothetic input isoquants. Central to the development of our estimator is stating the axioms as shape constraints and using shape constrained nonparametric regression methods. Under regularity conditions, we show the consistency of our estimator. Algorithms are also developed to facilitate the computation of the estimator. Finally, we implement this approach using data from the Japanese corrugated cardboard industry.

C0705: **On the properties of simulation-based estimators in high dimensions**
Presenter: **Mucyo Karemera**, Penn State, United States
Co-authors: Stephane Guerrier, Maria-Pia Victoria-Feser, Yanyuan Ma, Samuel Orso
Considering the increasing size of available data, the need for statistical methods that control the finite sample bias is growing. This is mainly due to the frequent settings where the number of variables is large and allowed to increase with the sample size bringing standard inferential procedures to incur significant loss in terms of performance. Moreover, the complexity of statistical models is also increasing thereby entailing important computational challenges in constructing new estimators or in implementing classical ones. A trade-off between numerical complexity (e.g. approximations of the likelihood function) and statistical properties is often accepted. However, numerically efficient estimators that are altogether unbiased, consistent and asymptotically normal in high-dimensional problems would be advantageous, especially for real data applications. We set a general framework from which such estimators can be easily derived for wide classes of models. The approach allows various extensions compared to previous results as it is adapted to possibly inconsistent estimators and is applicable to discrete models and/or models with a large number of parameters (compared to the sample size). We consider an algorithm, namely the Iterative Bootstrap, to efficiently compute simulation-based estimators by showing its convergence property.

CO470 Room P1 MODELLING SPATIAL DATA IN BUSINESS AND ECONOMICS Chair: Marzia Freo

C0261: **Spatial bootstrapped microeconometrics: Forecasting for out-of-sample geo-locations**
Presenter: **Katarzyna Kopczewska**, University of Warsaw, Poland
Spatial econometrics for mass point geo-locations has a limited possibility of forecasting with a calibrated model for a new out-of-sample geo-point. This is because the spatial weights matrix W is defined for in-sample observations only, as well as the computational complexity. The aim is to propose a novel methodology to calibrate both space and model relationships by using bootstrap and tessellation. Bootstrapping enables the calibration of the econometric model without the need for estimation on the whole dataset. Partitioning Around Medoids (PAM) algorithm finds the best points representation in the bootstrapped set of models and generates the medoids coefficients. Tessellated points in the selected best model allow for a representative division of space. New out-of-sample points are assigned to tiles and linked to W as a replacement for original point. The quality of the forecast is tested for the different scenarios of this bootstrap procedure. This efficient procedure supports the big data geo-located point data and makes feasible a usage of calibrated spatial models as a forecasting tool for out-of-sample data. This methodology will find its applications in real estate market forecasting as well as models of business location.

C0502: **Social capital and regions absorptive capacity in Europe**
Presenter: **Stefano Ghinoi**, University of Helsinki, Finland
Knowledge spillovers are fundamental for promoting regional development and for enhancing the innovation process. Scholars have demonstrated that these spillovers occur through a recombination of knowledge across industrial sectors, which in turn can have an impact on economic growth. However, regions absorptive capacity is moderated also by local social capital, which can affect firms innovativeness. The relationship between local social capital and regions absorptive capacity (measured by patents) is investigated, by focusing on European NUTS3 regions and the impact on employment and productivity. Building on previous work on regional social capital, we attempt to integrate social capital metrics within the framework of regions absorptive capacity at European level.

C0564: **Entropy based small area estimation for count data with spatial effects**
Presenter: **Rossella Bernardini Papalia**, University of Bologna, Italy
Co-authors: Esteban Fernandez Vazquez
Statistical information for empirical analysis is very frequently available at a higher level of aggregation than it would be desired. Small area estimation is important in light of a continual demand by data users for finer geographic detail of official statistics and for various subgroups. The spatial disaggregation of the socio-economic data is considered complex owing to the inherent spatial properties and relationships of the spatial data, namely, spatial dependence and spatial heterogeneity. The spatial dependence, spatial heterogeneity and the effect of scale produce major technical issues that largely impact on the accuracy of the small area estimates. We propose entropy-based small area estimation methods for count areal data that introduce spatial effects by using all available information at each level of aggregation even if it is incomplete. The proposed methods are validated through the Monte Carlo simulations using ancillary information. An empirical application to real data is also presented.

C0793: **Spatial comparisons of consumer prices using big data: Empirical evidence from Italy**
Presenter: **Tiziana Laureti**, University of Tuscia, Italy
Co-authors: Federico Polidoro
Accurate measurements of price level differences across geographical areas within a country is essential to better assess regional disparities, thus enabling policy makers to adequately identify and address areas of intervention. The increasing availability of new source of data concerning household consumption expenditure (i.e. scanner data) may change the current approach for estimating sub-national spatial price indexes from both a methodological and an empirical point of view. Sub-national spatial price indexes are estimated by focusing on the stochastic approach since it allows us to incorporate modern econometric tools and deal with the issue of spatial dependence which is inherent in consumer prices. We use a scanner dataset set up for experimental CPI computations for Italy in 2017 which includes information on prices, quantities and quality

characteristics of products at barcode level. This dataset refers to grocery products sold in a random sample of approximately 1,800 outlets across Italian provinces belonging to the most important retail chains (95% of modern retail trade distribution), covering 55.4% of total retail trade distribution for this product category. Estimates of provincial and regional spatial price indexes obtained using various index number formulae for specific product groups and for Food and Non-Food consumption aggregates are presented.

C1154: **Sub-national price indexes for housing: Methodological issue and empirical analyses for Italy**
Presenter: **Ilaria Benedetti**, University of Tuscia, Italy
Co-authors: Tiziana Laureti, Luigi Biggeri, Marco Brandi

Spatial price indexes (SPIs) measuring differences in price levels across regions within a country are essential for comparing real income and standards of living as well for measuring poverty and rural-urban differences. The issue of computing sub-national SPIs has been discussed extensively in literature and the importance of constructing sub-national SPIs has been acknowledged also in the research program of the international comparison program coordinated by the World Bank. Housing is the largest expenditure item in the household budget, especially for poor households, for which rents weight approximately 40%-50% of their total consumption expenditures. Therefore, computing spatial price indexes for house rents (SPIHRs) acquires considerable importance in the context of poverty measurement. However, systematic attempts to compile SPIHRs on a regular basis have been hampered by data unavailability. The aim is to focus on the methodological and empirical issues which arise when computing SPIHRs. By using data on house prices and rents produced by the Italian revenue and tax agency every six months, concerning all the Italian municipalities we estimated SPIHRs for different housing types. With the aim of accounting for spatial dependencies in house rents we include adjustments that incorporate spatial autocorrelation in the stochastic approach to spatial price comparisons.

CO110 Room A2 FINANCIAL MODELLING Chair: Genaro Sucarrat

C0637: **Negative skewness of asset returns with positive time-varying risk premia**
Presenter: **Dimitra Kyriakopoulou**, Universite Catholique de Louvain, Belgium
Co-authors: Christian Hafner

Portfolio selection and risk management are important problems that investors and portfolio managers face. The distributional characteristics of returns, for example the unconditional skewness, are able to create new challenges in the classical portfolio theory of Markowitz. Apart from the mean and variance predictability, portfolio choice can be also made with skewness information. Such a perspective can create for investors the notion of skewness corrected risk estimate, as their financial decisions can be affected by properties of the return distribution. It is well known that the marginal distribution of financial time series such as returns is often negatively skewed. We investigate the relation between positive time-varying risk premia and the unconditional skewness of returns. We show that if the error distribution is symmetric, the negative unconditional asymmetry of returns should be the outcome of a negative correlation between their first two conditional moments. Following one of the implications of the intertemporal capital asset pricing model (ICAPM), there is a positive and linear relationship between risk and expected returns. Under a fully parametric EGARCH-in-Mean specification, we propose to use an asymmetric error distribution in order to match the unconditional asymmetry of asset returns. Value-at-Risk prediction of the largest stock market indices is performed as an application.

C0494: **Realized peaks-over-threshold: A time-varying extreme value approach with high-frequency based measures**
Presenter: **Debbie Dupuis**, HEC Montreal, Canada
Co-authors: Marco Bee, Luca Trapin

Recent contributions to the financial econometrics literature exploit high-frequency (HF) data to improve models for daily asset returns. A new class of dynamic extreme value models is proposed that profit from HF data when estimating the tails of daily asset returns. Our realized peaks-over-Threshold approach provides estimates for the tails of the time-varying conditional return distribution. An in-sample fit to the S&P500 index returns suggests that HF data convey information on daily extreme returns beyond that included in low frequency (LF) data. Finally, out-of-sample forecasts of conditional risk measures obtained with HF measures outperform those obtained with LF measures.

C0598: **Dynamic properties and correlation structure of a large panel of cryptocurrencies**
Presenter: **Francesco Violante**, ENSAE ParisTech, France
Co-authors: Luc Bauwens, Jeroen Rombouts

The behaviour of a large portfolio of highly valued and most actively traded cryptocurrencies is studied. Unlike more traditional financial assets, the dynamic behaviour of cryptocurrencies returns is characterised by a particularly high level of volatility, by abnormally large variations, and is affected by extreme shocks to liquidity. We aim at investigating the dynamic properties of cryptocurrencies and particularly the correlation structure linking them, with the scope to identify whether and to what extent there exist diversification opportunities in these markets.

C1155: **Competition, fast growth and commercialization: A copula approach for systemic credit risk in microcredit markets**
Presenter: **Andreas Heinen**, Universite de Cergy Pontoise, France

The sources of systemic credit risk in the microfinance sector in a sample of countries are analysed by using techniques from risk management. More specifically, for all countries and years in our sample, we model the joint distribution of portfolio quality of all microfinance institutions (MFIs) using an equidependent copula and an empirical density for the marginal. Our methodology is based on the idea that a higher level of dependence among MFIs in a given country is an indicator of potential fragility of the sector. We find that penetration, market concentration, and accelerated growth of the sector make the microfinance sector more risky.

C0452: **Tackling infinity: A theory for infinite order autoregressions**
Presenter: **Menelaos Karanasos**, Brunel University, United Kingdom
Co-authors: Alexandros Paraskevopoulos, Philipp Sibbertsen, Richard Baillie

An integrated approach is developed in order to examine the dynamics of autoregressions of infinite order. We provide the closed form of the general solution for such processes. This enables us to characterize these models by deriving, first, its multistep ahead predictor, second, the first two unconditional moments, and third, its covariance structure. To illustrate the practical significance of our results we apply our approach to cyclical long-memory models.

CO124 Room B2 REGIME SWITCHING, FILTERING, AND PORTFOLIO OPTIMIZATION Chair: Joern Sass

C0318: **Deviations from triangular arbitrage parity in foreign exchange and bitcoin markets**
Presenter: **Julia Reynolds**, Universita della Svizzera italiana, Switzerland
Co-authors: Leopold Soegner, Martin Wagner, Dominik Wied

New econometric tools are applied to monitor and detect so-called "financial market dislocations", defined as periods in which substantial deviations from arbitrage parities take place. In particular, we focus on deviations from the triangular arbitrage parity for exchange rate triplets. Due to increasing media attention towards mispricing in the market for cryptocurrencies, we include the cryptocurrency Bitcoin in addition to fiat currencies. We do not find evidence for substantial deviations from the triangular arbitrage parity when only traditional fiat currencies are concerned. However, we document significant deviations from triangular arbitrage parities in the newer markets for Bitcoin.

C0384: **Utility maximization under model uncertainty**
Presenter: **Dorothee Westphal**, TU Kaiserslautern, Germany
Co-authors: Joern Sass

When modelling financial markets one is frequently confronted with model uncertainty. This is meant in the sense that parameters of the model, e.g. the drift of a stock, or the distributions of some factors in the model are only known up to a certain degree. Risk-averse investors in such a market try to maximize their worst-case expected utility. This naturally leads to considering robust optimization problems. We investigate optimal trading strategies for a robust utility maximization problem in a continuous-time Black-Scholes type financial market and impose a constraint that prevents a pure bond investment. The optimal strategy of an investor depends on how uncertain the parameters in the market are. As the degree of model uncertainty increases, investors tend to rely on robust strategies. We show that if the uncertainty exceeds a certain threshold simple strategies such as uniform portfolio diversification outperform more sophisticated ones due to being more robust. This generalizes previous results for a discrete-time model to continuous time.

C0760: **Particle filtering for truncated noise densities**
Presenter: **Elisabeth Leoff**, Fraunhofer ITWM, Germany
Co-authors: Tom Ewen

Particle filters are a popular class of Monte Carlo methods used for state estimation in general state space models, where the filter is not explicitly known. They work online to approximate the marginal distribution of the signal as observations become available. Importance sampling is used at each time point to approximate the distribution with a set of particles, each with a corresponding weight. When the density of the observation noise has finite support, it can happen that all weights are zero and the filter diverges. We present an approach where repeated sampling of all particles is applied, similar to the (partial) rejection control from literature. We demonstrate that this method is valid in the sense that it approximates the correct conditional expectation. If repeated sampling still leads to vanishing weights, it can be combined with smoothing out the noise density, which allows the filter to continue. Numerical examples for both concepts are presented.

C0769: **Market-timing in practice**
Presenter: **Michael Scholz**, University of Graz, Austria
Co-authors: Jens Perch Nielsen, Stefan Sperlich, Enno Mammen

In long-term investment products, it is important to understand the underlying financial risk of the optimal investment profile. Various performance measures, for example, the Sharpe-ratio, were proposed to evaluate those investment strategies in practice. We provide an improved estimator for the Sharpe-ratio which includes prior knowledge in the estimation process of conditional mean and variance function in a predictive regression model focusing on nonlinear relationships between a set of covariates. In an applied part, we compare different investment strategies based on our improved estimators using annual data of the S&P500 in a period from 1872 to 2015.

C1334: **Wavelet analysis applied to directors dealings**
Presenter: **Michaela Kiermeier**, University of Applied Sciences Darmstadt, Germany

The validity of efficient market hypotheses have been tested in a multitude of econometric settings. We investigate if insiders have superior information with which the market can be outperformed in a statistically significant way. For this purpose, we analyze data on insider trades from various European countries that have not been analyzed before and estimate returns on performances by insiders or outsiders who could use information on insiders trade activities. A common problem with this approach however is that expected returns have to be modelled using factor models like CAPM or APT. We use wavelet analysis to distinguish between expected returns and noise. A second application for wavelet analysis is concerned with the time period insider information might be able to generate outperformance. We therefore filter the return data and analyze the performance scale-by-scale.

CO478 Room C2 EMPIRICAL ANALYSIS OF BOND RISK PREMIA Chair: Andrea Berardi

C0373: **Long-term interest rate spillovers from the United States: Expectations, term premia and uncertainty**
Presenter: **Richhild Moessner**, Bank for International Settlements, Switzerland
Co-authors: Aaron Mehrotra, Chang Shu

The aim is to analyse how changes in the components of United States government bond yields affect long-term rates in other major advanced and in emerging market economies, and the role of financial market, policy and geopolitical uncertainty for such spillovers. We find significant spillovers from both the expectations and term premia components of long-term rates in the United States. Changes in the United States term premia affect long-term rates somewhat more in advanced economies than in emerging economies. But when global financial market uncertainty rises, spillovers from US term premia to long-term rates are amplified in both groups of countries. Policy uncertainty increases interest rate spillovers to advanced economies, while geopolitical risk boosts spillovers to emerging market economies.

C0666: **Tracing the impact of the ECB's expanded asset purchase programme on the yield curve**
Presenter: **Wolfgang Lemke**, European Central Bank, Germany
Co-authors: Fabian Eser, Ken Nyholm, Soeren Radde, Andreea Vladu

Since March 2015, the ECB has been purchasing public sector securities as part of its Expanded Asset Purchase Programme. The impact of those bond purchases on the term structure of euro area sovereign bond yields is quantified. The analysis deploys an estimated affine term structure model, in which central bank bond holdings compress term premia via a reduction of the market price of duration risk. We find that at the beginning of 2018 the stock of current and expected central bank bond holdings compressed ten-year term premia by about 100 basis points. The impact is persistent and expected to halve over around five years. The model is also utilised to quantify and interpret how modifications to the design of the purchase programme - such as changes in the re-investment policy - lead to changing yield impacts.

C1272: **A macroeconomic approach to the term premium**
Presenter: **Peter Williams**, International Monetary Fund, United States
Co-authors: Emanuel Kopp
A semi-structural dynamic term structure model is proposed which is augmented with macroeconomic factors to include cyclical dynamics with a focus on medium- to long-run forecasts. The results clearly show that a macroeconomic approach is warranted: While term premium estimates are in line with those from other studies, we provide (i) plausible, stable estimates of expected long-term interest rates and (ii) forecasts of short- and long-term interest rates as well as cyclical macroeconomic variables that are stunningly close to those generated from large-scale macroeconomic models.

C0299: **Nearly exact Bayesian estimation of non-linear no-arbitrage term structure models**
Presenter: **Marco Taboga**, Bank of Italy, Italy
A general method is proposed for the Bayesian estimation of nonlinear no-arbitrage term structure models. The main innovations we introduce are: 1) a computationally efficient method, based on deep learning techniques, for approximating no-arbitrage model-implied bond yields to any desired degree of accuracy; 2) computational graph optimizations for the acceleration of the MCMC sampling of the model parameters and of the unobservable state variables that drive the short-term rate. We apply the proposed techniques to the estimation of a shadow rate model with time-varying lower bound, where the shadow rate can be driven both by spanned unobservable factors and by unspanned macroeconomic factors.

C1040: **Term premia with macro expectations**
Presenter: **Andrea Berardi**, University of Venice, Italy
Term premia are sensitive to volatility and macroeconomic conditions. The aim is to estimate an affine term structure model with time-varying volatility where term premia are strictly interrelated to investors' expectations of future inflation and output growth. The empirical work is based on data for the US, the UK and the Euro Area. In contrast to previous yield-only models, we find that there is strong business cycle variation in term premia that is not revealed in the yield curve and that a significant percentage of the movements can be explained by macro expectations. We also provide evidence on the dynamics of the components of the term premium, that is the inflation risk premium and the real term premium, and on the degree of connectedness of these variables among the different countries.

CO158 Room D2 STATISTICAL MODELS FOR BANKING AND BUSINESS FAILURE PREDICTION Chair: Marcella Niglio

C0292: **Variable selection in proportional hazards cure model with time-varying covariates, application to US bank failures**
Presenter: **Alessandro Beretta**, HEC Liege, Belgium
Co-authors: Cedric Heuchenne
From a survival analysis perspective, bank failure data are often characterised by small default rates and heavy censoring. This empirical evidence can be explained by the existence of a subpopulation of banks likely immune from bankruptcy. In this regard, we use a mixture cure model to separate the factors with an influence on the susceptibility to default from the ones affecting the survival time of susceptible banks. We extend a semi-parametric proportional hazards cure model to time-varying covariates and we propose a penalized-likelihood variable selection technique. By means of a simulation study, we show how this technique performs reasonably well. Finally, we illustrate an application to commercial bank failures in the United States over the period 2006-2016.

C0669: **Contagion effects in small business failures: A spatial multilevel autoregressive model**
Presenter: **Raffaella Calabrese**, University of Edinburgh, United Kingdom
Co-authors: Robert Stine
The impact of nearby UK cities characteristics on small business defaults is studied. Credit scoring models would usually rely on city level fixed effects to capture the economic conditions of the cities where SMEs are located. However, this method ignores the contagion effects given by network ties between neighbouring cities. To include both contagion effects between and within cities, we propose a Bayesian multilevel model for binary data. We apply this model to data on SMEs located in the five biggest cities in the UK.

C0905: **A multivariate approach to measure the dimension of a bank**
Presenter: **Alessandro Berti**, Urbino University Carlo Bo, Italy
Co-authors: Cinzia Franceschini, Nicola Loperfido
The dimension of a bank is a central issue for practitioners, authorities and academics. In particular, policymakers often rely on the too big to fail theory, asserting that certain financial institutions are so large that their failure would be disastrous to the economic system. We measured banks dimensions by means of the common dominant principal component of several covariance matrices. We found that the proposed measure is stable, reliable and robust. Data are balance sheet ratios of Italian banks collected by the Bank of Italy during several years.

C1020: **Failure of small business enterprises: A competing risks analysis**
Presenter: **Francesca Pierri**, University of Perugia, Italy
Co-authors: Chrys Caroni, Elena Stanghellini
The time until closure of small business enterprises in Umbria, Italy and the factors that influence it, has been previously analyzed by using Cox regression with time-varying covariates. We considered only the event of failure (closure), from any cause. However, different routes to inactivity exist: court-ordered winding-up (790 of the 8999 firms in our data, 65.9% of the 1199 failures); bankruptcy (199 firms, 16.6%); and closure without action by creditors or courts (210 firms, 17.5%). These are competing risks - as if the various causes are racing to be the first to cause failure. The earlier analysis provides a valuable overall picture, but it is also interesting to examine the separate causes, the rates at which they operate and which factors influence them separately. Data for 2008-2013 provided by the Chamber of Commerce of Perugia included the firm's year of foundation, location, legal form and sector of activity. Financial indexes were constructed from annual balance sheets. Macroeconomic variables were obtained from the National Statistical Service. If the firm ceased activity, the date of, and reason for, closure were recorded. We carry out competing risk analysis using both the main regression methods, constructing cause-specific hazards and sub-distribution hazards.

C0400: **Variable selection in estimating bank default**
Presenter: **Marcella Niglio**, University of Salerno, Italy
Co-authors: Marialuisa Restaino, Francesco Giordano
The crisis of the first decade of the 21st century has definitely changed the approaches used to analyze data originated from financial markets. This break and the growing availability of information have lead to revise the methodologies traditionally used to model and evaluate phenomena related to financial institutions. In this context, we focus the attention on the estimation of bank defaults: a large literature has been proposed to model the binary dependent variable that characterizes this empirical domain and promising results have been obtained from the application of regression methods based on the extreme value theory. We consider, as dependent variable, a strongly asymmetric binary variable whose probabilistic structure can be related to the Generalized Extreme Value (GEV) distribution. Further, we propose to select the independent variables through proper penalty procedures and appropriate screenings of the data that could be of great interest in presence of large datasets.

CO244 Room G2 DEPENDENCE, EXTREMES AND ROBUST INFERENCE Chair: Rustam Ibragimov

C0538: **Volatility regression with fat tails**
Presenter: **Jihyun Kim**, Toulouse School of Economics, France
Co-authors: Nour Meddahi

Nowadays, a common practice to forecast integrated variance is to do simple OLS autoregressions of the observed realized variance data. However, nonparametric estimates of the tail index of this realized variance process reveal that its second moment is possibly unbounded. In this case, the behavior of the OLS estimators and the corresponding statistics are unclear. We prove that when the second moment of the spot variance is unbounded, the slope of the spot variance's autoregression converges to a random variable when the sample size diverges. Likewise, the same result holds when one consider either integrated variance's autoregression or the realized variance one. We also characterize the connection between these slopes whether the second moment of the spot variance is finite or not. Our theory also allows for a nonstationary spot variance process. We derive the results for the case of several lags in the autoregressions and multifactor volatility process. A simulation study corroborates our theoretical findings.

C0794: **Closure properties for heavy-tailed distributions: Some recent results**
Presenter: **Remigijus Leipus**, Vilnius University, Lithuania

The aim is to give a short overview of closure properties for heavy-tailed and related distributions, covering regularly, consistently varying, subexponential, long-tailed and other distributions. Recall that the closure property states that, assuming that two or more distributions are in some specific class, their transformation belongs to the same class of distributions. The main attention is focused to the sum-, mixture, min-, max- and power-convolution closure properties. Additionally, we present some recent results on random sum- and random max-closure properties for dependent and nonidentically distributed heavy-tailed random variables.

C1029: **Extreme returns and structural breaks in the Russian financial market**
Presenter: **Oleg Lebedev**, Innopolis University, Russia
Co-authors: Andrei Ankudinov

Structural breaks in the relationship between Russian financial indicators (the RTSI stock index, the rouble exchange rate and the Mosprime money market daily rate) and oil prices over the period of 2012-2015 are evaluated. To estimate the break dates, we employ a dynamic programming approach based on global minimizers of the sums of squared residuals and generalized fluctuation tests with moving estimates. The obtained results show that the 2014 sharp decline in stock prices and the rouble exchange rate and rise in interest rates cannot be explained by the fall of oil prices only. The peak of decoupling between the financial indicators and oil prices coincides with the period of the harshest anti-Russian sanctions.

C1121: **One country, two systems: The heavy-tailedness of Chinese A- and H- share markets**
Presenter: **Zhimin Chen**, Swiss Finance Institute and University of Lausanne, Switzerland
Co-authors: Rustam Ibragimov

Chinese A- and H- share markets operate in different institutional environments (emerging/developing v.s. developed) and thus may have different tail risk properties. The focus is on the analysis of heavy-tailedness properties of these two markets using recently developed robust inference methods. The equality of tail indices of returns for A and H dual-listed companies cannot be rejected, and some A- and H- share returns may have infinite second moments. Their heavy-tailedness properties did not change significantly with respect to the 2008 financial crisis and the date when the corresponding company starts to be dual-listed.

C1163: **Robust inference for predictive regressions under endogeneity and heteroskedasticity**
Presenter: **Anton Skrobotov**, Russian Presidential Academy of National Economy and Public Administration and Innopolis University, Russia
Co-authors: Rustam Ibragimov, Jihyun Kim

New simple approaches are proposed to robust inference in predictive regressions. The first approach is based on previous results. First, we utilize instrumental variable estimators such as Cauchy estimator to establish the asymptotic normality of the estimator regardless of the order of integration of the variables in regression models and endogeneity. Second, we show that, under general conditions, robust inference on unknown parameters of interest under heterogeneity and dependence may be conducted by partitioning the data into some number of groups and performing the standard t -test with asymptotically normal parameters group estimates and the critical values of Student-t distributions. The second approach is based on the fact that the limiting volatility process can be estimable as precise as possible asymptotically. Therefore, we can either correct directly the time series using the volatility estimate or use the time change method. These approaches provide standard normal inference in the limit. The proposed approaches to robust inference compare favorably with widely used inference procedures in terms of its finite sample properties and can be used under different general settings with dependence, volatility clustering, heavy tails and potentially nonstationary volatility observed in the real-world financial and economic markets.

CO096 Room N2 NEW METHODS FOR NONLINEARITIES IN TIME SERIES PANELS AND APPLICATIONS Chair: Peter Pedroni

C1088: **Nonlinear effects of inflation expectations on durable consumption: The role of household balance sheets**
Presenter: **Johannes Schuffels**, Maastricht University, Netherlands
Co-authors: Lenard Lieb

Research interest in the reaction of consumption to expected inflation has increased sharply in recent years due to efforts by central banks to kick start demand through higher inflation expectations. We contribute to this literature by analyzing whether various components of households balance sheets determine how consumption reacts to expected inflation. Theoretically, many channels are conceivable: an increase in inflation expectations can raise consumption through direct increases in expected wealth, e.g. for households with nominal debt contracts. By affecting the real interest rate, expected inflation can interact with wealth if only those households can adapt their consumption to current real interest rates that are not credit constrained or sufficiently liquid to shift funds between consumption and savings. To empirically assess these nonlinearities in the consumption effects of expected inflation we apply a lasso-based post-double variable selection model to a large dimensional household survey conducted by the Dutch central bank. The model selects those interactions between expected inflation and a large set of household balance sheet components that explain realized vehicle expenditures. By allowing for a wide range of functional forms of these interactions, we specifically investigate possible nonlinearities.

C0568: **Granger causality test in high dimensional VAR models: A post-double-selection procedure**
Presenter: **Luca Margaritella**, Maastricht University, Netherlands
Co-authors: Stephan Smeekes, Alain Hecq

An asymptotic F-test procedure is developed in order to test for Granger causality in high-dimensional VAR models. A post-double-selection setup is outlined and the asymptotic properties of the test statistics are studied. The extensive Monte Carlo simulations compare different ways of choosing the right tuning-parameter. Positive performances of the proposed procedure in highly-parametrized scenarios are found. The routine is applied to investigate the money-income causality relation using the FRED-QD dataset.

C1065: **Nonlinearities in financial development**
Presenter: **Peter Pedroni**, Williams College, United States
Co-authors: Diala Al Masri
New data and a fundamentally new and robust panel time series approach is employed to reexamine the nonlinear relationship between financial development and long run levels of per capita income and the implied relationship between financial innovation and economic growth. In support of our approach, we use a Schumpetarian growth model to motivate our thinking about the complex nature of the nonlinear relationships inherent in the process of financial development. The new empirical approach reveals that, consistent with such a modeling framework, financial development typically encompasses both convex and concave relationships between rates of financial innovation and economics growth, which can be accentuated or diminished depending on the mix of financial attributes and the presence of various economic conditions. We find that growth in the depth and general access to financial institutions is key to enhancing the convex, increasing returns to financial development, while the development of financial markets plays a supporting role, which can, however, lead to concave, declining returns when the rate of financial market growth is too slow or too fast relative to institutional growth. Furthermore, we find that the rate at which a country develops its domestic financial sectors relative to the rate at which it becomes financially open to the global economy can be important in determining whether the convex or concave aspects of financial development are accentuated.

C0516: **Inequality overhang**
Presenter: **Francesco Grigoli**, International Monetary Fund, United States
Co-authors: Adrian Robles
The linearity of the relationship between income inequality and economic development has been long questioned. While theory provides arguments for which the shape of relationship may be positive for low levels of inequality and negative for high ones, most of the empirical literature assumes a linear specification finding conflicting results. Employing an innovative empirical approach robust to endogeneity, we find pervasive evidence of nonlinearities. Similar to the debt overhang literature, we identify an inequality overhang level in that the slope of the relationship between income inequality and economic development switches from positive to negative. We also find that in an environment characterized by widespread financial inclusion and high income concentration, rising income inequality has a larger negative impact on economic development. On the positive side, a sufficiently high female labor participation can act as a shock absorber reducing such negative impact.

C0855: **The marginal product of private and public capital**
Presenter: **Alvar Kangur**, International Monetary Fund, United States
Co-authors: Francesco Grigoli, Peter Pedroni
Nonlinearities in private and public capital in determining the level of output are explored. Relying on a new dataset spanning 150 countries over more than three decades and a recently developed methodology, we draw conclusions on the optimal level of capital and therefore on the extent to which countries over- or under-invest. By reinterpreting the relationship between capital stocks and income levels in terms of marginal effects, we uncover nonlinearities in the marginal productivity of capital at different levels of capital stocks and analyze how they evolved over time. Furthermore, interacting the nonlinear relationship with theoretically complementary factors, such as quality of institutions, efficiency of public investment, proxies for international credit frictions, among others, we identify conditions that are conducive to higher marginal productivity of capital and consequently to endogenous growth. Finally, we also draw implications on reasons why capital does not flow from rich to poor countries where capital ratios are lower.

CO092 Room P2 BAYESIAN ECONOMETRICS Chair: Kaoru Irie

C0546: **High-frequency stochastic volatility models for the Japanese stock index**
Presenter: **Toshiaki Watanabe**, Hitotsubashi University, Japan
Co-authors: Jouchi Nakajima
A high-frequency stochastic volatility (SV) model is proposed for the Japanese stock index. Apart from the standard daily-frequency SV models, high-frequency SV models are fit to intraday returns by extensively capturing intraday volatility patterns. The proposed model consists of the persistent autoregressive stochastic volatility process, seasonal components of the intraday volatility patterns, and correlated jumps in prices and volatilities. A Bayesian method for the analysis of this model is developed using Markov chain Monte Carlo (MCMC) with the exact multi-move sampler for the SV process. Using this method, the proposed model is applied to the 5-minute returns of Nikkei 225 index. It is also examined whether the high-frequency SV model improves the predictive ability of volatility compared with the commonly-used realized volatility.

C0603: **Bayesian analysis of dependent functional data**
Presenter: **Silvia Montagna**, University of Turin, Italy
Co-authors: Surya Tokdar, Irina Irincheeva
The rapid evolution of data collection technologies has permitted vast quantities of data to be recorded densely over time or space. These data are usually regarded as error-prone measurements of underlying smooth functions, thus the name of functional data. A new Bayesian methodology is proposed for dependent functional data analysis. Dependency manifests across multiple trajectories at any given time point and, potentially, also temporally within each trajectory (e.g. gene expression trajectories). To accommodate for the dependence across curves, we will focus on hierarchical Bayesian factor models. Given the high dimensionality of the data, it will also be necessary to ensure parsimony and computational tractability of the proposed methods. We will explore the use of sparsity-inducing priors, such as latent threshold priors, for certain model parameters. The proposed methods will be implemented on real data applications.

C0676: **Realized stochastic volatility models with skew-t distributions**
Presenter: **Makoto Takahashi**, Hosei University, Japan
Co-authors: Yasuhiro Omori, Toshiaki Watanabe
Predicting volatility and quantiles of financial returns is essential to measure the financial tail risk such as value-at-risk and expected shortfall. There are two important aspects of volatility and quantile forecasts: the distribution of financial returns and the estimation of the volatility. Building on the traditional stochastic volatility model, the realized stochastic volatility model incorporates the realized volatility as the precise estimator of the volatility. Using the generalized hyperbolic skew-t and Azzalini skew-t distributions, the model is extended to capture the well-known characteristics of the return distribution, namely skewness and heavy tails. In addition to the normal and Student's t distributions included as the special cases of both distributions, the Azzalini skew-t contains the skew-normal, and hence allows flexible modeling of the return distribution. The Bayesian estimation scheme via a Markov chain Monte Carlo method is developed and applied to the US and Japanese stock indices, Dow Jones Industrial Average and Nikkei 225. The estimation results show that the negative skewness is evident for both indices whereas the heavy tail is largely captured by the realized stochastic volatility, and thus demonstrate that the model with the skew-normal distribution performs well. In addition, the prediction results with a range of tests and performance measures evaluating the volatility and quantile forecasts will be presented.

C1018: **Filtering for stochastic volatility models with leverage by mixture approximation**
Presenter: **Kaoru Irie**, University of Tokyo, Japan
Co-authors: Yasuhiro Omori, Naoki Awaya

The approximation of non-Gaussian distributions by finite mixture of normals is commonly used in Bayesian analysis of macroeconomic and financial time series models that are typically state space models with non-Gaussian observations such as stochastic volatility models. We apply this approximation to the SV models to realize its sequential analysis by the customized version of the existing sequential Monte Carlo methods of particle filtering/learning. The leverage parameter, which is the correlation of volatility and observation, can be sequentially sampled with the other parameters, utilizing the conditional normality of posteriors available under the mixture approximation. The approximation bias is also sequentially corrected through particle filtering. The proposed computational method is illustrated by the sequential analysis of the univariate and multivariate SV models with simulation and real data.

C0607: **An economic crisis indicator for emerging economies: Short term private external debt**
Presenter: **Oya Ekici**, Istanbul University, Turkey

The economic crisis analysis remains at the center stage of the economic policy discussions. Short term private external debt has crucial potential to foresee the crisis in emerging economies. A dynamic linear growth model and a Bayesian estimation method have been recently used to fit short term private external debt data. The model was outperformed to capturing unstable terms of the Turkey's economy. We move to the next step with the motives of producing an early warning system based on this model and searching its potential for other emerging economies. We develop a signal detecting indicator, which we inspired from the KLR model. We focus on 17 emerging economies data covering the period of 1998 to 2017 and we estimate an index based on the models estimated parameters and test if the index is successful in detecting the signal of the crisis terms. Our empirical findings reveal that the procedure has given signals for the crisis in considered period of the emerging economies data. Thus, we have more evidence that a signal detecting indicator serves for emerging economies.

CC652 Room Q1 CONTRIBUTIONS IN APPLIED ECONOMETRICS Chair: Roderick McCrorie

C0215: **An empirical investigation of credit cycles in advanced economies**
Presenter: **Mohammad Jahan-Parvar**, Federal Reserve Baord of Governors, United States
Co-authors: Daniel Beltran, Fiona Paine

Monetary authorities use the credit gap (devotions of private credit to GDP from its long-term trend) as a policy instrument. The size of credit gap varies based on the detrending method used. The credit gap is the accepted indicator of credit cycle. Credit cycles and business cycles do not coincide, the typical credit cycle is longer than a business cycle. The behavior of macroeconomic and financial quantities in various phases of the credit cycle, are not widely studied. BIS advocates using credit gaps as early warning indicators for financial distress and using a threshold for accuracy. The indicator is accurate if it flashes on only ahead of crises that do materialize. In practice, the threshold is set to maximize accuracy while capturing at least $2/3$ of crises since WWII. It must be reliable in real time: studies find that ex-post revisions to the credit gap are as large as the gap itself. Using 7 detrending methods, we derive credit gaps that optimally weigh accuracy (noise-to-signal ratio), relevance (fraction of crisis captured), and reliability (stability of real-time to full-sample estimates). Once we detect suitable detrending methods based on these criteria, we study the behavior of financial and macroeconomic quantities in various phases of the credit cycle across 17 economies. We then perform a Markov-switching VAR study of responses of financial and macroeconomic variables to financial or business cycle shocks, conditional on phases of the credit cycle.

C1663: **Elliptical subset VAR estimation and impacts on frequency causality measures**
Presenter: **Thibault Soler**, University Paris 1 – Pantheon-Sorbonne, France
Co-authors: Emmanuelle Jay, Christophe Chorro, Philippe De Peretti

Granger non-causality tests have received a great deal of attention over recent years, especially in economics, finance and neuroscience. Nevertheless, they report no information about the causal strength, which is often required. To correct this aspect, several measures in frequency domain has been proposed. These approaches are two-step ones consisting in first estimating a Vector AutoRegressive model (VAR), and then computing coherence of transfer function. While being very appealing, this two-stages procedure may suffer from an inaccurate estimation of the VAR coefficients, in particular for heavy-tailed time series, short sample size, or highly correlated innovations. We propose to focus on heavy-tailed estimation, which is more appropriate to modelling economic or financial time series, and our goal is twofold: first, extend Gaussian subset VAR estimation to elliptical using robust covariance matrix. Then, evaluate the impact on generalized partial directed coherence. The method uses Tyler's covariance matrix estimator and Yule-Walker equations to estimate VAR coefficients, which allows to reduce the estimation error of the coefficients due to extreme values. The empirical performance is demonstrated using Monte-Carlo simulations with different kind of multivariate systems and innovation assumptions (distribution, covariance structure, etc.), and the results obtained are then compared with those of other covariance matrix estimators (SCM, Q_n estimator, etc.).

C1538: **The impact of oil price volatility on the exchange rate in Russia**
Presenter: **Artem Aganin**, National Research University The Higher School of Economics, Russia
Co-authors: Anatoly Peresetsky

It is commonly discovered in literature that USD/RUB exchange rate depends on oil prices, but USD/RUB volatility does not attract same attention. The aim is to model dependence of USD/RUB volatility on oil price volatility. Another goal is to analyse USD/RUB volatility dependence from the potential macroeconomic factors, such as sanctions and Russian Central bank actions. One-dimensional GARCH models and two-dimensional VAR-BEKK models are used for the volatility modelling. Results from modelling suggest that USD/RUB volatility depends not only on oil volatility, but also on the macroeconomics factors. Sanctions increase the exchange rate volatility, but their impact decrease with time. Impact of oil volatility on the exchange rate volatility is higher at the periods with low oil prices.

E1660: **Modelling volatility in daily air temperature on Svalbard**
Presenter: **Sondre Holleland**, University of Bergen, Norway
Co-authors: Hans Arnfinn Karlsen

A lot of ink has been devoted to showing positive trends in air temperature over the last decades due to global warming and climate change. The effects are especially clear in the Arctic where the reduction in sea ice is a big issue. By considering daily average air temperature measurements from Svalbard airport dating back to 1975, we develop a model for the day-to-day conditional volatility. The model captures seasonal effects and a significant systematic decrease in the logarithmic volatility over the last four decades. A two-step iterative scheme is used between a mean model and the volatility model until convergence of the parameter estimates. The volatility model is related to exponential GARCH with seasonal time varying parameters and a linear trend in the logarithmic conditional volatility. We use the Template Model Builder (TMB) package in R to fit the model by maximum likelihood.

C1738: **Structural estimation of time-varying spillovers: An application to international market liquidity**
Presenter: **Lukas Boeckelmann**, Paris School of Economics, France
Co-authors: Arthur Stalla-Bourdillon
A structural version of the popular Diebold-Yilmaz spillover framework based on a single comprehensive empirical approach is proposed. Key to our approach is a SVAR-GARCH model that is identified by heteroskedasticity and allows for the construction of time-varying up-to-date forecast error variance decompositions. Building on these advances, we analyze the degree, time variation, direction and determinants of market liquidity spillovers on international equity markets. We find that liquidity spillovers (i) have a significant though time-varying impact (ii) are stronger between stock markets with more correlated fundamentals and (iii) increase in times of high risk aversion.

CC648 Room H2 CONTRIBUTIONS IN FORECASTING II Chair: Fallaw Sowell

C1332: **Regime-specific exchange rate predictability and the role of uncertainty**
Presenter: **Robinson Kruse-Becher**, University of Cologne, Germany
Co-authors: Joscha Beckmann
Exchange rates are notoriously difficult to predict. The benchmark findings that fundamental exchange rate models are unable to outperform naive benchmark forecasts has not been systematically overturned. The underlying reason is that predictability is strongly varying over time. Against this background, we contribute to the literature by analyzing whether regime-specific predictability can be traced back to specific highly persistent variables. We apply a threshold framework with two regimes, namely predictability and no predictability. This enables us to consider a model in which the switch between regimes is determined by an observable variable. In line with recent findings, we investigate several uncertainty and expectation measures as transition variables between the regimes of predictability and no predictability. Besides relying on established measures, we additionally obtain uncertainty from forecaster disagreement about real GDP, inflation, interest rates and the current account. Moreover, we extract yield curve factors. As exchange rate predictors for G10 currencies, we study persistent deviations from uncovered interest rate parity, purchasing power parity, the classical monetary exchange rate model and an asymmetric Taylor rule. Our findings suggest that various threshold effects are responsible for episodes of predictability with interest rate uncertainty being of special importance.

C1684: **Expectation formation, financial frictions, and forecasting performance of dynamic stochastic general equilibrium models**
Presenter: **Christoph Schult**, Halle Institute for Economic Research, Germany
Co-authors: Oliver Holtemoeller
The forecasting performance of estimated basic dynamic stochastic general equilibrium (DSGE) models is documented and compared to extended versions which consider alternative expectation formation assumptions and financial frictions. We also show how standard model features, such as price and wage rigidities, contribute to forecasting performance. It turns out that neither alternative expectation formation behaviour nor financial frictions can systematically increase the forecasting performance of basic DSGE models. Financial frictions improve forecasts only during periods of financial crises. However, traditional price and wage rigidities systematically help to increase the forecasting performance.

C0191: **Computation of reliable interval forecast for dynamic averaging of economic time series regression models**
Presenter: **Nikita Moiseev**, Plekhanov Russian University of Economics, Russia
A method to obtain reliable confidence intervals when conducting the dynamic averaging procedure for linear regression models of economic time series is presented. It is explicitly shown that, when we adapt models' weights each time new data occurs, traditional approach to computing an unbiased estimator of errors' variance systematically underestimates the true variance. Traditional approach is valid only in case of static weights, when optimization procedure is conducted just once. The same conclusion holds for dynamic and static specification procedure as well. However, when applying static weights to time series processes model accuracy is usually lower than for dynamically adapted weights. Thus, we make an attempt to solve this conundrum concerning the choice between either better prediction (adaptive weights) or reliable confidence intervals (static weights). To achieve this goal, we work out a substantial solution for obtaining an adjusted estimator of true errors' variance that yields a reliable result. Such an estimator is proposed to be computed numerically by simulating the errors' variance-covariance matrix by Wishart distribution with a prior equal to a sample variance-covariance matrix and, thus, obtaining the average bias of traditional estimator. To prove the efficiency of proposed approach, we also conduct a rigorous out-of-sample simulation and empirical testing.

C0573: **Out-of-sample performance of nonlinear models in commodities international price differential forecasting**
Presenter: **Nicola Rubino**, University of Barcelona, Spain
An analysis of a group of small commodity exporting countries' price differentials relative to the US dollar is presented. Using unrestricted self exciting threshold autoregressive models (SETAR), we model and evaluate sixteen national consumers' price index (CPI) differentials relative to the US dollar CPI. Out-of-sample forecast accuracy is evaluated through calculation of mean absolute errors measures on the basis of monthly rolling window and recursive forecasts and extended to three additional models, namely a logistic smooth transition regression (LSTAR), an additive nonlinear autoregressive model (AAR) and a simple neural network model (NNET). Our preliminary results confirm presence of some form of non linearity in the majority of the countries analyzed. The parsimonious AR(1) model does not appear to perform any worse than any nonlinear model in the rolling sample exercise. However, its validity in terms of a long run equilibrium driven by purchasing power parity is undermined by the results of the recursive estimates and the outcome of the Diebold-Mariano type tests, which more generally favor the Heckscher commodity points theory. As a policy advice to commodity exporting countries, we find no apparent reason to suggest commodity export price pegging as a generalized foreign exchange policy.

C1639: **Forecasting U.S. bank failures with machine learning techniques**
Presenter: **Alexander Kostrov**, University of St. Gallen, Switzerland
Co-authors: Lyudmila Grigoryeva, Stefanie Bertele
After consolidation and waves of failures, there are still about 5000 operating banks in the United States. Many of them remain weak and enter the official problem bank list. The main question is whether a Support Vector Machine (SVM) beats well-established models used to predict bank failures (namely, Naive Bayes classifier, discriminant analysis, and logit model). Using FDIC call reports with bank-specific statistics for 2004-2016 we construct a set of CAMELS proxies to predict failures. We consider the class imbalance problem in data, that is, when one class of observations is severely undersampled. We apply the synthetic minority oversampling technique (SMOTE) to solve the problem. The use of SMOTE brings a statistically significant improvement in the forecasting performance of all four models. It means that the class-imbalance problem must be addressed and mitigated in predicting bank failures. SVM is found to be very competitive in comparison to three classical classifiers. It is more accurate for different forecasting horizons. In line with the literature, improvements in classifying historical bank failures are likely to remain in the future.

CC649 Room I2 CONTRIBUTIONS IN ECONOMETRICS MODELLING Chair: Mikko Pakkanen

C1579: **A Bayesian analysis of extended Poisson distribution**
Presenter: **Haruhiko Shimizu**, Kobe University, Japan

A Bayesian analysis of extended Poisson distribution is considered, which can be applied to non-negative integers as well as the negative integers. The distribution has two parameters, lambda, which is related to the mean of the distribution, and p, which is the ratio of the positive integers. Maximum likelihood estimators of these parameters are analytically solved and shown that they are asymptotically independent. We can check the independence of the parameters by computing the correlation matrix of the parameters. However, if we try to compute the maximum likelihood estimators of these two parameters jointly, we often find that likelihood function as well as the parameters does not converge. We try to apply the Bayesian analysis and compute the parameters using Markov chain Monte Carlo, and compare with the maximum likelihood estimation. Based on the distribution histogram of the extended Poisson distribution, we use bimodal distribution as a proposal density. Poisson regression model can explain the non-negative integers. Using the extended Poisson distribution, we are able to use all the integers as explained variable of the regression model. As an extension, we consider the extended Poisson regression model.

C1600: **Stochastic frontier model choice with unobserved heterogeneity: A Monte Carlo study**
Presenter: **Antonio Carvalho**, Heriot-Watt University, United Kingdom
Co-authors: Jan Ditzen

Stochastic frontier models often assume the existence of a one-sided error term with an economic interpretation regarding technical or cost efficiency measurement. The true random and true fixed effects stochastic frontier models are popular solutions for dealing with unobserved heterogeneity in this context. Given the nature and assumptions of the models, it is natural to consider the Hausman test to make decisions on which model to use, as is done in random effects vs fixed effects models in panel data econometrics. However, in these specific models, an unbiased estimate of the efficiency term is potentially more important than an efficient estimate. We argue that using the Hausman test to make a decision on which model to use can potentially lead to an incorrect choice. A Monte Carlo simulation is set up to assess if the correlation between true and estimated efficiencies and the preservation of original efficiency rankings can be used as selection criteria, particularly if they are consistently favorable to the less restrictive fixed effects model, independently of the data generating process.

C1405: **The welfare-adequate measure of wage dispersion in the basic DSGE model with unemployment**
Presenter: **Przemyslaw Wlodarczyk**, University of Lodz, Poland

The emergence of the new class of DSGE models enabled assessment of the consequences of unemployment existence for the levels of aggregate welfare observed in an economy, as well as inference on the welfare implications of different strategies of monetary policy execution. However, as the analytical expressions of the welfare loss function are readily available only for relatively simple models, which exclude such complications as openness of the economy together with the role of exchange rates, in the majority of admissible and practically important cases we still have to resort to the numerical welfare comparisons in the vein of previous work. A welfare-adequate measure of wage dispersion in a standard DSGE model with wages set according to a previous formalism and unemployment introduced as in the literature is derived. The resulting social welfare function is then carefully analysed and welfare implications of unemployment are assessed using certain measures of welfare cost. We further compare the accuracy of the results with those obtained using the traditional Welfare Loss Function (WLF) based on the second-order approximation of the social welfare function around the steady state. The robustness of the results is checked by means of comparison with those obtained under different calibrations of the model and different monetary policy rules.

C0417: **Modelling unbalanced catastrophic health expenditure data**
Presenter: **Songul Cinaroglu**, Hacettepe University, Turkey

Traditional parametric statistical learning methods such as logistic regression (LR), perform poorly at predicting class-imbalanced data. Random Forest (RF) is an algorithmic statistical method to deal with unbalanced data. We compare performances of LR and RF classifiers predicting households faced with catastrophic out-of-pocket (OOP) health expenditure, while using a balanced oversampling procedure. Data came from nationally representative household budget data from the Turkish Statistical Institute for the year 2012. The number of households for which the surveys were valid was 9987 for the year 2012. WHOs methodology was employed to calculate catastrophic OOP health expenditure. The degree of imbalance is higher and the percentage of households faced with catastrophic OOP health expenditure is 0.14%. LR and RF models are compared based on eight common risk factors. A balanced oversampling was used and 31 artificial datasets were generated changing from 5% and 98% of original data size. Accuracy, sensitivity, specificity, precision and F-measure were used to evaluate classifiers. ROC curve was used to compare the performance of the classification models. Balanced oversampling data has more accurate predictions and RF is superior to identify households faced with catastrophic OOP health expenditure.

C1351: **Reducing model risk using Bayesian approach: Application to PD modelling of mortgage loans**
Presenter: **Zheqi Wang**, University of Edinburgh, United Kingdom
Co-authors: Jonathan Crook, Galina Andreeva

A new Bayesian informative prior selection method is proposed to reduce model risk of ignored information and improve model performances. We use logistic regression to model the probability of default of mortgage loans using both Bayesian approach with various priors and frequentist approach. In the Bayesian informative prior selection method we propose, we treat coefficients in the PD model as time series variables. We build ARIMA models to forecast the coefficient values in future time periods and use these ARIMA forecasts as priors. We find that the informative Bayesian models using this prior selection method outperform both frequentist models and Bayesian models with other priors in terms of model performances.

CG622 Room E2 CONTRIBUTIONS IN VALUE-AT-RISK Chair: Matei Demetrescu

C1710: **Forecasting value at risk and expected shortfall using a dynamic omega ratio**
Presenter: **James Taylor**, University of Oxford, United Kingdom

Value at Risk (VaR) and expected shortfall (ES) have become the standard measures of market risk. The recent development of joint scoring functions for the two measures enables joint models to be estimated. Previous work has shown promising results when an autoregressive model is used for the VaR, and the ES is modelled as the product of the VaR and a constant factor. We propose a time-varying multiplicative factor. It has previously been shown that the ES can be expressed as the product of an expectile and a constant multiplicative factor, which is a function of the expectile level. We rewrite this as the product of a quantile and a multiplicative factor that is a function of a time-varying expectile level. The expectile level is itself a simple function of the omega ratio, which is the ratio of expected gains to expected losses. This leads us to propose a new joint model in which the ES is modelled as the product of the VaR and a factor that is a function of a time-varying omega ratio, which we model using autoregressive expressions for the expected gain and expected loss. We provide empirical illustration using stock index returns.

C1546: **Realized GARCH model adding robust measures of skewness and kurtosis**
Presenter: **Cesar German Santamaria**, Universidad Nacional de San Martin, Argentina

Some researchers have started to incorporate higher moments into their volatility models, e.g., the GARCHSK model, that considers the conventional measures of the sample skewness and kurtosis based on daily data. With the availability of high-frequency data, the estimation of volatility has moved from traditional daily models to realized models. This shift aimed to provide more accurate short-term risk models. One of them is the RGARCH model. Following this trend, we propose the Realized Generalized Autoregressive Conditional Heteroskedasticity Robust Skewness and Robust Kurtosis (RGARCHRSRK) model, which incorporate not only the realized measure of volatility, but also robust measures of skewness and kurtosis, where the standardized residuals follow a modified Gram-Charlier expansion. We found empirically that the proposed model is statistically significant and empower the estimation of parametric Value at Risk (VaR) in comparison with the other RGARCH, GARCHSK and GARCH base models.

C1311: **Forecasting risk measures using intraday and overnight information**
Presenter: **Paula Tofoli**, Catholic University of Brasilia, Brazil
Co-authors: Douglas Gomes dos Santos, Osvaldo da Silva Filho

Risk measures such as value-at-risk (VaR) and expected shortfall (ES) are computed from forecasts of return volatility for the full day. When dealing with high-frequency data from markets which operate during a reduced time (e.g., six to seven hours a day), an approach to take into account the overnight return volatility is needed. In this context, we use heterogeneous autoregressions (HAR) to model the variation associated with the intraday activity, with, e.g., realized variance, bipower variation, realized semivariances and signed jump variation as regressors, and to model the overnight return variance, we use augmented GARCH type models. Then, we combine the forecasts from the two types of models to obtain forecasts for the total daily return volatility. We examine the aforementioned procedure in an extensive empirical study using high-frequency data sets (S&P 500 index and five individual stocks), where out-of-sample VaR and ES forecasts are compared with forecasts from traditional approaches. We benefit from recent results regarding the joint elicitability of VaR and ES, being able to assess the relative forecasting performance of the models in a simple manner. The overall results indicate that the combinations of HAR models with augmented GARCH type models generally produce the most accurate forecasts.

C1105: **Semi-parametric dynamic asymmetric Laplace models for tail risk forecasting, incorporating realized measures**
Presenter: **Richard Gerlach**, University of Sydney, Australia
Co-authors: Chao Wang

The joint Value at Risk (VaR) and expected shortfall (ES) quantile regression model is extended via incorporating a realized measure, to drive the tail risk dynamics, as a potentially more efficient driver than daily returns. Both a maximum likelihood and an adaptive Bayesian Markov chain Monte Carlo method are employed for estimation, whose properties are assessed and compared via a simulation study; results favour the Bayesian approach, which is subsequently employed in a forecasting study of seven market indices and two individual assets. The proposed models are compared to a range of parametric, non-parametric and semi-parametric models, including GARCH, realized-GARCH and the joint VaR and ES quantile regression models. The comparison is in terms of accuracy of one-day-ahead VaR and ES forecasts, over a long forecast sample period that includes the global financial crisis in 2007-2008. The results favor the proposed models incorporating a realized measure, especially when employing the sub-sampled realized variance and the sub-sampled realized range.

C1570: **Backtesting expected shortfall via multi-quantile regression**
Presenter: **Ophelie Couperier**, LEO CNRS - ENSAE - CREST, France
Co-authors: Jeremy Leymarie

A new approach to backtest Expected Shortfall (ES) exploiting the definition of ES as a function of Value-at-Risk (VaR) is proposed. The strategy aims at assessing the quality of multiple VaRs jointly along the tail distribution of the risk model, and encompasses the Basel Committee recommendation of verifying two given quantiles. Building on multi-quantile theory, we propose four backtests that focus on parameter estimates of an auxiliary regression model. Monte-Carlo simulations show that our tests are powerful to detect misspecified ES models. We provide an empirical application on S&P500 returns over the period 2007-2012 and demonstrate the good capability of our methodology to identify misleading ES forecasts. Our empirical results show that the detection abilities are higher with more than two quantiles, and should accordingly be taken into account in the current regulatory guidelines.

CG014 Room F2 CONTRIBUTIONS IN EMPIRICAL MACROECONOMICS Chair: Gunter Coenen

C1501: **Forecasting with large Bayesian VARs: On the importance of the prior**
Presenter: **Jamie Cross**, BI Norwegian Business School, Norway
Co-authors: Aubrey Poon, Chenghan Hou

Substantive empirical evidence has shown that large Bayesian VARs can provide better in- and out-of-sample fit compared to smaller scale models. When specifying such models, a multitude of hierarchical shrinkage priors on the autoregressive coefficients have been proposed. We provide a detailed comparison of six of these prior distributions: Minessota, lasso, SVSS, Dirichlet-Laplace, normal-gamma and horseshoe priors. Additionally, we show how each of these priors can be implemented in a stochastic volatility framework in a computationally efficient manner. At the beginning we compare the relative in-sample fit and out-of-sample forecast performance on a frequently used large data set of US macroeconomic and financial variables. The primary result is that the Horseshoe prior dominates all alternatives both in- and out-of-sample. We then conduct a simulation exercise which explores the possible reasons for this improvement.

C1533: **Analyzing asymmetric effects of monetary policy in the Euro Area**
Presenter: **Catalina Martinez Hernandez**, Free University of Berlin / DIW Berlin, Germany
Co-authors: Anton Velinov

After the introduction of the Euro in 1999, one of the main challenges of the European Central Bank (ECB) is to design common monetary policy among a group of countries that are heterogeneous. We assess the different responses of main macroeconomic and financial variables from eleven countries of the Euro Area (EA) to a contractionary monetary policy shock. We consider a Factor Augmented Vector Autoregression (FAVAR) in order to analyze responses in a data-rich environment. In particular, we focus on the reaction of inflation expectations of each country since it is a key component in the planning of monetary policy and in evaluating the credibility of the ECB. We obtain significant differences in the responses of real variables, inflation and inflation expectations between core and periphery countries.

C1559: **Highly frequent oil price shocks**
Presenter: **Fabrizio Venditti**, ECB, Germany
Co-authors: Giovanni Veronese, Fabrizio Venditti

A structural vector autoregression is constructed which uses information from financial markets to decompose daily changes in the price of crude oil into three structural drivers, namely a risk on/off shock, a global demand shock and an oil supply shock. We propose a novel identification strategy that rests on the use of instrumental variables blended with sign and narrative restrictions. Additional identification assumptions ensure

that the response of macro variables to our daily shocks is consistent with that obtained in monthly models commonly used in the macro literature. We find that, while demand shocks were mostly responsible for the increase in the price of oil before the crisis, and for its collapse during the crisis, oil supply shocks had a decisive role in driving down the price of crude oil in the 2014-2016 slump, as well as in its recovery in 2017-2018.

C1584: **Measuring financial cycle time**
Presenter: **Marco Lombardi**, Bank for International Settlements, Switzerland
Co-authors: Marek Raczko, Andrew Filardo
Motivated by the traditional business cycle approach, we explore cyclical similarities in financial conditions over time in order to improve our understanding of financial cycles. Looking back at 120 years of data, we find that financial cycles exhibit behaviour characterised by recurrent, endogenous swings in financial conditions, which result in costly booms and busts. Yet the recurrent nature of such swings may not appear so obvious when looking at conventionally plotted time-series data (that is, measured in calendar time). Using a pioneering framework, we offer a new statistical characterisation of the financial cycle using a continuous-time autoregressive model with time deformation, and test for systematic differences between calendar and financial cycle time. We find the time deformation to be statistically significant, and associated with levels of long-term real interest rates, inflation volatility and the perceived riskiness of the macro-financial environment. We conclude that past financial cycles do not appear to be a series of unique events but rather recurrent, inherent features of financially liberalised, free-market economies.

C1699: **Nowcasting disaggregated trade and real activity: a DFM approach**
Presenter: **Jan Bruha**, CNB, Czech Republic
An empirical model for nowcasting external trade and real activity is introduced. The model is based on model-based trend-cyclical decomposition of time series and the cyclical components are specified using a dynamic principal component model. We propose a stochastic EM algorithm to estimate the model parameters. The proposed approach is illustrated on data of selected Central European countries.

CP001 Room Ground Level Hall POSTER SESSION Chair: Panagiotis Paoullis

C0258: **Prior robustness and convergence analysis for MCMC output based on automated sensitivity computations**
Presenter: **Liana Jacobi**, University Melbourne, Australia
Co-authors: Dan Zhu
Bayesian inference relies heavily on numerical Markov chain Monte Carlo (MCMC) methods for the estimation of the typically intractable high-dimensional posterior distributions and requires specific inputs. We introduce a new general and efficient numerical approach to address important robustness concerns of MCMC analysis with respect to prior input assumptions, a major obstacle to wider acceptance of Bayesian inference, including MCMC algorithm performance (convergence) reflected in the dependence on the chain starting values. The approach builds on recent developments in sensitivity analysis of high-dimensional numerical integrals for classical simulation methods using automatic numerical differentiation methods to compute first order derivatives of algorithmic output with respect to all inputs. We introduce a range of new robustness measures based on Jacobian matrices of MCMC output w.r.t. to the two sets of input parameters, prior parameters and chain starting values, to enable researchers to routinely undertake a comprehensive sensitivity analysis of their MCMC results. The methods are implemented for a range of Gibbs samplers and illustrated using both simulated and real data examples. We show how to address issues of discontinuities that arise in the context of common random variable updates in Gibbs algorithms, the Gamma and Wishart updates.

C1475: **Recursive estimation of multivariate GARCH processes**
Presenter: **Radek Hendrych**, Charles University, Czech Republic
Recursive estimation methods suitable for univariate GARCH models have been recently studied in the literature. They undoubtedly represent attractive alternatives to the standard non-recursive estimation procedures with many practical applications (especially in the context of high-frequency financial data). It might be truly advantageous to adopt numerically effective techniques that can estimate, monitor, and control such models in real time. The aim is to extend this methodology to multivariate GARCH processes by applying general recursive estimation instruments. In particular, the suggested approach seems to be useful for various multivariate financial time series with (conditionally) correlated components. Monte Carlo experiments are performed in order to investigate the proposed algorithms. Moreover, real data examples are also discussed.

C1607: **How sensitive are VAR forecasts to prior hyperparameters: An automated sensitivity analysis**
Presenter: **Dan Zhu**, Monash University, Australia
Co-authors: Liana Jacobi, Joshua Chan
Vector autoregressions combined with Minnesota-type priors are widely used for macroeconomic forecasting. The fact that strong but sensible priors can substantially improve forecast performance implies VAR forecasts are sensitive to prior hyperparameters. But the nature of this sensitivity is seldom investigated. We develop a general method based on a new automatic differentiation approach for MCMC output to systematically compute the sensitivities of forecasts-both points and intervals-with respect to any prior hyperparameters. In a forecasting exercise using US data, we find that forecasts are relatively sensitive to the strength of shrinkage for the VAR coefficients, but they are not much affected by the prior mean of the error covariance matrix or the strength of shrinkage for the intercepts.

C1734: **Bayesian vector autoregressive models for forecasting inflation rate in Nigeria**
Presenter: **Joy Nwabueze**, Michael Okpara University of Agriculture, Umudike. Abia State, Nigeria, Nigeria
Co-authors: Uchechukwu George
Maintaining Price stability is a key function of central banks, there is therefore need for inflation forecasting. The focus is on using multiple time series method for forecasting of inflation rate in Nigerian. The Bayesian approach to the estimation of Vector Autoregressive (VAR) model is applied. This allows combination of prior information and data information. Forecasts of inflation rates in Nigeria are provided by using six different Bayesian VAR priors (Diffuse prior, Minnesota prior, Natural Conjugate prior, Independent Normal Wishart prior, Stochastic Search Variable selection prior-Wishart and Stochastic Search Variable selection prior-VAR). The forecast performance of various models is evaluated using root mean square error (RMSE). The Stochastic search variable selection-Wishart outperforms other methods. The forecast from this model together with the impulse response function is given. It is concluded that the stochastic search variable selection prior-Wishart gives a reliable forecast of Inflation rate in Nigeria and therefore, we recommend it for short term inflation forecasting in Nigeria.

C1745: **Revisiting the dynamics between CDS spreads and equity returns under a nonlinear approach**
Presenter: **Dolores Robles**, Universidad Complutense de Madrid, Spain
Co-authors: Jose-Luis Fernandez-Serrano
The relationship between equity and CDS markets between 2007 and 2018 is analyzed under a non-linear vector autoregression approach. We apply cointegration analysis to test the existence of stable long-term relationships between both markets in which the disequilibrium adjustment process may present nonlinearities. We study daily short and long term relationships between the equity price and the corresponding CDS spread of single-name issuers listed in the US and European equity markets. We study the lead-lag relationships with a non-linear vector error correction model and show that, although both variables respond to disequilibria in the long term relationship, equity returns leads changes in the CS spread.

This result indicates that the price discovery process mostly take place in the equity market. We find that the adjustment to disequilibrium in the long term relationship presents a significant nonlinear component. We also test for short and long term Granger causality between markets in our nonlinear setting. When we consider separately the analysis distinguishing the crisis period (2007 - 2010) and the subsequent period of economic recovery (2011 - 2018) we find that the equity market keep leading the CDS market in both periods, although the CDS has greater predictive ability in the recovery period. Overall, we find nonlinearities in the dynamic comovements and causality between equity and CDS markets that change in periods of market turmoil.

C1741: **An ARDL approach for income inequality: Case studies for France, Greece, UK and USA**
Presenter: **Alexandra Livada**, Athens University of Economics and Business, Greece
Co-authors: Kleanthis Natsiopoulos
The purpose is to explore the cointegrating relationships between the 1% top income share and the macroeconomic factors of credit, education, GDP, inflation, population growth and trade in order to reveal if there is a long-run relationship. This relationship is tested for four different countries: Greece, France, USA and UK. We are trying to see if the income inequality is driven by the same factors and in the same way for such different economies. The popular ARDL bounds test for cointegration is used for this analysis. There are strong indications supporting the existence of such a relationship for the cases of France and Greece. For the case of USA, the test suggests the existence of a long-run relationship but a simple graphical inspection is enough to tell us that this is a false positive alarm (type I error) as this is a degenerate relationship. Finally, for the case of UK a not well-defined model supports the longrun relationship hypothesis but a more carefully designed model is against this decision.

C1750: **Entropy measures in building Markowitz's efficient frontier: Evidence form Warsaw stock exchange**
Presenter: **Rumiana Gorska**, Warsaw School of Economics, Poland
The entropy of a probability distribution is a measure of the uncertainty of the distribution. The purpose is to compare efficient frontier build on different measures of entropy with Markowitz's mean-covariance efficient frontier. For this purpose, the efficient frontier and the corresponding entropy-based frontier is built for 30 corner portfolios. Two non-overlapping periods are considered: from 1.01.2000 to 31.12.2007 and from 1.01.2008 to 30.08.2018.

Sunday 16.12.2018 14:40 - 16:20 Parallel Session N – CFE-CMStatistics

EO546 Room A0 SOCIETAL IMPLICATIONS OF WORK IN STATISTICS AND DATA SCIENCE Chair: Jennifer Hill

E0348: **Ethical considerations to ensure that analytics help, never harm, students**
Presenter: **Amy Laitinen**, New America, United States
Co-authors: Iris Palmer
As higher education grapples with promoting student success using fewer resources, predictive analytics, the use of past data to forecast future outcomes, is a promising solution. But like all powerful tools, it must be used well. New America has conducted research into what it looks like to use predictive analytics ethically. We will present some of the challenges of implementing predictive analytics from recruiting and enrolment through graduation. It will also provide guiding practices to ensure these tools are used ethically.

E0677: **Omitted and included variable bias in tests for disparate impact**
Presenter: **Ravi Shroff**, New York University, United States
Co-authors: Sam Corbett-Davies, Sharad Goel, Jongbin Jung
Policymakers often seek to gauge discrimination against groups defined by race, gender, and other protected attributes. A common strategy is to estimate disparities after controlling for observed covariates in a regression model. However, not all relevant factors may be available to researchers, leading to omitted variable bias. Conversely, controlling for all available factors may also skew results, leading to so-called "included variable bias". We introduce a simple strategy, which we call risk-adjusted regression, that addresses both concerns in settings where decision makers have clear and measurable policy objectives. First, we use all available covariates to estimate the expected utility of possible decisions. Second, we measure disparities after controlling for these utility estimates alone, omitting other factors. Finally, we examine the sensitivity of results to unmeasured confounding. We demonstrate this method on a detailed dataset of 2.2 million police stops of pedestrians in New York City.

E0830: **A novel test for bias in decision-making**
Presenter: **Camelia Simoiu**, Stanford University, United States
Co-authors: Sharad Goel, Sam Corbett-Davies
In the course of conducting traffic stops, officers have discretion to search motorists for drugs and other contraband. Scholars and criminal justice advocates have raised concerns that search decisions are prone to racial bias, but it has proven difficult to empirically evaluate these claims due to well-known limitations of current tests. We develop a novel statistical method for testing for discrimination. Namely, we use a hierarchical Bayesian latent variable model to infer latent race-specific thresholds of evidence that officers apply when deciding to search motorists. On a data set of six million police stops in North Carolina from 2009 to 2014, we find that the threshold for searching blacks and Hispanics is significantly lower than the threshold for searching whites, suggestive of racial discrimination in these interactions.

E1050: **The need for standardized hospital screening systems and metrics to detect child maltreatment**
Presenter: **David Scheinker**, Stanford University, United States
Co-authors: Anneke Claypool, Margaret Brandeau
Maltreatment is one of the leading causes of injury and death of US children. No national standardized, validated protocols or metrics exist for how hospitals should screen for and report child maltreatment. Hospitals screen for child maltreatment with a variety of tools about the efficacy of which little published evidence exists. We review current screening tools and examine the challenges in creating standardized systems and performance metrics. We propose the outline of a system that is broadly standardized yet customizable to the needs and resources of each hospital and community. We illustrate our work with a case study at Lucile Packard Childrens Hospital Stanford. There we create a system to improve screening coverage, create protocols to standardize screening, and develop metrics to assess performance.

EO140 Room Aula 4 STATISTICS MEETS COMPUTING Chair: Binyan Jiang

E1047: **Penalized interaction estimation for ultrahigh dimensional quadratic regression**
Presenter: **Binyan Jiang**, The Hong Kong Polytechnic University, Hong Kong
Quadratic regression goes beyond linear model by simultaneously including main effects and interactions between the covariates. The problem of interaction estimation in high dimensional quadratic regression has received extensive attention in the past decade. We introduce a novel method which allows us to estimate the main effects and interactions separately. Unlike existing methods for ultrahigh dimensional quadratic regressions, the proposal does not require the widely used heredity assumption. In addition, the proposed estimates have explicit formulas and obey the invariance principle at the population level. We estimate the interactions of matrix form under penalized convex loss function. The resulting estimates are shown to be consistent even when the covariate dimension is an exponential order of the sample size. We develop an efficient ADMM algorithm to implement the penalized estimation. This ADMM algorithm fully explores the cheap computational cost of matrix multiplication and hence is much more efficient than existing penalized methods under heredity constraints. We demonstrate the promising performance of our proposal through extensive numerical studies.

E1281: **On generalization and computation of Tukey's depth**
Presenter: **Yiyuan She**, Florida State University, United States
Data depth provides a useful tool for nonparametric statistical inference and estimation but also encounters computational difficulties and scope limitations in modern statistical data analysis. The aim is to focus on the generalization and computation of Tukey's depth for supervised learning in multi-dimensions. A general framework of statistic-driven halfspace depth is presented, and on the basis of its connection to classification and M-estimation, we introduce polished data depth as a subspace pursuit problem. By use of generalized gradients and slack variables, we are able to generalize the concept significantly to accommodate restricted parameter spaces and non-smooth objectives in possibly high dimensions. The new formulation of Tukey's depth enables us to utilize state-of-the-art optimization techniques to develop algorithms with implementation ease and guaranteed fast convergence. Simulations and real data examples demonstrate the efficacy of the proposed methodology in statistical inference and estimation.

E1287: **Meta estimation of normal mean parameter: Seven perspectives of data integration**
Presenter: **Peter Song**, University of Michigan, United States
Data integration has recently drawn considerable attention in the statistical literature. We will present a synergic treatment on the estimation of mean parameter of a normal distribution from seven different schools of statistics, which sheds light on the future development of data integration analytics. They include best linear unbiased estimation (BLUE), maximum likelihood estimation (MLE), Bayesian estimation, empirical Bayesian estimation (EBE), Fisher's fiducial estimation, generalized methods of moments (GMM) estimation, and empirical likelihood estimation (ELE). Their properties of scalability and distributed inference will be discussed and compared analytically and numerically.

E1372: **Scalable Kernel-based variable selection**
Presenter: **Xin He**, Shanghai University of Finance and Economics, China
Co-authors: Junhui Wang, Shaogao Lyu
Variable selection is central to sparse modeling, and many methods have been proposed under various model assumptions. We will present a scalable framework for model-free variable selection in reproducing kernel Hilbert space (RKHS) without specifying any restrictive model. As opposed to most existing model-free variable selection methods requiring fixed dimension, the proposed method allows dimension p to diverge with sample size n. The proposed method is motivated from the classical hard-threshold variable selection for linear models, but allows for general variable effects. It does not require specification of the underlying model for the response, which is appealing in sparse modeling with a large number of variables. The proposed method can also be adapted to various scenarios with specific model assumptions, including linear models, quadratic models, as well as additive models. The asymptotic estimation and variable selection consistencies of the proposed method are established in all the scenarios. If time permits, the extension of the proposed method beyond mean regression will also be discussed.

EO450 Room Aula 5 RECENT DEVELOPMENTS IN HIGH-DIMENSIONAL MODELING AND INFERENCE Chair: Young Kyung Lee

E0255: **Semi-parametric hidden Markov model and large scale multiple testing under dependency**
Presenter: **Joungyoun Kim**, Chungbuk National University, Korea, South
Co-authors: Jong Soo Lee, Johan Lim
The optimal procedure for testing many hypotheses under dependence is known theoretically to depend on the local index of significance, called LIS, of each site, the conditional probability that the hypothesis of the site is non-true given the entire observed data. To evaluate the LIS, an assumption should be made for the dependence among observations and the finite state hidden Markov model (HMM) is popularly assumed. We study a two state HMM, denoted by semi-HMM, whose observational distribution for the null state is parametric but that for the non-null state is non-parametric. The main focus of the semi-HMM is on non-null distribution, the observational distribution of the non-null state. The observations from the non-null state are heterogeneous for many unknown reasons and no assumptions are made for the non-null distribution in the proposed semi-HMM. We show that the semi-HMM is not identifiable despite its model flexibility for non-null observations. To estimate the model, we adopt the recent results on the estimation of the semi-parametric mixture model and propose an EM type algorithm. The model and estimation procedure are numerically investigated and compared with existing parametric HMM in the context of multiple testing. Finally, it is applied to two real examples in the literature.

E0459: **Analysis of quantitative high throughput screening data using a robust nonlinear mixed effects model estimation**
Presenter: **Changwon Lim**, Chung-Ang University, Korea, South
Co-authors: Chorong Park
Quantitative high throughput screening (qHTS) data is used to assess the toxicity of a number of chemicals in short period of time by collectively analyzing them at several concentrations. The qHTS data can be analyzed by using a nonlinear mixed effects model that considers both intra-individual variability and inter-individual variability. Since qHTS data generated by repeating the same experiment several times for each chemical, it mainly analyzed using a nonlinear mixed model. In the nonlinear mixed model, one outlier can distort parameter estimates within each individual or overall estimates. We apply a one-step approach which is one of the robust estimation methods to estimate the fixed effect parameters and the variance-covariance structure. In addition, toxic chemicals were classified based on the significance of a parameter which means efficacy of the drugs.

E0728: **Ensemble estimation and variable selection with semiparametric regression models**
Presenter: **Sunyoung Shin**, University of Texas at Dallas, United States
Co-authors: Yufeng Liu, Stephen Cole, Jason Fine
Scenarios are considered in which the likelihood function for a semiparametric regression model factors into separate components, with an efficient estimator of the regression parameter available for each component. An optimal weighted combination of the component estimators, named an ensemble estimator, may be employed as an overall estimate of the regression parameter, and may be fully efficient under uncorrelatedness conditions. This approach is useful when the full likelihood function is difficult to maximize but the components are easy to maximize. As a motivating example, we consider proportional hazards regression with prospective doubly-censored data, in which the likelihood factors into a current status data likelihood and a left-truncated right-censored data likelihood. Variable selection is important in such regression modelling but the applicability of existing techniques is unclear in the ensemble approach. We propose ensemble variable selection using the least squares approximation technique on the unpenalized ensemble estimator, followed by ensemble re-estimation under the selected model. The resulting estimator has the oracle property such that the set of nonzero parameters is successfully recovered and the semiparametric efficiency bound is achieved for this parameter set. Simulations show that the proposed method performs well relative to alternative approaches. Analysis of the multicenter AIDS cohort study illustrates the practical utility of the method.

E1006: **Panel nonparametric MIDAS model: A clustering approach**
Presenter: **Yeonwoo Rho**, Michigan Technological University, United States
Co-authors: Yun Liu
The mixed data sampling regression (MIDAS) models are developed to handle different sampling frequencies in one regression model, preserving information in the higher sampling frequency. While a parametric MIDAS model provides a parsimonious way to summarize information in high frequency data, one parametric form may not necessarily be appropriate for all cross-sectional subjects. In the effort to identify groups in a panel data setting involving mixed frequencies, a flexible MIDAS model is proposed using a nonparametric approach. This nonparametric MIDAS model is further extended to a panel setting using a penalized regression idea. The estimated parameters can then be clustered using traditional clustering methods. The proposed clustering algorithm delivers reasonable clustering results both in theory and in simulations, without requiring prior knowledge about the true group membership information.

EO228 Room Aula B CHALLENGE AND NEW METHODS OF BIG DATA ANALYSIS Chair: HaiYing Wang

E0628: **Recent developments in variable selection and classification with presence-only data**
Presenter: **Garvesh Raskutti**, University of Wisconsin-Madison, United States
The problem of variable selection and classification in the context of presence-only responses is addressed. Such data naturally arises in biological applications due to the high-throughput sequencing technology used. We discuss issues of estimation, inference and debiasing, and optimization relating to this problem. In particular, the imperfect labels lead to a non-convex objective which presents both statistical and optimization issues. We address these challenges, present algorithms with statistical guarantees and validate our approach on a real data application.

E1012: **Bayesian shrinkage towards sharp minimaxity**
Presenter: **Qifan Song**, Purdue University, United States

Shrinkage prior becomes more and more popular in Bayesian modeling for high dimensional sparse problems due to its computational efficiency. Recent works show that a polynomially decaying prior leads to satisfactory posterior asymptotics under regression models. In literature, statisticians have investigated how the global shrinkage parameter, i.e., the scale parameter, in a heavy tail prior affects the posterior contraction. We explore how the shape of the prior, or more specifically, the polynomial order of the prior tail affects the posterior. We discover that, under sparse normal means models, the polynomial order does affect the multiplicative constant of the posterior contraction rate. More importantly, if the polynomial order is sufficiently close to 1, it will induce the optimal Bayesian posterior convergence, in the sense that the Bayesian contraction rate is sharply minimax, i.e., not only the order, but also the multiplicative constant of the posterior contraction rate are optimal. The above Bayesian sharp minimaxity holds when the global shrinkage parameter follows a deterministic choice which depends on the unknown sparsity s. Therefore, a beta modeling is further proposed, such that our sharply minimax Bayesian procedure is adaptive to unknown s. Our theoretical discoveries are justified by simulation studies.

E1075: **Optimal subsampling algorithms for big data generalized linear models**
Presenter: **HaiYing Wang**, University of Connecticut, United States

To fast approximate the maximum likelihood estimator with massive data, an Optimal Subsampling Method has been previously proposed under the A-optimality Criterion (OSMAC) for logistic regression. The scope of the OSMAC framework is extended to include generalized linear models with canonical link functions. The consistency and asymptotic normality of the estimator from a general subsampling algorithm are established, and optimal subsampling probabilities under the A- and L-optimality criteria are derived. Furthermore, using Frobenius norm matrix concentration inequality, finite sample properties of the subsample estimator based on optimal subsampling probabilities are derived. Since the optimal subsampling probabilities depend on the full data estimate, an adaptive two-step algorithm is developed. Asymptotic normality and optimality of the estimator from this adaptive algorithm are established. The proposed methods are illustrated and evaluated through numerical experiments on simulated and real datasets.

E1376: **Adaptive designs for optimal observed Fisher information**
Presenter: **Adam Lane**, Cincinnati Children's Hospital Medical Center, United States

Expected Fisher information can be found a priori and as a result its inverse is the primary variance approximation used in the design of experiments. This is in contrast to the common claim that the inverse of observed Fisher information is a better approximation of the variance of the maximum likelihood estimator. Observed Fisher information cannot be known a priori; however, if an experiment is conducted sequentially (in a series of runs) the observed Fisher information from previous runs is known. Two adaptive designs are proposed that use the observed Fisher information from previous runs to design the future runs.

EO086 Room Aula Magna RECENT INNOVATION IN MULTI-OMICS DATA ANALYSIS Chair: Taesung Park

E0992: **Inverse model based test robust to population structure in genetic association studies**
Presenter: **Minsun Song**, Sookmyung Women's University, Korea, South

A new statistical test of association between a trait and genetic markers is presented, which we theoretically and practically prove to be robust to arbitrarily complex population structure. The statistical test involves a set of parameters that can be directly estimated from large scale genotyping data, such as those measured in genome-wide associations studies. We also derive a new set of methodologies, called a genotype conditional association test, shown to provide accurate association tests in populations with complex structures, manifested in both the genetic and non-genetic contributions to the trait. We demonstrate the proposed method on a simulation study and on the real data. Our proposed framework provides a substantially different approach to the problem from existing methods.

E1136: **Patient-driven tumor xenograft based gene expression model to predict anti-cancer drug response in cancer patient**
Presenter: **Youngchul Kim**, H. Lee Moffitt Cancer Center and Research Institute, United States

Responses of cancer patients to anticancer drugs vary because of the substantial heterogeneity in molecular characteristics of their tumors. A successful personalized anticancer therapy will then greatly depend on cancer biomarkers that can accurately select patients who will benefit from the drugs. Patient-derived tumor xenograft (PDX) has been widely recognized to inform therapeutic development strategies. We developed a pipeline, PDXGEM, to construct a gene expression model (GEM) for predicting response to anti-cancer drugs in cancer patients on the basis of data on gene expression and dose-response curve from a pan-cancer PDX cohort. As a proof-of-cancer study, we applied the PDXGEM to build GEMs of paclitaxel and cetuximab for breast cancer and colorectal cancer, respectively. For paclitaxel, 66-genes based GEM was built by training the data of 13 breast cancer PDXs and it had a consistently significant prediction performance in multiple breast cancer cohorts. PDXGEM resulted in 882-genes based GEM for cetuximab and its retrospective validation on 70 colorectal cancer patients also yielded a significant prediction (AUC=0.769, p=0.031). PDXGEM can be used to discovery predictive cancer biomarkers and improve therapeutic response rates and therapeutic quality in cancer patients by enriching highly responsive patients.

E1220: **Integrative modeling based on fuzzy set of multiomics data**
Presenter: **Hyeyoung Jung**, Seoul National University, Korea, South

An integrative model of multiomics data is proposed to classify cancer. The proposed model uses fuzzy logic to integrate multiple types of molecular data with biological rules. An example using breast cancer subtypes illustrates the applicability of the proposed integrative modeling.

E1355: **Median-based multifactor dimensionality reduction methods for the survival phenotype**
Presenter: **Seungyeoun Lee**, Sejong University, Korea, South

With the development of high throughput technologies for genetic variants, genome-wide association studies on complex diseases such as hypotension, diabetes and cancers have been extensively developed for the last decades. The multifactor dimensionality reduction method has been originally proposed to reduce high order dimensions in gene-gene interaction analysis, in which the high-level genetic variants are classified into high and low risk groups by the ratio of the cases and controls for a case-control study. Many modifications for the multifactor dimensionality reduction methods have been proposed by allowing the various classifiers and generalizing the phenotypes. The quantitative multifactor dimensionality reduction method uses a t-test statistic to classify the high and low risk groups for the continuous phenotype. We propose to use a median survival time as a classifier for the survival phenotype, which called the median-based MDR. We compare the power of the proposed median-based MDR with the Surv-MDR in the simulation study and analyze a real example of the ovarian cancer patient data.

EO522 Room Sala Convegni ASTROSTATISTICS Chair: Mauro Bernardi

E0923: **Astronomical source detection and background separation via hierarchical Bayesian nonparametric mixtures**
Presenter: **Andrea Sottosanti**, University of Padua, Italy
Co-authors: Mauro Bernardi, Roberto Trotta, David van Dyk

The search of gamma-ray sources in the extra-galactic space is one of the main targets of the Fermi telescope project, which aims to identify and study the nature of high energy phenomena in the universe. This requires to separate their signal from a gamma-ray background component diffuse over the entire area observed by the telescope. From a statistical perspective, we can account for both these phenomena using a mixture of two densities: the first models the spread of photons around the sources, while the second includes the information from the background contamination. We propose a novel approach to the signal extraction of gamma-ray sources using a Dirichlet process mixture, that allows to discover and locate a possible infinite number of clusters in the map, and a new flexible Bayesian nonparametric model based on b-spline basis functions to account for the irregular shape of the background. The resultant is then a mixture of two Dirichlet process mixture models. From the results obtained on a region of the Fermi map we can conclude that the proposed approach both guarantees a posterior estimation of the number of sources and a complete separation of their signals from the background noise.

E0777: **Hunting for exoplanets around active stars**
Presenter: **David Stenning**, Imperial College London, United Kingdom
Co-authors: David Jones, Eric Ford, Tom Loredo, Jessi Cisewski, Robert Wolpert

The detection and characterization of exoplanets-planets that orbit stars other than the Sun-is one of the most active areas of research in modern astronomy. Many exoplanets are discovered using the radial velocity technique, which involves detecting the Doppler shift in a stars spectral lines resulting from the gravitational effects of an orbiting planet. A challenge to this approach is that measured radial velocity signals are often corrupted by stellar activity such as spots rotating across the stars surface. A principled method for recovering the underlying planetary radial velocity signal was previously proposed, which uses dependent Gaussian processes to jointly model the corrupted radial velocity signal and proxies for stellar activity. Our approach extends the previous one by (i) incorporating science- and data-driven dimension reduction techniques to extract more informative stellar activity proxies, and (ii) introducing a model comparison procedure to select the best model for the stellar activity proxies at hand from a larger class of models. Our methodology is tested on synthetic data generated by the SOAP 2.0 code, and initial results show substantially improved statistical power for planet detection compared to using existing models from the literature.

E0336: **Measuring precise radial velocities and cross-correlation function line-profile variations using a skew normal density**
Presenter: **Umberto Simola**, University of Helsinki, Finland
Co-authors: Jessi Cisewski, Xavier Dumusque

Stellar activity is one of the primary limitations to the detection of low-mass exoplanets using the radial velocity technique. Stellar activity can be probed by measuring time dependent variations in the shape of the cross-correlation function (CCF), often estimated using different parameters of the modelled CCF. A novel approach is to estimate the parameters of the CCF by fitting a skew normal density which, unlike the commonly employed normal density, includes a skewness parameter to capture the asymmetry of the CCF induced by stellar activity and also the natural asymmetry induced by convective blueshift. The performance of the proposed method is compared to the commonly employed normal density, using both simulations and real observations, with different levels of activity and signal-to-noise ratio levels. To use of the skew normal density results helpful to retrieve the different parameters of the CCF, since the correlations used to probe stellar activity are stronger than the ones retrieved by the common approach and because the uncertainties associated to the RV and the asymmetry of the CCF are both smaller than the ones retrieved by the commonly employed strategies.

E1044: **Frequentist or Bayesian: An exoplanet case study**
Presenter: **Alessandra Rosalba Brazzale**, University of Padova, Italy
Co-authors: Umberto Simola

The radial velocity method has for long been the by far most productive technique used to detect extrasolar planets orbiting around a star. We consider both, the frequentist (or neo-Fisherian) and the Bayesian frameworks, to estimate the orbital parameters of an extrasolar planet in a binary system. From the frequentist viewpoint, a likelihood-based fitting procedure is proposed. Bayesian computation uses adaptive Metropolis-Hastings sampling. Both approaches are compared. In particular we discuss how they address the number of methodological and computational issues inherent the radial velocity formula and, in particular, the strong correlation among some of the parameter estimates/posterior distributions. We focus also on the astronomical interpretation of the parameters and how the assumption of circularity of the orbit may be assessed.

EO304 Room A1 ADVANCES IN STATISTICAL NEUROIMAGING ANALYSIS Chair: Lexin Li

E0549: **Persistent homology on functional brain network**
Presenter: **Moo K Chung**, University of Wisconsin-Madison, United States

Advances in functional magnetic resonance imaging (fMRI) techniques enabled us to measure spontaneous fluctuations of neural signals in the brain. Many previous studies on resting-state fMRI have mainly focused on the topological characterization of static graph theory features that will not fluctuate over time. We present a simple but very effective data-driven approach to assess the dynamic pattern of resting state functional connectivity using persistent homology. Persistent homology has been successfully applied to various static brain networks by building graph filtrations by sequentially thresholding edge weights. The recently proposed exact combinatorial inference procedure for static network was adapted for statistically quantifying dynamic brain networks.

E0872: **A causal dynamic network model for functional MRI**
Presenter: **Xi Luo**, Brown University, United States
Co-authors: Xuefei Cao, Bjorn Sandstede

Functional MRI (fMRI) is a popular approach to investigate brain connections and activations when human subjects perform tasks. Because fMRI measures the indirect and convoluted signals of brain activities at a lower temporal resolution, differential equation modeling methods, for example dynamic causal modeling, are usually employed to model the neural processes and the resulting fMRI signals. However, this modeling strategy is computationally expensive and remains to be mostly a confirmatory or hypothesis-driven approach. A major statistical challenge is to infer, in a data-driven fashion, the underlying differential equation models from fMRI data. We propose a causal dynamic network (CDN) model to estimate brain activations and connections simultaneously. Our model links the observed fMRI data with the latent neural signals modeled by an ordinary differential equation (ODE) model. Using basis function expansions, we develop an optimization-based criterion that combines data-fitting errors and ODE fitting errors, and we develop a block coordinate-descent algorithm to compute the ODE parameters efficiently. We illustrate the numerical advantages of our approach using data from realistic simulations and a task-related fMRI experiment.

E0960: **Finding relevant communities in the brain with SPINNER**
Presenter: **Damian Brzyski**, Indiana University Bloomington, United States
Classical regression methods treat covariates as a vector and estimate a corresponding vector of regression coefficients. In medical applications, however, regressors in a form of multidimensional arrays can be often met. For example, one may be interested in identifying regions of the brain associated with an outcome of interest based on MRI images. Turning such image array into a vector is an unsatisfactory solution, since it destroys the inherent spatial structure of the image and could be very challenging from the computational point of view. We will present an alternative approach - the regularized matrix regression - where the matrix of regression coefficients is defined as a solution to the specific optimization problem. The method, called SParsity Inducing Nuclear Norm EstimatoR (SPINNER), simultaneously imposes two types of penalties on the matrix - the nuclear and the lasso-type norm - to encourage the low rank of the solution and its entry-wise sparsity. The alternating direction method of multipliers (ADMM) was used to build the fast and efficient numerical solver. SPINNER has been applied to investigate associations between brain's structural connections and HIV disease-related outcomes. Our approach outperforms others methods in the estimation accuracy in the considered situation - when the response-related entries (representing brain's connections) are arranged in well-connected communities.

E1296: **A new approach to Bayesian image analysis**
Presenter: **John Kornak**, University of California, San Francisco, United States
Co-authors: Karl Young
Bayesian image analysis can improve image quality, by balancing a priori expectations of image characteristics, with a model for the noise process via Bayes theorem. We will give a reformulation of the conventional Bayesian image analysis paradigm in Fourier space, i.e. the prior and likelihood are given in terms of spatial frequency signals. By specifying the Bayesian model in Fourier space, spatially correlated priors, that are relatively difficult to model and compute in conventional image space, can be efficiently modeled as a set of independent processes across Fourier space; the priors in Fourier space are modeled as independent, but tied together by defining a parameter function over Fourier space for the values of the pdf parameters. The originally inter-correlated and high-dimensional problem in image space is thereby broken down into a series of (trivially parallelizable) independent one-dimensional problems. We will describe the Bayesian image analysis in Fourier space (BIFS) modeling approach, illustrate its computational efficiency and speed, and demonstrate useful properties of isotropy and resolution invariance to model specification which are difficult to obtain in the conventional formulation. We will describe specific applications in medical imaging, and contrast with results based on more conventional Bayesian image analysis models. Finally, we will showcase a Python package that is under development to make the approach widely accessible.

EO534 Room Aula A RECENT ADVANCES IN ANALYSIS OF HIGH-DIMENSIONAL DATA Chair: Yunzhang Zhu

E0631: **Normalization and differential expression in single cell RNA-seq**
Presenter: **Hao Wu**, Emory University, United States
Co-authors: Zhijin Wu
Single cell RNA-seq (scRNA-seq) enables the transcriptomic profiling at individual cell level. This new level of resolution reveals inter-cellular transcriptomic heterogeneity and brings new promises to the understanding of transcriptional regulation mechanism. The special characteristics in scRNA-seq data, including excessive zeros, high variability, and multi-modal distribution, bring challenges in data analysis because typical assumptions made for bulk RNA samples are no longer hold. We will present a probabilistic model of sequencing counts that well explains the characteristics of single cell RNA-seq data. We will further present an adaptive normalization method that is robust to the bursting nature of expression in many genes, and a redefined differential expression procedure.

E0899: **Minimizing sum of truncated convex functions and its applications**
Presenter: **Hui Jiang**, University of Michigan, United States
A class of problems is studied where the sum of truncated convex functions is minimized. In statistical applications, they are commonly encountered when L_0-penalized models are fitted and usually lead to NP-hard non-convex optimization problems. We propose a general algorithm for the global minimizer in low-dimensional settings. We also extend the algorithm to high-dimensional settings, where an approximate solution can be found efficiently. We introduce several applications where the sum of truncated convex functions is used, compare our proposed algorithm with other existing algorithms in simulation studies, and show its utility in edge-preserving image restoration on real data.

E1053: **Network inference**
Presenter: **Peng Wang**, University of Cincinnati, United States
Network regression links a network's structures to covariates of interest, modeling pairwise conditional dependencies of interacting units as a function of covariates. For instance, in gene network analysis of a certain lung cancer, the network structures may vary over clinical attributes differentiating four different subtypes of the cancer. Within the framework of Gaussian structure equation models, we infer a network's structures, defined by an undirected graph, in in relation to covariates, through testing regression coefficients. To increase the power of hypothesis testing, we de-correlate the structure equation models, develop a combined constrained likelihood ratio test, combining independent marginal likelihoods and unregularizing hypothesized parameters whereas regularizing nuisance parameters through L_0-constraints controlling the individual degree of sparseness. On this ground, we derive asymptotic distributions of the combined constrained likelihood ratio, which is chi-square or normal depending on if the co-dimension of a test is finite or increases with the sample size. This leads to likelihood based tests in a high-dimensional situation permitting a network's size to increase in the sample size. Numerically, we demonstrate that the proposed method performs well in various situations. Finally, we apply the proposed method to infer a structural change of a gene network of a lung cancer with respect to four subtypes and other covariates of interest.

E1098: **An adaptive test on high-dimensional parameters in generalized linear models**
Presenter: **Wei Pan**, University of Minnesota, United States
Several tests for high-dimensional generalized linear models have been proposed recently, however, they are mainly based on a sum of squares of the score vector and only powerful under certain limited alternative hypotheses. In practice, since the associations in a true alternative hypothesis may be sparse or dense or between, the existing tests may or may not be powerful. We propose an adaptive test that maintains high power across a wide range of scenarios. To calculate its p-value, its asymptotic null distribution is derived. We conduct simulations to demonstrate the superior performance of the proposed test. Then we apply it and other existing tests to an Alzheimer's Disease Neuroimaging Initiative data set, detecting possible associations between Alzheimer's disease and sets of a large number of single nucleotide polymorphisms. As an end product, we put R package GLMaSPU implementing the proposed test on GitHub and CRAN.

EO669 Room Aula C MICROBIOME RESEARCH METHODS Chair: Alexander Alekseyenko

E0275: **A framework for multivariate causal mediation analysis with microbiome data**
Presenter: **Alexander Alekseyenko**, Medical University of South Carolina, United States
Translational microbiome research exists under the promise of uncovering microbial level interventions, which will promote health and eliminate disease. However, most of the available observational and experimental study designs do not allow for unambiguous determination of the direct causal role of the microbiome, as alternative interpretations are always plausible. For example, an antimicrobial intervention to manipulate the microbes may also have a direct effect on the measured outcome. For this reason, causal mediation analysis is desirable to estimate the extent to which microbes and intervention are responsible for observed study phenotypes. The univariate single mediator model framework achieves such inference by estimation of linear regressions. Even in the simplest case of univariate intervention and response, the microbiome measurements are highly multivariate, under-sampled, compositional and over-dispersed, which imposes modeling challenges in translating the single mediator model to microbiome data. Using distance-based energy statistics, we will derive a multivariate causal mediation framework suitable for application to microbiome data.

E0801: **Novel regression models for microbiome data**
Presenter: **Christian Lorenz Mueller**, Simons Foundation, United States
Targeted amplicon sequencing data, including 16S rRNA and ITS sequence data, are inherently compositional in nature. Using these data for regression tasks thus requires non-standard regression models that take compositionality into account. In addition, typical microbiome data are overdispersed and zero-inflated. To alleviate the challenges associated with these data, we present novel regression models for microbiome data that jointly model the underlying regression and scale vectors under a wide range modeling assumptions. The corresponding model estimation task can be formulated as non-smooth convex optimization problem which can be solved efficiently using a novel proximal algorithm formulation. We show improved prediction performance compared to state-of-the-art methods for regression tasks arising in microbiome data analysis from host-associated and environmental amplicon data.

E0983: **Multivariate logistic mixture regression for microbiome analysis**
Presenter: **Jack OBrien**, Bowdoin College, United States
Two standard approaches to understanding microbiome data are mixture models and regression techniques. Mixture models attempt to capture the structure and number of components of the data, with the aim of associating these with environmental or experimental conditions. Regression techniques seek to directly assess the connection between covariates and the specific features of the data (compositional, excess zeros, temporal or spatial structure). We show how recent advances in inference for multivariate logistic regression can be extended to a mixture context, determining the component structure as part of the regression. This translates the regression to a classification procedure and shows how the Voronoi tessellation can be used to understand microbial dynamics. A compositional approach is covered and details are given about how it may be extended to a non-compositional framework that may have wider applicability in marine ecology.

E1176: **Analyzing matched sets of microbiome data**
Presenter: **Yijuan Hu**, Emory University, United States
Matched data arise frequently in microbiome studies. For example, we may have gut microbiome data pre- and post-treatment from a set of individuals, or longitudinal microbiome samples (e.g., vaginal microbiome samples collected in each trimester of a pregnancy). We present a version of the Linear Decomposition Model (LDM) for analyzing matched datasets. The microbiome characterizing the set is treated as a 'nuisance parameter', allowing all effort to focus on the (common) differences within sets. We compare the power of the matched analysis with that of the standard (unmatched) analysis using the LDM.

EO683 Room B1 GRAPHICAL MARKOV MODELS V Chair: Kayvan Sadeghi

E1410: **Characterizing and learning equivalence classes of causal DAGs under interventions**
Presenter: **Karren Yang**, MIT, United States
Co-authors: Abigail Katcoff, Caroline Uhler
The problem of learning causal DAGs in the setting where both observational and interventional data is available is considered. This setting is common in biology, where gene regulatory networks can be intervened on using chemical reagents or gene deletions. It has been previously characterized the identifiability of causal DAGs under perfect interventions, which eliminate dependencies between targeted variables and their direct causes. We extend these identifiability results to general interventions, which may modify the dependencies between targeted variables and their causes without eliminating them. We define and characterize the interventional Markov equivalence class that can be identified from general (not necessarily perfect) intervention experiments. We also propose the first provably consistent algorithm for learning DAGs in this setting, and we evaluate our algorithm on simulated and biological datasets.

E1669: **On linear generating processes for joint distributions of binary variables having an undirected Markov graph structure**
Presenter: **Nanny Wermuth**, Chalmers University of Technology, Sweden
Co-authors: Giovanni Maria Marchetti
In general, linear relations among binary variables are not suitable to capture conditional dependences or independences among categorical variables. But, we can give necessary and sufficient conditions on undirected Markov graphs so that the associated joint distributions of standardised binary variables have recursive, linear generating processes. Surprisingly, graphs of chordless cycles belong to the class. The attractive properties of these models and of their parameters are summarised.

E1649: **A causal modeling framework in search of a graphical representation**
Presenter: **Joris Mooij**, University of Amsterdam, Netherlands
Co-authors: Tineke Blom
Structural Causal Models (SCMs) provide a causal modeling framework that is used in many fields such as economy, the social sciences, and biology. It offers appealing features, e.g., a graphical representation that simultaneously expresses conditional independence properties and causal properties of the model, which lies at the basis of many of the theoretical and algorithmic results in the area. We show that SCMs are not flexible enough to give a complete causal representation of equilibrium states of dynamical systems in general. We propose a generalization of the notion of SCM, that we call Causal Constraints Model (CCM), and prove that CCMs are flexible enough to capture the essential causal semantics of dynamical systems at equilibrium. As an illustration, we consider a simple and ubiquitous chemical reaction. The price one pays for the improved generality and flexibility of CCMs over SCMs is that no appropriate graphical representation of CCMs that simultaneously expresses conditional independence properties and causal properties is known. We challenge the graphical modeling community to invent such a representation.

E1733: **Testing for tetrad constraints in multivariate time series**
Presenter: **Michael Eichler**, Maastricht University, Netherlands
Tetrad constraints play a key role in the identification of latent variables structures. In the context of multivariate time series such tetrad constraints can be formulated in terms of the spectral density matrix. We present a test whether a tetrad constraint is satisfied and investigate its performance by a simulation study.

EO496 Room C1 ON RECENT DEVELOPMENT ABOUT TIME SERIES AND SPECTRAL ANALYSIS Chair: Yuexiao Dong

E0340: **Empirical band analysis of nonstationary time series**
Presenter: **Scott Bruce**, George Mason University, United States
Co-authors: Cheng Yong Tang, Robert Krafty
Power spectra of time series processes are defined over a continuous range of frequencies. However, time series data contain only a finite number of observations, so we must consider collapsed measures of power within local frequency bands that partition the frequency space. Frequency bands are used widely in the scientific literature and are often selected by manual observation of waveforms generated from a specific type of signal under particular settings. A standardized, unifying approach is provided to constructing customized frequency bands for different signals under study across different settings. A frequency-domain, iterative cumulative sum procedure is formulated to identify optimal frequency bands that best preserve nonstationary information. A formal hypothesis testing procedure is also developed to test which, if any, frequency bands remain stationary. This method is shown to consistently estimate the number of frequency bands and the location of the upper and lower bounds defining each frequency band.

E0368: **Bayesian spectral analysis of high-dimensional time series**
Presenter: **Zeda Li**, Baruch College, United States
Co-authors: Robert Krafty
A frequency-domain factor model is proposed which allows for complex-valued spectra which means that individual high-dimensional time series can propagate in a lagged fashion. Our model allows for different dynamics across the variates of the time series. The spectrum of the factors is assumed smooth as a function of frequency. The real and imaginary parts of the loading matrix are modeled by tensor products. Inference is performed by MCMC methods, and the method is illustrated with biomedical data.

E0612: **Conditional adaptive Bayesian spectral analysis of nonstationary biomedical time series**
Presenter: **Robert Krafty**, University of Pittsburgh, United States
Co-authors: Scott Bruce, Martica Hall
Many studies of biomedical time series signals aim to measure the association between frequency-domain properties of time series and clinical and behavioral covariates. However, the time-varying dynamics of these associations are largely ignored due to a lack of methods that can assess the changing nature of the relationship through time. We discuss a method for the simultaneous and automatic analysis of the association between the time-varying power spectrum and covariates, which we refer to as conditional adaptive Bayesian spectrum analysis (CABS). The procedure adaptively partitions the grid of time and covariate values into an unknown number of approximately stationary blocks and nonparametrically estimates local spectra within blocks through penalized splines. CABS is formulated in a fully Bayesian framework, in which the number and locations of partition points are random, and fit using reversible jump Markov chain Monte Carlo techniques. Estimation and inference averaged over the distribution of partitions allows for the accurate analysis of spectra with both smooth and abrupt changes. The proposed methodology is used to analyze the association between the time-varying spectrum of heart rate variability and self-reported sleep quality in a study of older adults serving as the primary caregiver for their ill spouse.

E1392: **Network analysis for time series data**
Presenter: **Brandon Park**, George Mason University, United States
Co-authors: Anand Vidyashankar, Tucker McElroy
Network based approaches for analyses of time series data can provide new insights concerning causality and forecasting. However, it is challenging to construct such networks using time series data, due to correlations and autocorrelations between the nodes. Additionally, the problem is getting more complicated when exogenous variables are present. We (i) describe theoretical guarantees for the regularization method for estimating the autoregressive parameters and the regression coefficients in an ARX Model, (ii) describe a new method for constructing an implicit network, and (iii) provide various network wide metrics (NWM) that are useful for identifying the active features of the implicit network. We also describe the asymptotic properties of NWM and utilize them to identify communities via the proposed implicit network.

EO663 Room E1 RECENT ADVANCES IN NETWORK DATA ANALYSIS Chair: Feng Liang

E0879: **Finite-graph superpopulation inference for random graphs with complex dependence**
Presenter: **Michael Schweinberger**, Department of Statistics, Rice University, United States
In practice, network scientists are often interested in superpopulation inference for random graphs with complex topological structures. In other words, there is a finite population of nodes and a population graph is generated by a population probability model capturing complex topological structures, including various forms closure in networks. We consider a finite population of nodes partitioned into subpopulations, e.g., armed forces partitioned into units of armed forces or school populations partitioned into school classes. Such data are increasingly widely collected in network science, and are called multilevel network data. We present non-asymptotic concentration and consistency results, assuming the population probability model restricts dependence to subpopulations, the number of subpopulations is large relative to the size of the largest subpopulation, and the subpopulation graphs are governed by curved exponential families with geometrically weighted terms capturing sensible forms of transitive closure.

E0930: **Heterogeneous susceptibilities in network influence models**
Presenter: **Daniel Sewell**, University of Iowa, United States
Network autocorrelation models are widely used to evaluate the impact of influence on some variable of interest or diffusion through a network. This is a large class of models that parsimoniously accounts for how one's neighbors influence one's own behaviors, opinions, or states by incorporating the network adjacency matrix into the joint distribution of the data. These models assume homogeneous susceptibility to influence through the network, however, which may be a strong assumption in many contexts. A hierarchical model is proposed which allows the influence parameter to be a function of individual attributes and/or of local network topological features. An approximation of the posterior distribution is derived in a general framework that is applicable to the Durbin, network effects, network disturbances, or network moving average autocorrelation models. The proposed approach can also be applied to investigating determinants of influence in the context of egocentric network data.

E0729: **Applying a network modeling approach to the analysis of binary item response data**
Presenter: **Ick Hoon Jin**, University of Notre Dame, United States

Item response theory (IRT) is one of the most widely utilized tools for item response analysis; however, local item and person independence, which is a critical assumption for IRT, is often violated in real testing situations. We propose a new type of analytical approach for item response data that does not require standard local independence assumptions. By adapting a latent space joint modeling approach, the proposed model can estimate pairwise distances to represent the item and person dependence structures, from which item and person clusters in latent spaces can be identified. We provide an empirical data analysis to illustrate an application of the proposed method. A simulation study is provided to evaluate the performance of the proposed method in comparison with existing methods.

E0919: **Spectral clustering with higher-order structures under a superimposed stochastic block model**
Presenter: **Subhadeep Paul**, The Ohio State University, United States

Higher-order subgraphs or motif structures are increasingly important in network studies to understand functionality, regulation, and control of complex networks. We consider the problem of community detection in a complex network based on such higher-order structures. We propose a Superimposed Stochastic Block Model (SupSBM), a random graph model with community structure that allows dependence among the network edges and leads to the presence of higher-order structures yet remaining mathematically tractable. The model is based on a signal and noise superimposition framework where certain higher-order structures are directly generated from a signal random graph model, and superimposed with edges generated randomly from a noise random graph model. We then analyze the performance of the recently proposed higher-order spectral clustering method. We prove non-asymptotic upper bounds for the error of community detection using the method under a SupSBM where the signal component consists of triangle motifs and the noise component consists of undirected edges. As part of our analysis, we also derive error bounds of the higher-order spectral clustering method under the ordinary SBM and the triangle hypergraph SBM. Finally, under the non-uniform hypergraph SBM where one observes edges and triangle hyperedges distinctly, we obtain a criterion to choose between spectral clustering using edges or triangle hyperedges.

EO348 Room F1 DYNAMIC MODELS AND STRUCTURAL CHANGES Chair: Zuzana Praskova

E0601: **Monitoring non-stationary processes**
Presenter: **Wolfgang Schmid**, European University Viadrina, Germany
Co-authors: Taras Lazariv

In the literature related to the statistical process control for time-dependent data it is assumed that the underlying process is stationary. However, in finance and economics we are often faced with situations where the process is close to non-stationarity or it is even non-stationary. A target process is modeled by a multivariate state-space model which may be non-stationary. The aim is to monitor its mean behavior. The likelihood ratio method, the sequential probability ratio test, and the Shiryaev-Roberts procedure are applied to derive control charts signaling a change from the supposed mean structure. These procedures depend on certain reference values which have to be chosen by the practitioner in advance. The corresponding generalized approaches are considered as well, and generalized control charts are determined for state-space processes. These schemes do not have further design parameters. In an extensive simulation study the behavior of the introduced schemes is compared with each other using various performance criteria as the average run length, the average delay, the probability of a successful detection, and the probability of a false detection.

E1115: **Change detection and estimation for the covariance matrices of a high-dimensional time series**
Presenter: **Ansgar Steland**, University Aachen, Germany

New results about inference and change point analysis of zero mean high dimensional vector time series are discussed. Applications cover sparse principal component analyses, financial portfolio management and signal processing. The results deal with change-point procedures that can be based on an increasing number of bilinear forms of the sample variance-covariance matrix as arising, for instance, when studying change-in-variance problems for projection statistics and shrinkage covariance matrix estimation. Contrary to many known results, e.g. from random matrix theory, the results hold true without any constraint on the dimension, the sample size or their ratio. The large sample approximations are in terms of (strong resp. weak) approximations by Gaussian processes. They hold not only without any constraint on the dimension, the sample size or their ratios, but even without any such constraint with respect to the number of bilinear form. Further, distributional approximations for shrinkage covariance matrix estimators are provided including a confidence interval for the shrinkage intensity and lower and upper bounds for the covariance matrix. These bounds lead to novel bounds for portfolio risks in financial portfolio analysis. The accuracy of the methods is investigated by simulations. Lastly, we discuss an application to asset returns from financial markets.

E1224: **Change point detection in multivariate two-sample setup**
Presenter: **Zdenek Hlavka**, Charles University, Czech Republic
Co-authors: Marie Huskova, Simos Meintanis

New methods are discussed for detecting structural breaks in a series of multivariate observations but, instead of considering general alternatives, we concentrate on a two-sample setup. In other words, we assume that two random subvectors have identical distribution until the unknown change-point. Most often, these random subvectors will consist of the same variables observed for two different populations and the goal will be to estimate the unknown change-point. The procedures are based on L_2-type criteria utilizing multivariate empirical characteristic functions leading to computationally attractive closed-form expressions. We present asymptotic and Monte-Carlo results for both on-line and off-line type procedures.

E0936: **Testing structural breaks in large dynamic models**
Presenter: **Zuzana Praskova**, Charles University, Czech Republic

A linear dynamic panel data model with cross-sectional dependence is considered. A procedure to detect changes in coefficients of lagged variables is proposed and asymptotic distribution of the test statistic is studied as the number of panels and the number of observations converge to infinity. The asymptotic distribution is a functional of a Gaussian process and the critical values of the test can be obtained by simulations only. A bootstrap variant of the test statistic is considered that takes into account both the temporal and the cross-sectional dependences.

EO442 Room G1 RECENT ADVANCES OF STATISTICAL METHODS IN SURVIVAL ANALYSIS AND MISSING DATA Chair: Sy Han Chiou

E1100: **A flexible joint longitudinal-survival modeling framework for incorporating multiple longitudinal biomarkers**
Presenter: **Daniel Gillen**, University of California, Irvine, United States
Co-authors: Sepehr Akhavan, Alexander Vandenberg-Rodes, Babak Shahbaba

When monitoring the health of subjects it is common for multiple biomarkers to be measured longitudinally over time. While the associations between the collected biomarkers and a time-to-event endpoint are often of primary scientific interest, modeling the longitudinal risk factors simultaneously can be beneficial, particularly when the density of measurements is differential across biomarkers or when data on some biomarkers are intermittently missing. We propose a joint longitudinal-survival framework with the longitudinal component modeled via a Gaussian process that allows for the correlation between biomarker trajectories to be estimated and utilized. Biomarker trajectories are then linked to survival times via a multiplicative hazards model. Joint estimation of the longitudinal and survival models is performed to account for uncertainty in the estimated time-dependent biomarkers. The proposed methodology is robust to strong parametric assumptions on the mean and covariance structure of the longitudinal component, while also allowing for subject-specific baseline hazard functions in the survival component. Simulation studies are presented to illustrate the performance of the proposed method. We further use the approach to estimate the association between multiple serum-based nutritional biomarkers and survival among end-stage renal disease patients.

E1271: **Drawing inference for high-dimensional models via a selection-assisted partial regression approach**
Presenter: **Yi Li**, University of Michigan, United States

Drawing inferences for high-dimensional models is challenging as regular asymptotic theories are not applicable. A new framework of simultaneous estimation and inferences for high-dimensional linear models is proposed. By smoothing over partial regression estimates based on a given variable selection scheme, we reduce the problem to a low-dimensional least squares estimation. The procedure, termed as Selection-assisted Partial Regression and Smoothing (SPARES), utilizes data splitting along with variable selection and partial regression. We show that the SPARES estimator is asymptotically unbiased and normal, and derive its variance via a nonparametric delta method. The utility of the procedure is evaluated under various simulation scenarios and via comparisons with the de-biased lasso estimators, a major competitor. We apply the method to analyze two genomic datasets and obtain biologically meaningful results.

E1365: **An alternative sensitivity analysis approach for missing not at random data**
Presenter: **Chengcheng Hu**, University of Arizona, United States
Co-authors: Chiu-Hsieh Hsu, Yulei He

Missing mechanism is unverifiable. Often researchers perform sensitivity analysis to evaluate the impact of various missing mechanisms. All the existing sensitivity analysis approaches for missing not at random (MNAR) data require fully specifying the relationship between the missing value and the missing probability or simply use the delta adjustment approach. The relationship is specified using a selection model, a pattern-mixture model, or a shared parameter model. We propose an alternative sensitivity analysis approach for MNAR using a nonparametric multiple imputation approach, which only requires specifying the correlation between the missing value and the missing probability. The correlation is a standardized measure and can be directly used to indicate the magnitude of MNAR. We perform simulation studies to compare the proposed sensitivity analysis approach and the delta adjustment procedure. Numerical results indicate that the proposed approach performs well and can be used as an alternative approach for MNAR. The proposed sensitivity analysis approach is demonstrated on hemoglobin A1c data of open heart surgery patients, which are subject to missing especially for non-diabetic patients.

E1373: **Applications of survival analysis towards building a value-driven pre-emptive genotyping program**
Presenter: **Jonathan Schildcrout**, Vanderbilt University, United States

Currently, there are more than 250 medications containing genetics-based guidance within the USFDA label inserts, and in recent years, Vanderbilt University Medical Center developed a quality improvement program to incorporate genetic information into the electronic health record that, when appropriate, permits genotype guided prescribing. Since a large percentage of patients are prescribed medications with pharmacogenetic (PGx) effects and many patients are prescribed multiple such medications, the program involves both pre-emptive and multiplexed genetic testing. Pre-emptive genotyping allows physicians to use the genetic information seamlessly at the time they decide to prescribe a PGx medication, and multiplexing permits cost efficiencies when genetic data are reused. It is cost-prohibitive to genotype all patients, and we describe our approach to identifying patients for genotyping based on anticipated benefit to the patient. We will detail 1) a survival analysis-based, predictive modeling approach using clinical data to estimate patient-level risk of being prescribed each PGx medication, 2) a discrete event simulation that uses literature-based estimates of adverse event rates, variant allele frequencies, and secular death rates to capture the impact of genotype guided therapy in patients once prescribed, and 3) a decision theoretic approach to combine 1) and 2) in order to develop genotyping rules.

EO314 Room H1 RECENT ADVANCES IN FLEXIBLE DIRECTIONAL MODELING Chair: Jose Ameijeiras-Alonso

E0593: **WeiSSVM model and its applications to cylindrical data**
Presenter: **Toshihiro Abe**, Nanzan University, Japan

By focusing on cylindrical distributions, examples of cylindrical data are shown. After a brief review of probability distributions on the line and on the circle, we introduce WeiSSVM distribution. Using a statistical model of forest tree data in Finland, we demonstrate an application of the cylindrical distributions to quantify the factors that affect asymmetric crown expansion.

E0773: **Nonparametric methods for circular data with errors in variables**
Presenter: **Agnese Panzera**, University of Florence, Italy
Co-authors: Marco Di Marzio, Stefania Fensore, Charles C Taylor

Nonparametric approaches are considered to estimate circular densities and regression curves involving circular variables when data are affected by measurement errors. Generalizations to the case when the density of the error is unknown are performed via a double smoothing version of the previous methods. Asymptotic properties along with some simulation results are provided.

E0837: **A directional proposal to solve a chronobiological problem**
Presenter: **Yolanda Larriba**, University of Valladolid, Spain
Co-authors: Cristina Rueda, Miguel Fernandez, Shyamal Peddada

Biological processes, such as cell cycle, circadian clock or blood pressure, are governed by oscillatory systems whose components exhibit periodic patterns over time. The study of these periodic patterns, or rhythms, and how they change under different conditions, is called chronobiology. An interesting problem in chronobiology is to reconstruct the temporal order at which samples were taken, when it is unknown. For instance, when dealing with human organ biopsies or autopsies, the exact time of the day (e.g. 11:00 A.M.) at which each sample was taken is often unrecorded, as it might suppose a risk for health or a cost increase. A directional proposal is proposed to analyse chronobiological rhythms. The intrinsic circular nature of these data suggests a mathematical formulation in the circular space and the use of circular statistics. The specific problem described above, the temporal order estimation, is solved as a circular order aggregation problem within the framework of a general methodology

based on order restrictions. In particular, we recover the temporal order among different cell cycle stages for unsynchronized single cell RNA-seq (scRNA-seq) transcriptome data. Simulation experiments are also illustrated to validate the procedure.

E0964: **Circulas obtained through a Fourier series based approach**
Presenter: **Shogo Kato**, Institute of Statistical Mathematics, Japan
Co-authors: Arthur Pewsey, Chris Jones

Circular data are a set of observations which can be expressed as angles $[-\pi,\pi)$. Bivariate circular data, comprised of pairs of circular observations $[-\pi,\pi)^2$, arise in numerous contexts. A general method is proposed to obtain copulas for bivariate circular data which are called circulas. This is achieved first by representing probability density functions of bivariate circular distributions in terms of Fourier series. With this representation, some conditions on the Fourier coefficients which produce a general family of circulas are presented. Then, as special cases of the general family, some classes of circulas arising from different patterns of non-zero Fourier coefficients are considered. The shape and sparsity of such arrangements are found to play a key role in determining the properties of the resultant models. All the special cases of the considered circulas have simple closed-form expressions for their densities and display different dependence structures between variables.

EO214 Room I1 NEW ADVANCES ON STATISTICAL MODELING OF COMPLEX DATA II Chair: Tsung-I Lin

E0489: **Improving the precision of oncology trials analysis using progression-free-survival as an endpoint**
Presenter: **Chien-Ju Lin**, MRC Biostatistics Unit University of Cambridge, United Kingdom
Co-authors: James Wason

In many oncology trials, patients are followed up until progression or death and the time at which this happens is used as the efficacy endpoint. This is known as progression-free-survival (PFS). Typical analyses consider tumour progression as a binary event, but in fact it is defined by a certain change in tumour size. This additional information on continuous tumour shrinkage at multiple times is discarded. We propose a method to make use of this information to improve the precision of analyses using PFS. We use joint modelling of the continuous tumour measurement, death and progression for other reasons (such as new tumour lesions) to construct survival curves. We present how to compute confidence intervals for quantities of interest, such as the median or mean PFS. We assess the properties of the proposed method by using simulated data and real data from a real phase II cancer trial. We also showcase a R-Shiny app to implement the proposed method.

E0896: **Multivariate measurement error models based on Student-t distribution under censored responses**
Presenter: **Mauricio Castro**, Pontificia Universidad Catolica de Chile, Chile
Co-authors: Larissa Avila Matos, Celso Cabral, Victor Hugo Lachos Davila

Measurement error models constitute a wide class of models, that include linear and nonlinear regression models. They are very useful to model many real life phenomena, particularly in the medical and biological areas. The great advantage of these models is that, in some sense, they can be represented as mixed effects models, allowing to us the implementation of well-known techniques, like the EM-algorithm for the parameter estimation. We consider a class of multivariate measurement error models where the observed response and/or covariate are not fully observed, i.e., the observations are subject to certain threshold values below or above which the measurements are not quantifiable. Consequently, these observations are considered censored. We assume a Student-t distribution for the unobserved true values of the mismeasured covariate and the error term of the model, providing a robust alternative for parameter estimation. Our approach relies on a likelihood-based inference using an EM-type algorithm. The proposed method is illustrated through some simulation studies and the analysis of an AIDS clinical trial dataset.

E0922: **Autoregressive skew-normal/independent linear mixed models**
Presenter: **Victor Hugo Lachos Davila**, University of Connecticut, United States

In longitudinal studies, the repeated measures of each subject are collected over time and hence tend to be serially correlated. An extension of skew-normal/independent linear mixed models previously introduced is considered, where the error term has Ar(p) dependence. The proposed model provides flexibility in capturing the effects of skewness and heavy tails simultaneously when continuous repeated measures are serially correlated. We present an efficient EM-type algorithm for the computation of maximum likelihood estimation of parameters. The observed information matrix is derived analytically to account for standard errors, in addition, the technique for the prediction of future responses under this model is also investigated. The methodology is illustrated through an application to schizophrenia data and some simulation studies.

E1193: **Testing the equality of two object parameters in two populations of symbolic data**
Presenter: **Anuradha Roy**, The University of Texas at San Antonio, United States
Co-authors: Daniel Klein

Advances in computing power in the past few decades greatly encouraged the collection of hundreds of thousands of data (big data) in our everyday lives. Big data in an object format such as histograms or intervals provide a more complete picture and the dynamics of the data. Big data can be explored by symbolic data analysis in which the object of the analysis is not a single-valued variable, but an object variable including histogram-valued and interval-valued variables. We will consider a method of testing the equality of two mean intervals for interval-valued data, and consider a Mushroom data set to illustrate our proposed method. The key point we want to emphasize is that we do not treat the Mushroom data as the classical single-valued data, nor as a very large collection of individual observations (also known as 'Big Data'), but rather some type of structured aggregated data such as interval-valued data.

EO432 Room L1 LATENT VARIABLE MODELS WITH APPLICATIONS Chair: Sara Taskinen

E0414: **The analysis of longitudinal mixed data via multivariate latent variable models: An analysis on alcohol use disorder**
Presenter: **Silvia Cagnone**, University of Bologna, Italy
Co-authors: Cinzia Viroli

Alcohol abuse is a dangerous habit in young people. The National Youth Survey is a longitudinal American study in part devoted to the investigation of alcohol disorder over time. The symptoms of alcohol disorder are measured by binary and ordinal items. In the literature it is well recognized that alcohol abuse can be measured by a latent construct; therefore generalized latent variable models for mixed data represent the ideal framework to analyse these data. However, it might be desirable to cluster individuals according to the different severity of their alcohol use disorder and to investigate how the groups vary over time. We present a new methodological framework that includes two levels of latent variables: one continuous hidden variable for dimension reduction and clustering and a discrete random variable accounting for the dynamics of alcohol disorder symptoms. The effect of covariates is also measured and a testing procedure for the temporal assumption is developed. Three important issues are discussed. First, it represents a unified framework for the analysis of longitudinal multivariate mixed data. Secondly, it captures and models the unobserved heterogeneity of the data. Finally, it describes the dynamics of the data through the definition of latent constructs.

E0978: **Approximate inference in latent variable models based on dimension-wise quadrature**
Presenter: **Silvia Bianconcini**, University of Bologna, Italy
Co-authors: Silvia Cagnone

Approximate methods are considered for likelihood inference to longitudinal and multidimensional data within the context of health science studies. The complexity of these data necessitates the use of sophisticated statistical models that can pose significant challenges for model fitting in terms of computational speed, memory storage, and accuracy of the estimates. Our methodology is motivated by a study that examines the temporal evolution of the mental status of the US elderly population between 2006 and 2010. We propose modeling the individual mental status as a latent process also accounting for the effects of individual specific characteristics, such as gender, age, and years of educational attainment. We describe the specification of such a model within the generalized linear latent variable framework, and its efficient estimation using a recent technique, called dimension-wise quadrature. The latter allows a fast and streamlined analytical approximate inference for complex models, with better or no degradation in accuracy compared with the standard techniques, such as Laplace approximation and adaptive quadrature. The model and the method are applied in the analysis of cognitive assessment data from the health and retirement study combined with the asset and health dynamic study.

E1293: **Output-sparse latent Gaussian processes**
Presenter: **Markus Heinonen**, Aalto University, Finland

Zero-inflated datasets, which have an over-abundance of zero outputs, are commonly encountered in problems such as climate or rare event modelling. Conventional machine learning approaches tend to overestimate the non-zeros leading to poor performance. We propose a novel family of zero-inflated Gaussian processes (ZiGP), produced by sparse kernels through learning latent probit processes that can zero out kernel rows and columns whenever the signal is absent. The ZiGPs are particularly useful for making the powerful Gaussian process networks more interpretable. We introduce sparse GP networks where variable-order latent modelling is achieved through sparse mixing of latent signals. To scale the models to large datasets, we demonstrate that the variational inference of probit-sparsified networks of latent Gaussian processes is remarkably tractable.

EO188 Room M1 RECENT ADVANCES ON FUNCTIONAL DATA ANALYSIS AND APPLICATIONS Chair: Ana Maria Aguilera

E0997: **Functional regression for estimating probability density functions: An application to electricity price forecasting**
Presenter: **Jose Portela**, Universidad Pontificia Comillas, Spain
Co-authors: Antonio Bello

Probabilistic forecasting of electricity prices in the medium term is highly important for operational scheduling, fuel purchasing, trading and profit management. In this context, fundamental models are frequently used, which obtain a probabilistic forecast based on market equilibrium simulations. While they provide insights when structural and regulatory changes are expected to happen in the market, these are not well calibrated to actual data. That is why hybrid methods are a growing research field, whose objective is to aggregate the fundamental forecasts with statistical methods to increase predictive capability. The proposed hybrid approach is to use a functional regression model that estimates the probability density function of the electricity price for each hour using, as explanatory variables, the probabilistic forecasts from the fundamental model. The functional parameters used in the regression are integral operators in the L^2 space and, in this approach, the kernels of the operators are modeled as a linear combination of sigmoid functions. The novelty of the method is that, as the endogenous variable is unobserved (only price realizations are known), the parameters are estimated by maximizing the likelihood of the price realizations over the estimated density functions.

E0723: **Functional data analysis of resistive switching processes**
Presenter: **Ana Maria Aguilera**, University of Granada, Spain
Co-authors: M Carmen Aguilera-Morillo, Juan Bautista Roldan , Francisco Jimenez-Molinos

The current flash memories have several problems that prevent a further reduction so that it is necessary a technological substitute. One of the strong candidates for future nonvolatile applications are Resistive Random Access Memories (RRAMs) due to their excellent properties of good scalability, long endurance, fast switching speed, and ease of integration in the CMOS processing. These devices have a simple physical structure: two metal plates acting as electrodes with a dielectric in between. The physical mechanisms behind resistive switching are stochastic and present important numerical problems for the correct modelling. The conduction takes place trough conductive filaments and a reasonable electric current is obtained under an applied voltage. The filaments are formed (set) and destroyed (reset) within the resistive switching device operation. The device resistance switches from a High Resistance State (HRS) to a Low Resistance State (LRS) so that the result is a sample of current-voltage curves corresponding to the reset/set cycles. Because of this, functional data analysis (FDA) methodologies are applied for modelling and explaining the variability associated with the stochastic process generating these curves. Data registration, P-spline approximation and Functional Principal Component Analysis (FPCA) are considered to provide a simple model that explains most variability in terms of one scalar variable highly correlated with the voltage to reset/set.

E1225: **Banach-valued multivariate mixed effects model with weakly dependent errors**
Presenter: **Javier Alvarez-Liebana**, University of Granada, Spain
Co-authors: Maria Dolores Ruiz-Medina

A mixed effect model in function spaces, under weakly dependent functional errors, is introduced. Specifically, the response, the fixed and random effect parameters are valued in a separable Banach space. A simulation study is undertaken to illustrate the performance of the methodology adopted.

E1341: **Tidyfun: A new framework for representing and working with function-valued data**
Presenter: **Fabian Scheipl**, Ludwig-Maximilians-Universitaet Muenchen, Germany
Co-authors: Jeff Goldsmith

A new R package "tidyfun" (https://fabian-s.github.io/tidyfun/) for working with function-valued data is presented which implements a unified interface for dealing with regularly or irregularly observed function-valued data and functional data in basis representation. The package follows the tidyverse design philosophy of R packages and provides idiomatic functions for quickly and easily wrangling and exploring functional data and, specifically, datasets that contain both scalar and functional data or multiple types of functional data, potentially measured over different domains. We discuss the available feature set as well as forthcoming extensions and show some application examples.

EO364 Room N1 HIGH DIMENSIONAL EXTREMES Chair: Chen Zhou

E0294: **Bayesian model averaging over tree-based dependence structures for multivariate extremes**
Presenter: **Raphael Huser**, King Abdullah University of Science and Technology, Saudi Arabia
Co-authors: Sabrina Vettori, Johan Segers, Marc Genton

Describing the complex dependence structure of extreme phenomena is particularly challenging. To tackle this issue we develop a Bayesian model that describes extremal dependence by taking advantage of the inherent hierarchical dependence structure of the max-stable nested logistic distribution. The proposed algorithm can identify possible clusters of extreme variables using reversible jump Markov chain Monte Carlo techniques. Parsimonious representations are achieved when clusters of extreme variables are found to be completely independent. Moreover, we significantly decrease the computational complexity of full likelihood inference by deriving a recursive formula for the nested logistic model likelihood. The algorithm performance is verified through extensive simulation experiments which also compare different likelihood procedures. The new methodology is used to investigate the dependence relationships between extreme concentration of multiple pollutants in California and how these pollutants are related to extreme weather conditions. Overall, we show that our approach allows for the representation of complex extremal dependence structures and has valid applications in multivariate data analysis, such as air pollution monitoring, where it can guide policymaking. Bayesian modeling of multivariate spatial extremes will also be discussed.

E0690: **Estimation of high-dimensional extreme conditional expectiles**
Presenter: **Gilles Stupfler**, The University of Nottingham, United Kingdom
Co-authors: Stephane Girard

The concept of expectiles is a least squares analogue of quantiles. It has, over the last decade, received a fair amount of attention due to its potential for application in financial, actuarial, and economic contexts. Some very recent work has focused on the application of extreme expectiles to assess tail risk, and on their estimation in a heavy-tailed framework. We investigate the estimation of extreme conditional expectiles of a heavy-tailed random variable Y given a finite-dimensional covariate X, whose dimension is allowed to be large. We derive generic conditions under which the limiting behaviour of our estimators can be investigated. We then present applications of our results to certain regression models of particular interest, as well as a finite-sample study to get a grasp of the behaviour of our procedures in practice.

E0761: **Modelling extreme European windstorms with functional peaks-over-threshold analysis**
Presenter: **Raphael de Fondeville**, EPFL, Switzerland
Co-authors: Anthony Davison

Estimating the risk of extreme windstorms has become important in recent decades, but up until now it has been largely limited to re-using catalogues of historical events, which usually do not exceed 40 to 50 years in length, and to numerical models, which require heavy computation but are often unreliable for extrapolation. Extreme value theory provides statistical methods for estimating the frequency of past extreme events as well as for extrapolating beyond observed severities, but natural hazards cannot be modelled using only univariate results. We present a statistical methodology based on functional peaks-over-threshold analysis which allows one to define complex extreme events as special types of exceedances, and then obtain their limit distribution for increasingly high thresholds, namely the generalized r-Pareto process. We describe a model based on log-Gaussian functions, which enables to use classical Gaussian covariance structures to model extremal dependence and then develop a stochastic weather generator for extreme windstorms over Europe, capable of quantifying the recurrence of past events as well as generating completely new ones.

E1404: **On binary classification in extreme regions**
Presenter: **Anne Sabourin**, Telecom ParisTech, France
Co-authors: Stephan Clemencon, Hamid Jalalzai

In pattern recognition, a random label Y is to be predicted based upon observing a random vector X valued in R^d with $d \geq 1$ by means of a classification rule with minimum probability of error. In a wide variety of applications, extreme observations X are of crucial importance, while contributing in a negligible manner to the (empirical) error however, simply because of their rarity. As a consequence, empirical risk minimizers generally perform very poorly in extreme regions. The aim is to develop a general framework for classification of extreme data. Precisely, under heavy-tail assumptions for the class distributions, we introduce a natural and asymptotic notion of risk, accounting for predictive performance in extreme regions of the input space. We show that minimizers of an empirical version of a non-asymptotic approximant of this dedicated risk, based on a fraction of the largest observations, lead to classification rules with good generalization capacity, by means of maximal deviation inequalities in low probability regions. Beyond theoretical results, numerical experiments are presented in order to illustrate the relevance of the approach developed.

EO500 Room O1 COPULAS AND DEPENDENCE MODELLING Chair: Piotr Jaworski

E0223: **On bivariate copula mappings**
Presenter: **Martynas Manstavicius**, Vilnius University, Lithuania
Co-authors: Gediminas Bagdonas

Inspired by many examples from the literature, we are concerned with a particular method of bivariate copula construction, namely, for a given copula $C:[0,1]^2 \to [0,1]$ and a function $f:[0,1] \to \mathbb{R}_+$ we let $H_f(C)(x,y) := C(x,y) f(\overline{C}(x,y))$, where $\overline{C}(x,y) := 1-x-y+C(x,y)$, $(x,y) \in [0,1]^2$, is the survival function associated with copula C. To classify the functions f based on the properties of $H_f(\cdot)$, we call a particular function f *eligible* if $H_f(C)$ is a copula for any bivariate copula C, *conditionally-eligible* if $H_f(C_1)$ is a copula but $H_f(C_2)$ is not a copula for some bivariate copulas C_1, C_2, and *non-eligible* if $H_f(C)$ is not a copula for any bivariate copula C. We then provide necessary and sufficient conditions for f to be eligible, and illustrate with examples. Some of the statistical properties of $H_f(C)$ for eligible f are also discussed.

E0875: **The Markov product of copulas revisited**
Presenter: **Wolfgang Trutschnig**, University of Salzburg, Austria

It is well known that the so-called star-product of two-dimensional copulas can equivalently be expressed in terms of the standard composition of the underlying Markov kernels. Building on this simple fact and on recent results we show that idempotence (i.e. invariance with respect to the star product) is a very rare property for copulas in commonly used classes. In particular, we prove that in the class of extreme-value copulas, in the class of Bernstein copulas, and in some special class of copulas represented by Lebesgue-measure-preserving transformations only the usual suspects - the product copula and minimum copula - are idempotent. Additionally, we prove a conjecture saying that the only idempotent strict Archimedean copula is the product copula.

E1524: **Variational inference for high dimensional structured factor copulas**
Presenter: **Hoang Nguyen**, Universidad Carlos III de Madrid, Spain
Co-authors: Pedro Galeano, Concepcion Ausin
Factor copula models have been recently proposed for describing the joint distribution of a large number of variables in terms of a few common latent factors. We employ a Bayesian procedure to make a fast inference for multi-factor and structured factor copulas. To deal with the high dimensional structure, we apply a Variational Inference (VI) approximation to estimate the different specifications of factor copula models. Compared to the Markov chain Monte Carlo (MCMC) approach, the VI approximation is much faster and could handle a sizeable problem in a few seconds. Another issue of factor copula models is that the bivariate copula functions connecting the variables are unknown in high dimensions. We derive an automated procedure to recover the hidden dependence structure. By taking advantage of the posterior modes of the latent variables, we select the bivariate copula functions based on minimizing Bayesian information criterion (BIC). The simulation studies in different contexts show that the procedure of bivariate copula selection could be at least 80% accuracy in comparison to the true generated copula model. We illustrate our proposed procedure with high dimensional real dataset.

E0851: **Discrete copulas and stochastic monotonicity**
Presenter: **Elisa Perrone**, Massachusetts Institute of Technology, United States
Discrete copulas serve as useful tools for modeling dependence among random variables. The space of discrete copulas admits a representation as a convex polytope which has for instance been exploited in entropy-copula methods relevant to environmental sciences. We further analyze geometric features of discrete copulas with prescribed stochastic properties. In particular, we focus on studying geometrically the class of componentwise convex discrete copulas, i.e., ultramodular discrete copulas, which capture joint behavior of mutually stochastically decreasing random variables. First, we identify the minimal collection of bounding affine hyperplanes of the convex space of ultramodular discrete copulas. Then, we analyze some of the extremal points of the class. Finally, we show how our geometric findings can be used to conduct hypothesis testing for stochastic decreasingness of bivariate random vectors.

EO062 Room Q1 DATA DEPTH AND HIGH-DIMENSIONAL DATA Chair: Sara Lopez Pintado

E0274: **On robust nonparametric techniques for dealing with the analysis of high dimensional data**
Presenter: **Francesca Ieva**, Politecnico di Milano, Italy
Co-authors: Anna Maria Paganoni, Juan Romo, Nicholas Tarabelloni
The aim is to gather recently proposed statistical methods that deal with the robust inferential analysis of univariate and multivariate functional data. Functional Data Analysis (FDA) has seen an impressive growth in the statistical research due to the more and more frequent production of complex data in many different contexts (healthcare, environmental, engineering, etc.). According to the FDA model, data can be seen as measurements of a given quantity (or set of quantities) along an independent and continuous indexing variable (time or space). Observations are treated as random functions and can be viewed as trajectories of stochastic processes defined on a given infinite dimensional functional space. Even if the research in FDA dates back to 1970s, the major achievements have been reached in the last decade, especially in the multivariate setting. Despite the usefulness of robust statistics in this field, their generalization to the functional framework is definitely not straightforward, due to the infinite-dimensional nature of the spaces embedding data. A possibility is then to leverage on the concept of depth measures in order to create proper order statistics to be used in a suitable robust inferential framework. Topics like efficient methods for outlier detection and related graphical tools, as well as inferential tools for testing differences and dependency among families of curves will be discussed, presenting challenging applications.

E0354: **Local depth measures for trajectories with a common origin**
Presenter: **Marcela Svarc**, Universidad de San Andres, Argentina
Co-authors: Lucas Fernandez Piana, Ana Justel
The trajectories of common origin or back-trajectories appear very commonly in problems of meteorology and ecology, where the aim is to determine the origin of particles dragged by the winds that reach a certain point. We want to study arriving to the Byers Peninsula, located at the western coast of the Livingston Island (South Shetland Islands, Antarctica), to establish the dispersal capability of microorganisms susceptible of colonizing newly exposed locations in a climate change scenario. Local depths are extremely useful, because they allow us to adequately describe data which may not be unimodal. We will give an adequate definition of local depth for the problem in question based on considering local depth on circular data which is then properly integrated into a local depth measure for the back-trajectories. We will show its main properties and analyze its performance and computational aspects.

E0740: **Total variation depth and its decomposition for functional-data outlier detection**
Presenter: **Huang Huang**, NCAR, United States
Co-authors: Ying Sun
There has been extensive work on data depth-based methods for robust multivariate data analysis. Recent developments have involved infinite-dimensional objects such as functional data. We propose a notion of depth, the total variation depth, for functional data, which has many desirable features and is well suited for outlier detection. The proposed depth is of the form of an integral of a univariate depth function. We show that the novel formation of the total variation depth leads to useful decomposition associated with shape and magnitude outlyingness of functional data. Compared to magnitude outliers, shape outliers are often masked among the rest of samples and more difficult to identify. We then further develop an effective procedure and visualization tools for detecting both types of outliers, while naturally accounting for the correlation in functional data.

E1666: **Nonparametric depth based methods for analyzing health data**
Presenter: **Sara Lopez Pintado**, Northeastern University, United States
Technological development in many emerging research fields has led to the acquisition of large collections of data of extraordinary complexity. In neuroscience for example, brain-imaging technology has provided us with complex collections of signals from individuals in different neurophysiological states in healthy and diseased populations. The development of statistical tools to analyze this type of high-dimensional data sets is very much needed. New robust methodologies for analyzing functional and imaging data based on the concept of depth are presented. Functional depth provides a rigorous way of ranking from center-outward a sample of functions. This ordering allows the definition of robust descriptive statistics such as medians, trimmed means and central regions for functional data. Moreover, data depth is often used as a building block for developing robust statistical methods and outlier-detection techniques. Permutation depth-based tests for comparing the location and dispersion of two groups of functions or images are proposed and calibrated. The performances of these methods are illustrated in simulated and real data sets. In particular, we tested differences between brain images of healthy controls and patients with severe depression. We also used these methods to analyze differences between the growth pattern of normal and obese children.

EO054 Room O2 ECOSTA JOURNAL PART B: STATISTICS II Chair: Ana Colubi

E0297: **Valid and consistent adaptive multiple tests**
Presenter: **Arnold Janssen**, Heinrich-Heine University Duesseldorf, Germany
Co-authors: Marc Ditzhaus
Simultaneous hypotheses testing for big data sets is a very difficult affair. First, the modern concept of multiple testing is introduced and examples are illustrated. Then, new results are presented. The pioneer multiple test, with up to date more than 42000 citations, is a basic tool in high dimensional data analysis, for instance in genomics, when a huge amount of tests are carried out simultaneously for the same data set. This test, and also improved data dependent adaptive tests, controls the so called FDR. The FDR is the expectation of the ratio of the number of false rejections and all rejections. Although the FDR can be controlled by some given level alpha, the false discovery proportion (FDP) may have stochastic fluctuations. We discuss the consistency for general adaptive multiple tests. We present finite sample and asymptotic results in order to bound deviations of the FDP from the present FDR level.

E1709: **HAC standard errors for robust estimators**
Presenter: **Christophe Croux**, Edhec Business School, France
Robust regression methods give reliable estimates in presence of outliers. Most of the research in robust regression focuses on point estimation, although the statistical inference part has been developed as well and is available in statistical software. We discuss the problem of obtaining valid standard errors for robust regression estimators when the error terms are heteroscedastic or serially correlated. Such standard errors are called HAC standard errors. We collect existing formulas and discuss up to what extent the HAC standard errors are robust with respect to outliers, including leverage points. We also obtain new results, as we could quantify the loss in efficiency when using a HAC standard error when in fact there is no need for it.

E0508: **Small area estimation of proportions under an area-level compositional mixed model**
Presenter: **Maria Jose Lombardia**, Universidade da Coruna, Spain
Co-authors: Domingo Morales, Maria-Dolores Esteban, Esther Lopez Vizcaino, Agustin Perez
An area-level compositional mixed model is introduced by applying an additive logistic transformation to a multivariate Fay-Herriot model. Small area estimators of the proportions of the categories of a classification variable are derived from the new model and the corresponding mean squared errors are estimated by parametric bootstrap. Several simulation experiments designed to analyze the behaviour of the introduced estimators are carried out. An application to real data from the Spanish Labour Force Survey of Galicia (north-west of Spain), in the first quarter of 2017, is given. The target is the estimation of domain proportions of people in the four categories of the variable labour status: under 16 years, employed, unemployed and inactive.

E1028: **Stochastic block models for social network data: Inferential developments**
Presenter: **Silvia Pandolfi**, University of Perugia, Italy
Co-authors: Francesco Bartolucci, Maria Francesca Marino
Stochastic Block Models (SMBs) have known a flowering interest in the social network literature. They provide a tool for discovering communities and identifying clusters of individuals characterized by similar social behaviors. According to the SBM specification, each individual in the network belongs to one of k distinct blocks, corresponding to the categories of a discrete latent variable, and the probability of observing a connection between two units only depends on their block memberships. In this framework, full maximum likelihood estimates are not achievable due to the intractability of the likelihood function. For this reason, several approximate solutions are available in the literature. These alternative approaches are mainly based on classification likelihood, composite likelihood or variational approximation. We propose a new and more efficient approximate method for estimating model parameters, which has a hybrid nature as it is based on a classification likelihood but has features in common also with full likelihood and composite likelihood inference. Moreover, it relies on an optimization algorithm with structure and numerical complexity similar to that of the variational approach, while being typically faster to converge. We illustrate the potential of the proposed approach by an intensive simulation study.

EO396 Room P2 BAYESIAN ANALYSIS AND APPLICATIONS VIA PARTITION AND UNIFICATION APPROACHES Chair: Hongsheng Dai

E0185: **Confusion: A confidential fusion approach to statistical secret sharing**
Presenter: **Murray Pollock**, University of Warwick, United Kingdom
Co-authors: Louis Aslett, Hongsheng Dai, Gareth Roberts
A surprisingly challenging problem in computational statistics is how to unify distributed statistical analyses and inferences into a single coherent inference. This problem arises in many settings (for instance, combining experts in expert elicitation, incorporating disparate inference in multi-view learning, and recombining in distributed big data problems), but a general framework for conducting such unification has only recently been addressed. A particularly compelling application is in statistical cryptography. Consider the setting in which multiple (potentially untrusted) parties wish to share distributional information (for instance in insurance, banking and social media settings), but wish to ensure information theoretic security (in particular, information is shared in such a way that another party with unbounded compute power could not determine secret information or data of any other party). Confusion, a confidential fusion approach to statistical secret sharing, is the first approach which addresses this important statistical application.

E1023: **A joint modeling approach for baseline matrix-valued imaging data and treatment outcome**
Presenter: **Bei Jiang**, University of Alberta, Canada
A unified Bayesian joint modeling framework is proposed for studying association between a binary treatment outcome and a baseline matrix-valued predictor, such as imaging data. Under this framework, a theoretically implied relationship can be established between the treatment outcome and the matrix-valued imaging data, although the imaging data is not directly considered in the model. The proposed joint modeling approach provides a promising framework for both association estimation and prediction. Properties of this method are examined using simulated datasets. In particular, our simulations show good performance of the proposed method under even difficult scenarios in which the sample size is small and/or the signal-to-noise (STN) in the imaging data is poor. Finally, a detailed illustration of the proposed modeling approach is provided using a motivating depression study that aims to explore the association between the baseline EEG data and the probability of a favorable response to an antidepressant treatment.

E1061: **Modular modelling: Joining and splitting models with Markov melding**
Presenter: **Robert Goudie**, University of Cambridge, United Kingdom
Integrating multiple sources of data into a joint analysis yields more precise estimates and reduces the risk of biases introduced by using only partial data. However, it can be difficult to conduct a joint analysis. Often it is only feasible in practice to take a modular approach, with each data source modelled separately, but this leads to information and uncertainty not being propagated. We propose to address this problem using a simple, general method, which requires only small changes to existing models and software. We first form a joint model based upon the original submodels

using a generic approach called Markov melding. We then show that this model can be fit in stages, rather than as a single, monolithic model. The approach also enables splitting of large joint models into smaller submodels, allowing inference for the original joint model to be conducted via our multi-stage algorithm. We demonstrate how the approach can be used to integrate longitudinal latent class models with Dirichlet process-based clustering models; to integrate intensive care unit A/H1N1 influenza data and other information sources; and jointly model ecological census and mark-recapture-recovery data.

E1162: **Bayesian fusion**
Presenter: **Hongsheng Dai**, University of Essex, United Kingdom
Co-authors: Murray Pollock, Gareth Roberts
An exact Bayesian fusion algorithm is presented, which can carry out perfect inferences for the unification of distributed data analysis. The new method uses parallel but coalesced Markov processes to drive distributed Monte Carlo draws to a Monte Carlo sample from the posterior of the full data. The Markov processes are simulated via path-space rejection sampling for diffusion processes. The methodology of this exact Bayesian fusion algorithm explained why existing methods do not provide good results and how to correct approximated draws of existing methods in order to obtain exact samples. Its approximate version, the sequential Bayesian fusion algorithm, can be implemented in parallel for big data analysis.

EO222 Room Q2 BAYESIAN SEMI- AND NONPARAMETRIC MODELLING III Chair: Matteo Ruggiero

C0941: **Bayesian nonparametric sparse VAR models**
Presenter: **Roberto Casarin**, University Ca Foscari of Venice, Italy
Co-authors: Monica Billio, Luca Rossini
In a high dimensional setting, vector autoregressive (VAR) models require a large number of parameters to be estimated and suffer from inferential problems. We propose a nonparametric Bayesian framework and introduce a new two-stage hierarchical Dirichlet process prior (DPP) for VAR models. This prior allows us to avoid overparametrization and overfitting issues by shrinking the coefficients toward a small number of random locations and induces a random partition of the coefficients, which is the main inference target of nonparametric Bayesian models. We use the posterior random partition to cluster coefficients into groups and to estimate the number of groups. Our nonparametric Bayesian model with multiple shrinkage prior is well suited for extracting Granger causality networks from time series, since it allows us to capture some common features of real-world networks, which are sparsity, blocks or communities structures, heterogeneity and clustering in the strength or intensity of the edges. In order to fully capture the richness of the data, it is therefore crucial that the model used to extract network accounts for weights associated to the edges. We illustrate the benefits of our approach by extracting network structures from panel data for shock transmission in business cycles and in financial markets.

E0240: **Verifiable posterior consistency conditions for jump diffusions**
Presenter: **Jere Koskela**, University of Warwick, United Kingdom
Co-authors: Dario Spano, Paul Jenkins
Jump diffusions are a flexible class of stochastic models for time series data, with abundant applications in a variety of fields. They are natural to specify in terms of function- and measure-valued coefficients, which motivates the use of nonparametric inference methods which are able to retain this level of modelling flexibility. However, likelihoods under jump diffusions are almost always intractable, which makes the development and analysis of methods challenging. We will introduce jump diffusions, as well as the machinery of Bayesian nonparametric inference for discretely observed jump diffusion trajectories. We will then present tractable sufficient conditions on the prior for posterior consistency, and examples of standard nonparametric priors which satisfy the consistency conditions.

E1358: **Bayesian Wasserstein deconvolution**
Presenter: **Catia Scricciolo**, University of Verona, Italy
The problem of recovering a distribution function from n i.i.d. observations additively contaminated with random errors whose distribution is known to the observer is considered. Implicit measurements occur quite often and one wishes to undo the errors inflicted on the signal. We investigate whether a nonparametric Bayes approach for modelling the latent distribution may result in valid asymptotic frequentist inferences under the 1-Wasserstein loss, which, originated in the optimal transport literature, has now applications in a broad spectrum of research areas such as statistics, economics, image processing, etc. The approach we adopt uses inversion inequalities relating distances between mixtures to the 1-Wasserstein distance between the corresponding mixing distributions to translate posterior contraction rates for mixtures into rates for mixing distributions. For finite mixtures, when the number of components is known up to an upper bound, Bayesian estimation of the mixing distribution can be performed at the best possible rate $n^{-1/4}$ (up to a log-term). If the mixing distribution is completely unknown, for convolutions with ordinary smooth errors, the rate $n^{-1/8}$ is obtained, up to a log-factor, for the Laplace error distribution. We discuss whether the prior law can be chosen to act as an efficient approximating scheme leading to the lower bound rate $n^{-1/5}$ recently obtained in the literature for a minimum distance estimator.

E1591: **Improper priors for nonparametric Bayes estimation of Poisson intensity functions**
Presenter: **Fumiyasu Komaki**, The University of Tokyo, Japan
The nonparametric Bayes estimation of intensity functions of inhomogeneous Poisson processes is investigated. It is shown that reasonable nonparametric Bayes estimators can be constructed by using improper priors, although improper priors have not been widely used for nonparametric Bayes estimation. We propose a class of improper priors for Poisson intensity functions. Nonparametric Bayes estimators based on the priors in our class are admissible under the Kullback-Leibler loss.

EG600 Room D1 CONTRIBUTIONS IN METHODOLOGICAL STATISTICS AND APPLICATIONS I Chair: Satish Iyengar

E1730: **A discrete survival model with a smooth baseline hazard for age at alcohol intake debut**
Presenter: **Alphonce Bere**, University of Venda, South Africa
Co-authors: Godfrey Sithuba
A discrete logit survival model is employed to investigate on the risk factors for early alcohol intake among students at two tertiary institutions in Thohoyandou, South Africa. Data were collected from a sample of 744 students using a self-administered questionnaire. Significant covariates were arrived at through a regularization algorithm implemented using the glmmLasso package. The tuning parameter was determined using cross-validation. The baseline hazard was modeled as a smooth function of time through the use of spline functions. The results show that the hazard of initial alcohol intake peaks at the age of about 16 years and that at any given time, being of a male gender, prior use of other drugs, having drinking peers, having experienced negative life events and physical abuse are associated with a higher risk of alcohol intake debut.

E1739: **A parallel regularized optimization approach to hedging a portfolio of collateralized mortgage obligations**
Presenter: **Yury Gryazin**, Idaho State University, United States

Collateralized Mortgage Obligations (CMO) can exhibit different degrees of cash flow variability depending on tranche structure and realized prepayment speeds. In order to have a meaningful comparison across structures and collateral types, an Option-Adjusted Spread (OAS) methodology is typically used. While OAS provides a meaningful improvement to yield to maturity for CMO, it does not fully account for embedded optionality. We introduce a related methodology which significantly improves valuation metric for CMOs. In this method, the construction of the optimal hedging policy is considered an essential part of the valuation procedure. To ensure the convergence of the resulting ill-conditioned optimization problem, the standard Tikhonov type regularization technique is applied. The convergence of this method and the stability of the underlying control optimization problem are discussed. The numerical results on the hedging of the portfolios of CMO and European swaptions based on parallel Monte-Carlo simulation on multicore PCs are presented.

E1751: **Refined measures of dynamic connectedness based on time-varying parameter vector autoregressions**
Presenter: **David Gabauer**, Johannes Kepler University, Austria

The dynamic connectedness measures are combined with a time-varying parameter vector autoregressive model (TVP-VAR) with a time-varying variance-covariance structure. This framework allows us to capture possible changes in the underlying structure of the data in a very flexible and robust manner. Specifically, there is neither need to arbitrarily set the rolling-window size nor a loss of observations in the calculation of the dynamic measures of connectedness as no rolling-window analysis is involved. Since this TVP-VAR-based connectedness framework rests on multivariate Kalman filters it is less outlier-sensitive than the originally proposed rolling-window VAR approach. Those merits are illustrated by conducting various Monte Carlo simulations. Moreover, we are investigating the dynamic transmission mechanism of the four most-traded foreign exchange rates by comparing the results of the TVP-VAR model with three different rolling-window VAR models. Finally, we introduce confidence intervals for TVP-VAR-based and rolling-window VAR-based dynamic connectedness measures.

CO410 Room P1 SPATIO-TEMPORAL MODELS FOR PREDICTION OF CLIMATE IMPACTS ON SOCIETIES Chair: Nourddine Azzaoui

C1593: **Efficient spatial designs for monitoring environmental phenomenon**
Presenter: **Pierre Druilhet**, Universite Clermont Auvergne, France
Co-authors: Sylvain Coly

Gaussian processes are often used as surrogate models for computer experiments or as current prior knowledge of spatio-temporal phenomenon. Space-filling designs aim to provide accurate estimate in the whole space. However, in a monitoring context, it is more relevant to focus on crucial areas. We propose three strategies to build spatial designs, depending on the purpose. The first one aim to focus on areas with high expected values, the second one on areas exceeding a given threshold and the third one on level set detection. We also propose some algorithms to construct efficient designs and we compare our designs with space-filling designs.

C1596: **Climate change mitigation management using an optimal stochastic control framework**
Presenter: **Daisuke Murakami**, The Institute of Statistical Mathematics, Japan
Co-authors: Matthew Ames, Pavel Shevchenko, Tor Andre Myrvoll, Tomoko Matsui, Yoshiki Yamagata

In climate change mitigation management, it is crucial to understand the uncertainty in carbon dioxide emissions and atmospheric concentration, temperature, and damage caused by climate change. We investigate a stochastic control framework in order to find an optimal strategy for mitigating climate change using the Bellman equation. In the proposed method, the state variables consist of carbon dioxide emissions, carbon dioxide concentration and temperature, whereas the control variable is the mitigation cost. The optimal mitigation strategy that minimizes the discounted sum of damage and mitigation cost functions is estimated. While the damage cost function links temperature with economic influence, the mitigation cost function links carbon dioxide emissions and mitigation cost with economic influence. Since there may exist some unknown factors and uncertainty in both functions and it is difficult to specify them beforehand, we stochastically model them so that we do not need to define them explicitly. Furthermore, we compare methods of solving the Bellman equation via reinforcement learning and deep reinforcement learning. In the simulation experiments for future several dozens of years with different scenarios, we study the characteristics of the strategy which is the most suitable for mitigating climate change.

C1601: **A comparative study on autoregressive models: An application to several financial assets**
Presenter: **Bing Xiao**, UCA University Clermont Auvergne, France
Co-authors: Marie-Eliette Dury

It has become more important for financial institutes to capture the volatility of a financial asset. The autoregressive conditional heteroscedasticity models assume that volatility is not constant. Actually, some of the results of research in the models performance seem conflicting and confusing. It seems that the Students t distribution characterizes better the heavy-tailed returns than the Gaussian distribution. Assets with higher kurtosis are better predicted by a GARCH model with Students distribution while assets with lower kurtosis are better forecasted by using an EGARCH model. Moreover, stochastic models such as stable processes appear as good candidates to take heavy-tailed data into account. These observations lead us to determine how well these different models perform in terms of forecasting volatility and will be assessed based on the forecasts they make. We attempt to model and forecast the volatility of different asset Index during the recent period. The questions to ask are: Which models capture the volatility better? If one model captures the volatility better, would it lead to a more efficient forecast accuracy? The assets concerned are the following: the gold market, the nickel market, the equity market and the maize market. This study contributes to the existing finance literature by investigating the U.S. stock market during the recent period.

C1634: **Spatial prediction based on kernel method and dimension reduction**
Presenter: **Anne Francoise Yao**, University Clermont Auvergne, France
Co-authors: Liliana Forzani, Maria Antonella Gieco, Pamela Llop

The problem of regression (or prediction) is a main concern when dealing with spatial data modelling. In this framework, kriging (and related linear or spatial generalized linear model) is a well-known and widely used method. However, krigging is suitable mainly when dealing with the spatial Gaussian process. In the case of non-Gaussian process, some alternatives have been proposed. Among them, the spatial kernel approach which is the subject of a large and dynamic literature of the last years. We are interested in the problem of predicting unknown spatial values of the process using both dependence information due to the spatial location on one hand, and to some exogenous variables on the other hand. In this setting, in some non-Gaussian cases, the kernel approaches outperform the krigging methods, but with small difference in the prediction errors. As a solution, we suggest combining the spatial kernel predictors with a dimension reduction methods. Then, we study this new spatial predictor. We illustrate our purpose through some simulations.

CO094 Room A2 SENTIMENT AND FINANCIAL MARKETS Chair: Francesco Audrino

C0228: **Application of multivariate Hawkes graphs to uncover Granger causality of financial news**
Presenter: **Anastasija Tetereva**, University St Gallen, Switzerland

A considerable amount of current research in financial business addresses the influence of news and social media on stock returns and volatility. Although news data are used in many applications, the mutual relationship among public announcements remains unclear. Moreover, the majority of studies are conducted using aggregated data, which are less effective in detecting causal links than observations of higher frequency. Evidence of self and mutual triggering of news announcements in the financial sector is provided. It is proposed that the news arrival times be modelled as a multivariate Hawkes process to test the Granger causality of company-specific news and to detect the most influential companies. Based on this information, a novel method of constructing a composite news intensity index (NII) is presented. The NII demonstrates the ability to timeously describe the uncertainty in financial markets. The proposed measure Granger causes VIX at 6-month lag and can therefore be used to diagnose the health of a financial system.

C0296: **Using large and heterogeneous sources of sentiment and attention data for predicting stock market volatility**
Presenter: **Fabio Sigrist**, Lucerne University of Applied Sciences, Switzerland
Co-authors: Francesco Audrino, Daniele Ballinari

The impact of sentiment and attention variables on volatility is analyzed by using a novel and extensive dataset that combines social media, news articles, information consumption, and search engine data. Applying a state-of-the-art sentiment classification technique, we investigate whether sentiment and attention measures contain additional predictive power for realized volatility when controlling for a wide range of economic and financial predictors. Using a penalized regression framework, we identify investors' attention, as measured by the number of Google searches on financial keywords (e.g. "financial market" and "stockmarket"), and the daily volume of company-specific short messages posted on StockTwits to be the most relevant variables. In addition, our study shows that attention and sentiment variables are able to significantly improve volatility forecasts, although the improvements are of relatively small magnitude from an economic point of view.

C0460: **Cryptocurrency-specific lexicon and sentiment projection**
Presenter: **Thomas Renault**, IESEG, France

A non-classical asset, cryptocurrencies, draws great attention to a particular subset of investors who possess higher risk preference in order to ride the trend. Due to limited knowledge for its fundamental value, investor opinions (sentiment) in this digital asset class could convey incremental information in terms of price discovery. Using a novel dataset of 1069K messages related to 375 cryptocurrencies posted on the microblogging platform Stocktwits during a 4-year period, we construct a "crypto-specific lexicon" in order to precisely capture the semantic orientations and avoid misspecification. The results show that in comparison with the financial lexicon used in the literature, the crypto-specific lexicon achieves 32% higher accuracy in terms of out-of-sample classification. The opinions quantified through crypto-specific lexicon strikingly drive the movement of cryptocurrency market. A time series return predictability further suggests our tone measure positively predicts Top 200 cryptocurrency index up to 3 days without return reversal.

C0495: **Country news sentiment indices**
Presenter: **Svetlana Borovkova**, Vrije Universiteit Amsterdam, Netherlands

Country-specific indices which measure the sentiment in the news about the regional stock markets are constructed and investigated. From all news items that hit Thomson Reuters newswire, we filter out those that are relevant for a country's large cap stock index and aggregate sentiment of these items into one comprehensive index, taking into account novelty and volume of news. This index can be thought of as a thermometer of the perception of a particular country's stock market in the media. We construct such indices for the US, Europe, the UK, Japan and China. We investigate the relationships of these indices to the stock market returns and volatilities, foreign exchange rates, sovereign debt spreads. We find evidence that the derived Country News Sentiment Indices ("CNSI") have positive contemporaneous and one-week-ahead relations to the corresponding stock market returns for all the considered markets apart from China. We also demonstrate strong relationships of these sentiment indices to FX, sovereign CDSs as well as country-specific economic indicators.

CO130 Room B2 HOUSING MARKETS Chair: Ivan Paya

C0209: **Speculative bubbles in regional housing markets**
Presenter: **Andre Anundsen**, Norges Bank, Norway
Co-authors: Bjoernar Kivedal

Quarterly data for the 182 metropolitan statistical areas in the US for the period 1985q1-2015q4 are used in order to test for rational bubbles in local housing markets. Applying tests for explosive roots in price-to-income ratios, we find evidence suggesting that 81 percent of the areas experienced a bubble in the period between 2000 and 2015. Conditional on bubble-detection, we date - stamp the start and end of bubbles across metro areas. We find that and both the dating and the duration of the bubble differ substantially. We also calculate a measure of the amplitude of the bubble. We find that areas with more restrictions on land supply were more prone to stronger bubble dynamics. We also find that population size, house price level and subprime lending affected the duration of the bubble.

C0419: **The impact of the UK real estate sector on systemic risk**
Presenter: **Ivan Paya**, Lancaster University, United Kingdom
Co-authors: Efthymios Pavlidis, Alexandros Skouralis

The extent to which real estate (RE) firms contribute to the systemic risk of the banking and financial industries in the UK is examined. We employ a CoVaR/DeltaCoVaR approach to identify not only the contribution of individual firms to the risk of the overall banking and financial system, but also what are the main firm characteristics driving the result. Larger, riskier and more leveraged RE firms appear to be the ones with a larger impact on systemic risk. We extend our analysis to a dynamic framework using four different methods that allow us to identify the periods where the effect on systemic risk is higher. We also employ an alternative, structural approach, to disentangle the marginal effects of housing market conditions, firm characteristics and macro variables on systemic risk.

C0765: **A leading indicator of house-price bubbles**
Presenter: **Simon Juul Hviid**, Danmarks Nationalbank, Denmark

Prior to the financial crisis in the mid-2000s, house prices increased dramatically and most economists agree that part of the increase in Danish house prices can be characterized as a house-price bubble. The emergence of a house-price bubble can have sizeable implications for macroeconomic as well as financial stability. A house-price bubble is often a result of self-exciting beliefs, leading to explosiveness of the developments in house prices. The dynamics of house prices in Denmark are investigated in order to identify emerging bubbles in due time. We develop a fundamentals-adjusted house price index and apply a previous testing procedure to date-stamp house-price bubbles. The empirical results identify developments in line with a price bubble from mid-2005 in Denmark. When applied to flats in Copenhagen, real price developments in 2015-16 indicate speculative behaviour but it cannot be ruled out that developments are driven by fundamental economic factors.

C1693: **Commercial real estate, housing and the business cycle**
Presenter: **Kostas Vasilopoulos**, Lancaster University Management School, United Kingdom
Co-authors: William Tayler
The 2007-08 financial crisis exposed the magnitude in which real estate markets and the construction sector can spillover to goods market, and the real sector. The aim is to develop a two-sector RBC model, consumption good and a construction sector, where both commercial and residential real estate forms the construction sector. We introduce income-producing real estate in a model with housing to investigate the property-price and investment dynamics, and their implications to the macroeconomic fluctuations. We allow for common characteristics for commercial and residential real estate production, where the inputs of technology, capital, and labor have the same structure and are determined separately. Whilst we treat the production of real estate independently they share risks, and they compete for their final input, land, which is fixed on aggregate. The borrowing capacity of the entrepreneurs is constrained by the value of their collateral assets, where they differ depending on the type of assets available in each sector. We show that a housing demand shock generates a complementarity effect between commercial and residential real estate. However, the degree of complementarity depends on the land regime i.e. land regulations that determine the substitutability of residential and commercial land (e.g. zoning and land use). Policies that focus on land allocations can mitigate this effect and thus restrict the amount of spillover to the real economy.

CO466 Room C2 ASSET PRICING WITH FINANCIAL FRICTIONS Chair: Luca Benzoni

C1523: **SONOMA: a Small Open ecoNOmy for MAcrofinance**
Presenter: **Mariano Croce**, Bocconi, Italy
A new small open economy model with recursive preferences is developed in which international and corporate debts are jointly determined at the equilibrium. The economy faces both internal credits shocks and shocks to the cost of external debt. We quantify the relevance of credits shocks and credit frictions in the determination of the equity risk premium and offer a novel benchmark model designed to replicate key properties of small EU countries. Our setting is flexible and can be easily extended to address fiscal, monetary, and macro-prudential policies in an economy in which equity risk matters.

C1608: **Comparative valuation dynamics in models with financing restrictions**
Presenter: **Fabrice Tourre**, Copenhagen Business School, Denmark
Co-authors: Lars Hansen, Paymon Khorrami
A theoretical framework is developed to nest many recent dynamic stochastic general equilibrium economies with financial frictions into one common generic model. The goal is to study the macroeconomic and asset pricing properties of this class of models, identify common features and highlight areas where these models depart from each other. In order to characterize the asset pricing implications of this family of models, we study their term structure of risk prices and risk exposures, the natural extension of impulse response functions in economic environments exhibiting non-linear behaviors.

C1107: **Monetary policy and reaching for income**
Presenter: **Lorenzo Garlappi**, University of British Columbia, Canada
Co-authors: Kent Daniel, Kairong Xiao
The impact of monetary policy on investors' portfolio choice and asset prices is studied. Using data on mutual fund flows and individual trading records, we find that low-interest-rate monetary policy increases investors' demand for high-dividend stocks and drives up their asset prices. The increase in demand is more pronounced among investors who live off dividend income for consumption. To explain these empirical findings, we develop a portfolio choice model in which investors have quasi-hyperbolic time preferences and use dividend income as a commitment device to curb their tendency to over-consume in the short-run. When accommodative monetary policy lowers interest rates, it reduces the income stream from bonds and induces investors who want to keep a desired level of consumption to "reach for income" by tilting their portfolio towards high-dividend stocks. Our finding suggests that low-interest-rate monetary policy may lead to under-diversification of investors' portfolios and may cause redistributive effects across firms that differ in their dividend policy.

C1046: **The term structure of credit spreads with dynamic debt issuance and asymmetric information**
Presenter: **Luca Benzoni**, Federal Reserve Bank of Chicago, United States
Co-authors: Lorenzo Garlappi, Robert Goldstein
Credit spreads and capital structure dynamics are investigated in a model in which management has private information regarding firm value and is able to issue both equity and debt to service existing debt. Rather than choosing to default, managers of investment-grade (IG) firms who receive bad private signals conceal this information by servicing existing debt via new debt issuance. As such, firms with IG-commensurate spreads have zero jump-to-default risk (and hence, command zero jump-to-default premium), at least until their debt capacity is fully utilized and spreads have increased to "fallen angel" status. These predictions match observation well.

CO462 Room G2 MACHINE LEARNING TECHNIQUES FOR TIME SERIES FORECASTING Chair: Lyudmila Grigoryeva

C1731: **Tipping point analysis of dynamical**
Presenter: **Valerie Livina**, National Physical Laboratory, United Kingdom
Tipping point analysis helps anticipate, detect and forecast tipping points in a dynamical system. The methodology combines monitoring memory in a time series with potential analysis that visualizes and extrapolates the system states. Early warning signal indicators are based on autocorrelation, power-law scaling exponent of detrended fluctuation analysis, and recently developed power-spectrum-based indicator. When indicators rise monotonically, this signals an upcoming transition or bifurcation. By combining several indicators, it is possible to distinguish different types of tipping, such as forced transitions and genuine bifurcations. The potential analysis detects a transition or bifurcation in a series at the time when it happens, which is illustrated in a colour plot mapping the potential dynamics of the system. Potential analysis is also used in forecasting time series by extrapolation of Chebyshev approximation coefficients of the kernel distribution, with reconstruction of correlations in the data. The methodology has been extensively tested on artificial data and on observed datasets, and proved to be applicable to trajectories of dynamical systems of arbitrary origin.

C1691: **Inference and time series analysis with artificial neural networks**
Presenter: **Gerhard Fechteler**, Universitat Konstanz, Germany
Multilayer Perceptrons (MLP) have become a popular tool for nonlinear data analysis and proved to be useful for forecasting economic variables, outperforming not only linear models, but also other machine learning frameworks. Firstly, the asymptotic normality of the parameters of an MLP is discussed and an algorithm for estimating the asymptotic covariance even under misspecification of the network structure is provided. Moreover, also the marginal effects from an MLP are shown to be normally distributed. An algorithm to estimate the asymptotic covariance matrix is presented, such that confidence bounds for the marginal effects can be constructed. Based on the estimated distribution of the marginal effects, a local Granger causality test is proposed. It allows us to detect causal relationships between time series variables that are only present in local

regions in the parameter space. To study high-dimensional time series data, a generalization of dynamic factor models (DFM) is proposed, which relaxes two critical assumptions of DFMs that are estimated via principal component analysis: linear dependency of the variables on the factors and orthogonality of the factors. The proposed framework is therefore less prone to misspecification.

C1631: **Forecasting and the universality problem in dynamic machine learning**
Presenter: **Juan-Pablo Ortega**, University St. Gallen, Switzerland
Co-authors: Lyudmila Grigoryeva
A relatively recent family of dynamic machine learning paradigms known collectively as reservoir computing is presented which is capable of unprecedented performances in the forecasting of deterministic and stochastic processes. We will then focus on the universal approximation properties the most widely used families of reservoir computers in applications. These results are a much awaited generalization to the dynamic context of previous well-known static results obtained in the context of neural networks.

C1662: **Forecasting of high-dimensional realized covariances with reservoir computing**
Presenter: **Lyudmila Grigoryeva**, University of Konstanz, Germany
Co-authors: Oleksandra Kukharenko, Juan-Pablo Ortega
The problem of forecasting high-dimensional realized covariance (RV) matrices computed out of intraday returns of the components of the S&P 500 market index is considered. The study focuses on a novel machine learning paradigm known as reservoir computing (RC) for producing multistep ahead forecasts for time series of realized covariances. Various families of reservoir computers have been recently proved to have universal approximation properties when processing stochastic discrete-time semi-infinite inputs. The goal is to implement with reservoir computers the forecasting of realized covariances. We examine the empirical performance of RC in comparison with many conventional state-of-the-art econometric models for various RV estimators, periods, and dimensions. We show that universal RC families consistently demonstrate superior predictive ability for various designs of empirical exercises.

CO532 Room H2 EMPIRICAL MACROECONOMICS Chair: Nora Traum

C0197: **Low-frequency fiscal uncertainty**
Presenter: **Zhao Han**, College of William and Mary, United States
Fiscal variables' steady-state values, or synonymously, the fiscal targets, are usually assumed to be known to households inside the economy. In reality, this is rarely the case. The aim is to investigate effects of low-frequency fiscal uncertainty in an incomplete information, anticipated utility environment in which households are learning unknown fiscal targets. Highly persistent fiscal movements cause households to suspect fiscal targets may be time-varying, even though the underlying targets are all assumed time-invariant. An RBC model with a detailed fiscal section that features government spending, lump-sum transfers, risk-free debt and capital and labor taxes is fit to US data to estimate deep structural parameters. Small deviations of households' beliefs on fiscal targets generate vastly different dynamics on how shocks influencing the real economy. Ignoring low-frequency fiscal uncertainty could lead to qualitatively and quantitatively misleading policy evaluations across all horizons.

C0204: **On the convexity of supply curves convex: Implications for state-dependent responses to shocks**
Presenter: **Christoph Boehm**, UT Austin, United States
Co-authors: Nitya Pandalai Nayar
The aim is to study whether responses to shocks are state-dependent. To guide our empirical analysis we develop a putty-clay model in which short-run capacity constraints generate a convex supply curve at the industry level. Using a sufficient statistics approach, we estimate the model and find strong support for state-dependent responses to shocks. Industries with low initial capacity utilization rates expand production much more after dollar depreciations or defense spending shocks than industries that produce close to their capacity limit. Further, prices rise after such demand shocks only if the initial level of capacity utilization is high. Our evidence supports the view that supply curves are convex at the industry level and suggests that policies that raise demand are more effective during slumps.

C0213: **International linkages and the changing nature of international business cycles**
Presenter: **Wataru Miyamoto**, University of Hong Kong, Hong Kong
Co-authors: Thuy Lan Nguyen
The effects of changes in international input-output linkages on the nature of business cycles are quantified. We build a multi-sector multi-country international business cycle model that matches the input-output structure within and across countries. We find that, in our 23 countries sample with manufacturing and non-manufacturing sectors, changes in the input-output linkages within and across borders between 1970 and 2007 causes a 21% drop in output volatility in a median country and a small increase in cross-country output correlations. Our decomposition attributes about 11% of the drop in output volatility to changes in international input-output linkages.

C0193: **Trade flows and fiscal multipliers**
Presenter: **Matteo Cacciatore**, HEC Montreal, Canada
Co-authors: Nora Traum
Novel insights on the role of international trade following unanticipated government spending and income tax changes in a flexible exchange rate environment are presented. We show fiscal multipliers can be larger in economies more open to trade, even when fiscal expansions imply a trade deficit. For a given trade openness and trade elasticity, the details of the fiscal environment—the relative composition of public and private imports and how the government finances its budget—dictate whether trade linkages enhance a fiscal stimulus. We demonstrate these ambiguous effects analytically in a simple two-country, two-good model. Using a Bayesian prior-predictive analysis, we show a quantitative international business-cycle model—featuring a rich fiscal specification and microfounded trade structure—bears the same agnostic predictions. We estimate the model on Canadian and U.S. data and find medium-run Canadian government spending multipliers are higher than in a counterfactually closed economy. Income tax cuts generate lower multipliers but are more effective in inducing positive cross-country comovement.

CO198 Room M2 MACROECONOMIC UNCERTAINTY Chair: Svetlana Makarova

C0330: **Uncertainty, perception and internet**
Presenter: **Roberto Golinelli**, University of Bologna, Italy
Co-authors: Maria Elena Bontempi, Matteo Squadrani
Macroeconomic uncertainty comprises unobservable, heterogeneous and uncertain components. We do not know when economic agents perceive uncertainty and which component of uncertainty may arise or not their interest (and reactions). We created the EURQ index measuring the volumes of economic uncertainty related queries on the Web. The EURQ catches peoples interest/need to gather more information when they are worried and uncertain, and that some topics spontaneously drive more interest. Results can be exploited to improve forecasting and policy evaluation, and to create separate EURQ indexes (macro-real, financial, political) helpful in solving identification and endogeneity of the commonly used uncertainty indexes.

C0585: **Identifying uncertainty shocks using geopolitical swings in Korea**
Presenter: **Seohyun Lee**, Bank of Korea, Korea, South
Co-authors: Jongrim Ha, Inhwan So
Using a novel set of instrumental variables in a structural VAR framework, we investigate the economic impact of uncertainty shocks from geopolitical swings in South Korea. We construct robust instrumental variables for geopolitical swings by observing high-frequency changes in financial asset returns and their volatilities around such geopolitical events. Our empirical results show that heightened (reduced) geopolitical uncertainty has negative (positive) impact on macroeconomic outcomes in South Korea. We provide evidence that financial and capital markets - fluctuations in exchange rates and sovereign spreads, changes in financial asset prices and market volatility, and swings in foreign investments - play an important role in the transmission of uncertainty shocks.

C0605: **On the effect of the Bank of England expected inflation uncertainty on private forecasters' risk assessments**
Presenter: **Carlos Diaz**, University of Leicester, United Kingdom
There is a large body of literature on the effect of central bank forecasting and communication on the formation of private agents' expectations. Most studies focus on the extent by which central bank point forecasts act as focal points for private forecasts, or for their dispersion. The aim is to determine whether measures of expected inflation uncertainty drive to some extent private agents' inflation risk assessment. A measure of inflation uncertainty based on the revisions of the Bank of England's density forecasts has been previously proposed. This measure accounted for the level of uncertainty that the bank expects that information perceived in a given quarter will have on future inflation. The determinants of this measure is analysed using a structural dynamic factor model formed of 50 key macroeconomic and financial variables of the British economy that tries to mimic the bank's process of information acquisition. Once the reaction function of the bank's measure of uncertainty is analysed, we study whether it has any effect on the responses to the survey of external forecasters. Given the importance of private inflation risk assessment on the asset pricing mechanisms, knowing the extent to which central banks can affect private expectations in this way could be used as another measure of unconventional monetary policy.

C0891: **Inflation forecast uncertainty in the three transitional Central and East European countries**
Presenter: **Svetlana Makarova**, University College London, United Kingdom
Co-authors: Wojciech Charemza, Carlos Diaz
It is argued that for some countries simple measures of uncertainty based on inflation forecast errors might be more useful than the uncertainty indices based on the external text search indices which use the simple phrase and term counting method. We consider the three post-communist countries – Poland, Russia, and Ukraine – and the US as a benchmark for comparison. The distribution of inflation forecast uncertainty has been modelled with the use of the weighted skew-normal (WSN) distribution. The fit of the WSN distribution has been compared with that of other frequently used distributions, like the two-piece normal and generalized beta. The Monte Carlo study has shown that it is difficult to distinguish, on the statistical grounds, between these three distributions. Our results show that the WSN provides a good fit for Poland and the US, that is, for countries with a clearly defined monetary policy, but not for Russia and Ukraine. The estimates of the distributional parameters suggest the prevalence of anti-inflationary policy over the output-stimulating policy for Poland, while the opposite seems to be true for the US.

CO538 Room N2 LONG MEMORY Chair: Christian Leschinski

C0481: **Forecasting long memory via a VAR model**
Presenter: **Guillaume Chevillon**, ESSEC Business School, France
Co-authors: Luc Bauwens, Sebastien Laurent
A large dimensional vector autoregressive (VAR) model can generate long memory in its components under conditions which restrict the VAR parameters. We compare the forecasting performance of univariate ARFIMA and HAR models, a VAR estimated by ML under the CHL constraints, and a VAR estimated by MCMC. The latter is based on a Gaussian prior density that incorporates the CHL restrictions through the prior mean of the VAR parameters, while the prior variances control the tightness of the restrictions. The forecast comparisons are done on simulated and real data.

C0641: **Spurious fractional cointegration**
Presenter: **Christian Leschinski**, Leibniz University Hannover, Germany
Co-authors: Philipp Sibbertsen
For univariate time series it is well documented that low frequency contaminations generate spurious long memory. This analysis is extended to vector valued processes. We propose a rigorous definition of spurious fractional cointegration, we show that such a behavior will occur in processes with joint low frequency contaminations and standard estimation of the cointegrating rank can spuriously indicate fractional cointegration in these situations. To deal with multivariate low frequency contamination, we derive a robust multivariate local Whittle (RMLW) estimator for the memory parameters and the cointegrating vector that is consistent and asymptotically normal in the presence of low frequency contaminations and spurious fractional cointegration.

C0949: **Estimation pitfalls when the noise is not i.i.d.**
Presenter: **Liudas Giraitis**, Queen Mary University of London, United Kingdom
Co-authors: Masanobu Taniguchi, Murad Taqqu
Whittle estimation is extended to linear processes with a general stationary ergodic martingale difference noise. We show that such an estimation is valid for standard parametric time series models with smooth bounded spectral densities, e.g. ARMA models. Furthermore, we clarify the impact of the hidden dependence in the noise on such estimation. We show that although the asymptotic normality of the Whittle estimates may still hold, the presence of dependence in the noise impacts the limit variance. Hence, the standard errors and confidence intervals valid under i.i.d. noise may not be applicable and thus require correction. The goal is to raise awareness to the impact of a non i.i.d. noise in applied work.

C1411: **Fixed bandwidth CUSUM tests for change-in-mean under long memory**
Presenter: **Kai Wenger**, Institute of Statistics, Germany
Co-authors: Christian Leschinski
Testing for mean-shifts in time series regression where the errors may exhibit long memory is considered. We propose four modified versions of the CUSUM test that apply kernel-based fixed-b and fixed-M long-run variance estimators. It is shown that the test statistics have a well-defined limiting distribution under long-range dependence that only depends on the long-memory parameter. We further discuss the bandwidth choice of all tests and show in an extensive Monte Carlo simulation study that our procedures perform best among all existing change-in-mean tests under long memory.

CC646 Room D2 CONTRIBUTIONS IN RISK ANALYSIS Chair: Marc Paolella

C1217: **Peer to peer personal credit risk assessment based on survival model**
Presenter: **Rui Liang**, University of Chinese Academy of Sciences, China
The credit risk problem of peer to peer online lending is increasingly prominent, and the default rate is a key parameter for quantifying credit risk. Therefore, it is particularly important to effectively calculate the methods and models for default events. Taking a large amount of transaction data as a sample, through survival analysis to determine the key factors affecting default and construct a loan default model, and use Cox regression to analyze when the borrower defaults and draw a loan survival curve. The empirical results show that education, credit rating, credit limit, number of loans, academic qualifications, real estate certification and default rate are negatively correlated, and positively correlated with loan survival time.

C1481: **A new approach to backtesting and risk model selection**
Presenter: **Ilaria Peri**, Birkbeck-University of London, United Kingdom
Co-authors: Jacopo Corbetta
Backtesting risk measures represent a challenge and complex methods are often required. We propose a new framework for backtesting that can be applied to every law invariant risk measures. We base our approach on the formalization of the concept of level of coverage associated with the risk model as defined in the original Basel Accord. Thus, we propose two simple hypothesis tests based only on results of probability theory without requiring any approximation or simulation. In addition, within this new framework, we introduce a methodology for selecting the best performing risk model among all the existing alternatives. This proposal adds value to the current state of the art, since using the traditional loss function approach, any comparison among forecasting outcomes of different risk models appeared to be meaningless. A series of simulation studies show that our hypothesis tests provide similar size and power to the classical binomial tests of value at risk and well-known tests of expected shortfall. A final experiment on real data allows determining the best risk measure procedures among the value at risk, expected shortfall, expectiles and lambda value at risk in different time windows over more than 40 years of daily data.

C1571: **Option implied tail risk and expected stock returns**
Presenter: **Conall OSullivan**, University College Dublin, Ireland
Co-authors: Yan Wang
The use of a Cornish-Fisher expansion is proposed in order to estimate the value-at-risk and expected shortfall of a risk-neutral distribution using model-free option-implied moments. We extract these risk-neutral tail risk measures from SP-500 index options on a daily basis and average over each month to obtain a monthly time series of forward-looking value-at-risk and expected shortfall across different tenors. These risk-neutral tail measures predict the equity risk premium at longer horizons (6 months or more) and are robust to the inclusion of other option implied predictors such as the variance risk premium. The risk-neutral tail measures are a priced factor in the cross-section of stock returns as stocks with higher loadings on the RN tail measures earn lower returns on average in subsequent months. This is consistent with these stocks acting as a hedge against perceived tail risk.

C1229: **Nonparametric risk-neutral density estimation using local cubic polynomials applied to intraday data**
Presenter: **Ana Monteiro**, University of Coimbra, Portugal
Co-authors: Antonio Santos
A new approach is considered to estimate risk-neutral densities (RND) within a kernel regression framework, through local cubic polynomial estimation using intraday data. There is a new strategy for the definition of a criterion function used in nonparametric regression that includes calls, puts, and weights in the optimization problem associated with parameters estimation. No-arbitrage restrictions are incorporated in the problem through equality and bound constraints. This yields directly density functions of interest with minimum requirements needed. Within a simulation framework, it is demonstrated the robustness of proposed procedures. Additionally, RNDs are estimated through option prices associated with two indices, S&P500 and VIX.

CC653 Room E2 CONTRIBUTIONS IN EMPIRICAL FINANCE Chair: Massimo Guidolin

C1433: **Oil price shocks and debt in the oil industry: An empirical analysis**
Presenter: **Christoph Funk**, Justus-Liebig-University Giessen, Germany
Co-authors: Karol Kempa, Johannes Lips
The effects of oil price shocks on the world economy have been extensively studied over the last decade. Yet, little is known of the effects of these on the U.S. oil industry on a firm level basis. We fill this gap by examining the relationship between (adverse) oil price shocks and the response of oil firms, the market for corporate loans and the impact on a firms capital structure. In particular, we will give an answer to the question of how oil firms respond to oil-price shocks and how these shocks affect their borrowing decision and creditworthiness. We combine data on individual syndicated loans with data from corporate financial statements for an analysis of a companies' financial decisions in great detail. First, we evaluate companies' borrowing behaviour and the loan characteristics, depending on the company's financial situation. Thereby, we gain insight in the relationship between an energy firms capital structure and their reaction to oil price shocks. Moreover, we differentiate companies along the oil industry's value chain. This allows us to determine the impact and exposure to price shocks depending on the position in the value chain. The overall findings of our research highlight the importance to monitor all financing channels of companies in order to be able to react to unforeseen deteriorations of market conditions.

C1320: **On the market model as a counterfactual for event studies in finance**
Presenter: **Carlos Castro**, Universidad del Rosario, Colombia
A common framework is provided that relates traditional event study estimation methods in finance to a modern approach for causal event studies. The framework provides a model for abnormal returns that nest the market model (the traditional approach) and more recent approaches based on difference-in-differences and synthetic control methods. We show that a synthetic control method in this context can be understood as a synthetic portfolio. We provide a simulation exercise and an empirical application, using mergers and acquisitions as the event of interest, to evaluate the performance of the different models within the framework. The results indicate that the performance of the market model, as a good counterfactual, depends on the distribution of the weights of the index, which is unfortunately overlooked in many empirical applications.

C1544: **ESG transparency and investment: Signaling and the power of social responsibility on performance**
Presenter: **Ramon Bermejo Climent**, Universidad Pontificia Comillas (Madrid, Spain), Spain
Co-authors: Isabel Catalina Figuerola-Ferretti Garrigues, Alvaro Santos Moreno
The impact of Environmental Social and Governance (ESG) disclosure in corporate equity performance is illustrated. Over the past decade there has been an increasing concern about the ethical impact of investment decisions. Hence, ESG factors have gained increased relevance in the corporate management process. We use an extensive data set of European ESG disclosure scores to demonstrate that ESG investing is associated with improved stock performance. Using a sample covering the 2005-2016 period of European Corporate and ESG compliance data a factor analysis demonstrate that ESG compliance scores become highly important in the asset allocation process.

C1716: **Banks' credit ratings: Impact of the business lines**
Presenter: **Patrycja Chodnicka - Jaworska**, University of Warsaw, Poland

The purpose is to analyze the impact of the business lines using linear decomposition of credit ratings on banks notes. A literature review was made and the following research hypothesis has been put: There exists the impact of the business lines on the banks' credit ratings. There were used ordered logit panel data models, and as a dependent variable there were taken the long-term issuer credit ratings published by the three largest rating agencies, Fitch, S&P and Moody. There have been collected data from the Thomson Reuters Database. The data were collected for the years 1990 - 2017. As independent variables there have been used CAMEL factors. Data was obtained from the World Bank and Thomson Retuters database.

CG059 Room F2 CONTRIBUTIONS IN TIME SERIES II Chair: Michael Thornton

C0217: **CAMPLET: Seasonal adjustment without revisions**
Presenter: **Barend Abeln**, Investment consultant, Netherlands
Co-authors: Jan Jacobs

Seasonality in macroeconomic time series can 'obscure' movements of other components in a series that are operationally more important for economic and econometric analyses. Indeed, in practice, one often prefers to work with seasonally adjusted data to assess the current state of the economy and its future course. A seasonal adjustment program called CAMPLET, an acronym of its tuning parameters, is presented which consists of a simple adaptive procedure to separate the seasonal and the non-seasonal component from an observed time series. Once this process is carried out there will be no need to revise these components at a later stage when new observations become available. Recently, two most widely used seasonal adjustment methods, Census X-12-ARIMA and TRAMO-SEATS, merged into X-13ARIMA-SEATS to become a new industry standard. The main features of CAMPLET are described, and a brief review of X13ARIMA-SEATS is provided. We compare and contrast CAMPLET with X-13ARIMA-SEATS. We evaluate the outcomes of both methods in a controlled simulation framework using a variety of processes.

C1641: **On the reliability of bootstrapped cointegration test findings**
Presenter: **Sven Schreiber**, Macroeconomic Policy Institute IMK and Free U Berlin, Germany

As applied cointegration analysis faces the challenge that (a) potentially relevant variables are unobservable and (b) it is uncertain which covariates are relevant, partial systems are often used and potential (stationary) covariates are ignored. By simulating hypothesized larger systems earlier findings seemed to suggest that a nominally significant cointegration outcome using a bootstrapped rank test in the bivariate sub-system might be due to test size distortions. We review the issue and the claim systematically. We find only mild size distortions, except when the specified data-generating process includes a large borderline-stationary root, reflecting an earlier insight from the literature. It turns out that when revisiting the earlier application of a long-run Phillips curve (inflation and unemployment), the interpretation of the bivariate cointegration test result (for the euro area) hinges on the assumed persistent output gap measure in the background.

C1636: **Estimation of permanent and temporary shocks in a factor model framework**
Presenter: **Carlos Montes-Galdon**, European Central Bank, Germany

An efficient algorithm is presented to recover permanent and temporary shocks from a factor model. There are several contributions. First, the algorithm makes use of a new way to introduce sign restrictions in a reduced form model to identify structural shocks. Moreover, via a state space model with restrictions on the persistence of the shocks, we can disentangle the impact of, for example, permanent productivity shocks or temporary ones. Also on the technical side, it is well known that in factor models the estimates of the loadings have very slow convergence properties. This problem is also present (and worsened) when we introduce sign restrictions. We propose an extension of very recent econometric techniques to overcome this problem. The final algorithm is extremely efficient and reliable. Then, based on restrictions that come from a DSGE model, looking at both changes in the steady state and the cyclical components, we estimate a factor model under the presence of both permanent and temporary developments. One of the main findings is that the great recession was mostly driven by permanent structural shocks (mainly investment specific shocks) which could explain the slow recovery.

C1627: **Term structure of variance risk premium and returns' predictability**
Presenter: **Giacomo Bormetti**, University of Bologna, Italy
Co-authors: Fulvio Corsi, Adam Aleksander Majewski

An analytic relation between equity risk premium and the term structure of variance risk premium (VRP) is derived. Motivated by this result, we estimate the VRP term structure using a general and fully analytical discrete-time option pricing framework featuring multiple volatility components and multiple risk premia. We confirm the importance of VRP in improving option pricing performances and show the ability of multi-component GARCH models to produce realistic hump-shaped VRP term structure. We finally uncover the strong predictive power of the shape of the VRP term structure on future stock-index returns.

CG057 Room I2 CONTRIBUTIONS IN AUTOREGRESIVE MODELS Chair: Yohei Yamamoto

C1370: **Testing for observation-dependent regime switching in mixture autoregressive models**
Presenter: **Mika Meitz**, University of Helsinki, Finland
Co-authors: Pentti Saikkonen

Testing for regime switching when the regime switching probabilities are specified either as constants ("mixture models") or are governed by a finite-state Markov chain ("Markov switching models") are long-standing problems that have also attracted recent interest. Testing for regime switching is considered when the regime switching probabilities are time-varying and depend on observed data ("observation-dependent regime switching"). Specifically, we consider the likelihood ratio test for observation-dependent regime switching in mixture autoregressive models. The testing problem is highly nonstandard, involving unidentified nuisance parameters under the null, parameters on the boundary, singular information matrices, and higher-order approximations of the log-likelihood. We derive the asymptotic null distribution of the likelihood ratio test statistic in a general mixture autoregressive setting using high-level conditions that allow for various forms of dependence of the regime switching probabilities on past observations, and we illustrate the theory using two particular mixture autoregressive models. The likelihood ratio test has a nonstandard asymptotic distribution that can easily be simulated, and Monte Carlo studies show the test to have satisfactory finite sample size and power properties.

C0234: **Autoregressive tempered fractionally integrated moving average time series: Theory and applications**
Presenter: **Farzad Sabzikar**, Iowa State University, United States

The autoregressive tempered fractionally integrated moving average (ARTFIMA) time series model applies a tempered fractional difference to the standard ARMA time series. The ARTFIMA model can also be interpreted as an extension of the ARFIMA model. ARTFIMA time series exhibit semi-long range dependence: Their covariance function resembles long range dependence for a number of lags, depending on the tempering parameter, but eventually decays exponentially fast. The mathematical foundation for ARTFIMA parameter estimation will be discussed. A new R package artfima to fit data will be presented. Several examples from finance, geophysics, turbulence, and climate illustrate the fitting procedure,

and the utility of the ARTFIMA model. Finally, an invariance principles for ARTFIMA times series will be given when the tempering parameter depends on the sample size.

C1675: **Bernoulli vector autoregressive model**

Presenter: **Carolina Euan**, King Abdullah University of Science and Technology, Saudi Arabia
Co-authors: Ying Sun

Categorical time series appear in many fields such as biology, industry, stocks markets and environmental sciences. Even for univariate binary time series, the analysis is usually more challenging than time series analysis for continuous variables. In a multivariate setting, modeling the dynamics in multiple binary time series is not an easy task. Most existing methods model the joint transition probabilities from marginals pairwisely. However, the resulting cross dependency may not be flexible enough. We propose a vector autoregressive (VAR) model for multivariate binary time series. The model is constructed by latent multivariate Bernoulli random vectors. The Bernoulli VAR model represents the instantaneous dependency between components via latent processes, and the autoregressive structure represents a switching between the hidden vectors depending on the past. Our proposed model provides an intuitive interpretation when analyzing real data sets. We derive the mean and matrix valued autocovariance function for the Bernoulli VAR model analytically and develop a likelihood based inference.

C0556: **Semi-parametric realized nonlinear conditional autoregressive expectile and expected shortfall models**

Presenter: **Chao Wang**, The University of Sydney, Australia
Co-authors: Richard Gerlach

A joint conditional autoregressive expectile and expected shortfall framework is proposed. The framework is extended through incorporating a measurement equation which models the contemporaneous dependence between the realized measures and the latent conditional expectile. Nonlinear threshold specification is further incorporated into the proposed framework. A Bayesian Markov Chain Monte Carlo method is adapted for estimation, whose properties are assessed and compared with maximum likelihood via a simulation study. One-day-ahead VaR and ES forecasting studies, with seven market indices and two individual assets, provide empirical support to the proposed models.

Sunday 16.12.2018 16:50 - 18:05 Parallel Session O – CFE-CMStatistics

EO488 Room Aula 4 MODEL SELECTION AND INFERENCE Chair: Ulrike Schneider

E0362: **Valid confidence intervals for post-model-selection predictors**
Presenter: **Francois Bachoc**, Universite Paul Sabatier, France
Co-authors: Hannes Leeb, Benedikt Poetscher
Inference post-model-selection in linear regression is considered. In this setting, it has been recently introduced a class of confidence sets, the so-called PoSI intervals, that cover a certain non-standard quantity of interest with a user-specified minimal coverage probability, irrespective of the model selection procedure that is being used. We generalize the PoSI intervals to confidence intervals for post-model-selection predictors.

E1258: **Inference in data with high-dimensional dependence structures**
Presenter: **Damian Kozbur**, University of Zurich, Switzerland
Co-authors: Christian Hansen, Jianfei Cao, Lucciano Villacorta
An inference approach is presented for dependent data in spatial applications. We consider a setting in which a high-dimensional parametric SAR model approximates the score process of a statistical model of interest. The method selects from among a large set of candidate spatial weight matrices which characterize dependence in the score process across observations. In a second step, we estimate standard errors that are robust to cross-sectional correlation structures implied by the selected spatial weight matrices. We show that the resulting procedure defines an inferential strategy which is robust against a flexible class of spatial dependence structures. We provide simulation evidence that shows the procedure outperforms conventional inference procedures.

E1270: **Post-selection inference in correlation learning**
Presenter: **Kory Johnson**, University of Vienna, Austria
Forward stepwise regression provides an approximation to the sparse feature selection problem and is used when the number of features is too large to manually search model space. In this setting, we desire a rule for stopping stepwise regression using hypothesis tests while controlling a notion of false rejections. That being said, forward stepwise regression is commonly considered to be “data dredging" and not statistically sound. As the hypotheses tested by forward stepwise are determined by looking at the data, the resulting classical hypothesis tests are not valid. We present a simple solution which leverages classical multiple comparison methods in order to test the stepwise hypotheses using the max-t test proposal. The resulting procedures are fast enough to be used in high-dimensional settings and can be tailored to control the family-wise error rate or FDR. Other procedures estimate new, computationally difficult p-values and have significant lower power. We provide both step-up and step-down variants of our procedure. Furthermore, our proofs readily extend to more general correlation learning methods such as sure independent screening.

EO202 Room Aula B ANALYSIS OF LARGE DATA SETS: THEORY AND APPLICATIONS Chair: Malgorzata Bogdan

E0969: **Analysis of Langevin Monte Carlo via convex optimization**
Presenter: **Blazej Miasojedow**, University of Warsaw, Poland
Co-authors: Alain Durmus, Szymon Majewski
New insights are provided on the unadjusted Langevin algorithm. We show that this method can be formulated as a first order optimization algorithm of an objective functional defined on the Wasserstein space of order 2. Using this interpretation and techniques borrowed from convex optimization, we give a non-asymptotic analysis of this method to sample from logconcave smooth target distribution. Our proofs are then easily extended to the stochastic gradient Langevin dynamics, which is a popular extension of the unadjusted Langevin algorithm. Finally, this interpretation leads to a new methodology to sample from a non-smooth target distribution, for which a similar study is done.

E1507: **Machine learning methods for initial orthonormal basis selection for functional data**
Presenter: **Heddy Bellout**, Lund University, Sweden
Co-authors: Krzysztof Podgorski
In most current implementations of the functional data methods, the effect of the initial choice of an orthonormal basis that is used to analyze data is typically has not been studied. As a result, trigonometric (Fourier), wavelet, or polynomial bases are most popularly used by default. No formal criteria are developed to give a researcher indication which of the bases is preferable for initial transformation of the data. On the other hand it is well known that the choice of the basis affects efficiency in retrieving stochastic structure of a studied model. A classical result in this context is the Karhunen-Loeve expansion of the covariance. The basis associated with this expansion has the optimality in the total mean square error sense. We will propose quantitative criteria in terms of the computational efficiency and the mean square error that will allow for comparison performances of different bases in a given problem. The convenience of a priori chosen orthonormal basis is mostly mathematical, however typically for a given functional data set it maybe computationally more effective to work with a data driven basis. We plan to implement machine learning algorithms for the choice of basis uniformly for all samples and study its efficiency against arbitrary choice of the basis. The optimality criterion, like the total mean square error, discussed above would be utilized, both in the learning algorithms, and in comparison studies.

E1648: **Online adaptive and anytime Mondrian forests**
Presenter: **Stephane Gaiffas**, Universite Paris-Diderot, France
Co-authors: Jaouad Mourtada, Erwan Scornet
Random Forests (RF) is one of the algorithms of choice in many supervised learning applications, be it classification or regression. The appeal of such methods comes from a combination of several features: a remarkable accuracy in a variety of tasks, the small number of parameters to tune, the ability to handle both numerical and categorical outputs, their reasonable computational cost at training and prediction time, and their suitability in high-dimensional settings. The most commonly used RF variants however are offline algorithms, which require the availability of the whole dataset at once. We introduce an online anytime random forest algorithm based on Mondrian forests. Using a suitable adaptation of the context tree weighting algorithm, we show that it is possible to efficiently perform an exact aggregation over all labelled prunings of the trees; in particular, this enables to obtain a truly online parameter-free algorithm, which is adaptive to the unknown regularity of the regression function. Numerical experiments show that our algorithm is competitive, compared to Breimans original random forests.

EO294 Room Aula Magna ADVANCE IN STATISTICAL METHODS FOR BIG AND COMPLEX DATA Chair: Linbo Wang

E0196: **Integrative multi-view reduced-rank regression: Bridging group sparsity and low-rank models**
Presenter: **Gen Li**, Columbia University, United States
Co-authors: Kun Chen

Multi-view data have been routinely collected in various fields of science and engineering. A general problem is to study the predictive association between multivariate responses and multi-view predictor sets, all of which can be of high dimensionality. It is likely that only a few views are relevant to prediction, and the predictors within each relevant view contribute to the prediction collectively rather than sparsely. We cast this new problem under the familiar multivariate regression framework and propose an integrative reduced-rank regression (iRRR), where each view has its own low-rank coefficient matrix. As such, latent features are extracted from each view in a supervised fashion. For model estimation, we develop a convex composite nuclear norm penalization approach, which admits an efficient algorithm via alternating direction method of multipliers. Extensions to non-Gaussian and incomplete data are discussed. Theoretically, we derive non-asymptotic oracle bounds of iRRR under a restricted eigenvalue condition. Our results recover oracle bounds of several special cases of iRRR including lasso, group lasso and nuclear norm penalized regression. Therefore, iRRR seamlessly bridges group-sparse and low-rank methods and can achieve substantially faster convergence rate under realistic settings of multi-view learning. Simulation studies and an application in the longitudinal studies of aging further showcase the efficacy of the proposed methods.

E0451: **Dynamic tracking and screening in massive data streams**
Presenter: **Lilun Du**, HKUST, China

The aim is to construct a large-scale dynamic tracking and screening (DTS) procedure capable of rapidly identifying irregular individual streams whose behavioral patterns deviate from that of the majority. By fully exploiting the sequential feature of datastreams, we first develop a robust estimation approach under a framework of varying coefficient model. The procedure naturally accommodates unequally-spaced design points and updates estimates as new data arrive without the need to store an ever-increasing data history. A data-driven choice of an optimal smoothing parameter is accordingly proposed. Then, we suggest a new model-specification test tailored to the streaming environment. The resulting DTS scheme is able to adapt time-varying structures appropriately, track changes in the underlying models, and hence maintain high identification accuracy in detecting irregular individuals. Moreover, we derive the asymptotic properties of the procedure and investigate its finite sample performance by means of a simulation study and a real data example.

E0958: **Sampling latent states for high-dimensional non-linear state space models with the embedded HMM method**
Presenter: **Alexander Shestopaloff**, The Alan Turing Institute, United Kingdom

A new scheme is proposed for selecting pool states for the embedded Hidden Markov Model (HMM) Markov Chain Monte Carlo (MCMC) method. This new scheme allows the embedded HMM method to be used for efficient sampling in state space models where the state can be high-dimensional. Previously, embedded HMM methods were only applicable to low-dimensional state-space models. We demonstrate that using our proposed pool state selection scheme, an embedded HMM sampler can have similar performance to a well-tuned sampler that uses a combination of Particle Gibbs with Backward Sampling (PGBS) and Metropolis updates. The scaling to higher dimensions is made possible by selecting pool states locally near the current value of the state sequence. The proposed pool state selection scheme also allows each iteration of the embedded HMM sampler to take time linear in the number of the pool states, as opposed to quadratic as in the original embedded HMM sampler.

EO032 Room A1 RECENT DEVELOPMENT IN STATISTICAL ANALYSIS OF BRAIN DATA Chair: Guofen Yan

E0479: **Inference for first passage times of the Feller process**
Presenter: **Satish Iyengar**, University of Pittsburgh, United States

The Feller diffusion process has linear drift and a state dependent diffusion coefficient that vanishes at zero. Earlier studies have shown that it provides a better fit for neural activity than the Ornstein-Uhlenbeck under certain conditions. We describe inference based on maximum likelihood for this model when the available data are spike trains rather than the neurons subthreshold voltage traces.

E0878: **Lagged hierarchical semiparametric models for task-based dynamic functional connectivity (dFC) estimation**
Presenter: **Jaroslaw Harezlak**, Indiana University School of Public Health-Bloomington, United States
Co-authors: Zikai Lin, Maria Kudela, Brandon Oberlin, Joaquin Goni, David Kareken, Mario Dzemidzic

Functional Magnetic Resonance Imaging (fMRI) studies are utilized to assess both brain activation and co-activation among brain regions. Data produced in an MRI scanner consist of hundreds of thousands of time series indicating changes in the blood oxygenation level. We developed a method to estimate the co-activation of hundreds of brain regions at the task-, subject- and population-level at both concurrent time points and at the lagged time intervals. Our method utilizes dynamic functional connectivity approximation, time series bootstrap-based uncertainty evaluation and semiparametric mixed model estimation. We assess our methodology via its application to the study of social and heavy alcohol drinkers reaction to different gustatory cues, including beer, Gatorade and water.

E0937: **Mixed-effect time-varying stochastic blockmodel and application in brain connectivity analysis**
Presenter: **Lexin Li**, University of California Berkeley, United States

Time-varying networks are fast emerging in a wide range of scientific and business disciplines. Most existing dynamic network models are limited to a single subject and discrete-time setting. We propose a mixed-effect multi-subject continuous-time stochastic blockmodel that characterizes the time-varying behavior of the network at the population level, meanwhile taking into account individual subject variability. We develop a multi-step optimization procedure for a constrained stochastic blockmodel estimation, and derive the asymptotic property of the estimator. We demonstrate the effectiveness of our method through both simulations and an application to a study of brain development in youth.

EO594 Room Aula A ADVANCES IN ANALYSIS OF COMPLEX TIME SERIES DATA Chair: Seyed Yaser Samadi

E0587: **Simultaneous inference for curve estimation in time-varying models**
Presenter: **Sayar Karmakar**, University of Florida, United States
Co-authors: Stefan Richter, Wei Biao Wu

A general class of time-varying regression models which cover general linear models as well as time series models is considered. We estimate the regression coefficients by using local linear M-estimation. For these estimators, Bahadur representations are obtained and are used to construct simultaneous confidence bands. For practical implementation, we propose a bootstrap based method to circumvent the slow logarithmic convergence of the theoretical simultaneous bands. The results substantially generalize and unify the treatments for several time-varying regression and auto-regression models. The performance for ARCH and GARCH models is studied in simulations and a few real-life applications of our study are presented through analysis of some popular financial datasets.

E1134: **Learning stochastic dynamical systems via bridge sampling**
Presenter: **Harish Bhat**, University of Utah, United States

Algorithms are developed to automate discovery of stochastic dynamical system models from noisy, vector-valued time series. By 'discovery, we mean learning both a nonlinear drift vector field and a diagonal diffusion matrix for a d-dimensional Ito stochastic differential equation. We parameterize the vector field using tensor products of Hermite polynomials, enabling the model to capture highly nonlinear and/or coupled dynamics. We solve the resulting estimation problem using expectation maximization (EM). This involves two steps. We augment the data via diffusion bridge sampling, with the goal of producing time series observed at a higher frequency than the original data. With this augmented data, the resulting expected log likelihood maximization problem reduces to a least squares problem. Through experiments on systems with dimensions one through eight, we show that this EM approach enables accurate estimation for multiple time series with possibly irregular observation times. We study how the EM method performs as a function of the noise level in the data, the volume of data, and the amount of data augmentation performed.

E1528: **Time series analysis for symbolic interval-valued data**
Presenter: **Seyed Yaser Samadi**, Southern Illinois University Carbondale, United States

While many series record a single value for each time point, many other series record the observations as intervals. This is particularly so with financial data, where, e.g., assets have two prices (bid and ask prices) and the interval between them represents all possible prices at which the asset can be traded. There are countless examples. Therefore, in comparison with standard classical data, they are more complex and can have structures (especially internal structures) that impose complications that are not evident in classical data. As a result of dependency in time series observations, it is difficult to deal with symbolic interval-valued time series data and take into account their complex structure and internal variability. In the literature, the proposed procedures for analyzing interval-valued time series data used either midpoint or radius that are inappropriate surrogates for symbolic interval variables. All previously available methods in the literature fail in some way to use all the variations inherent in the interval-valued data; there is a loss of information. We develop a methodology using the information contained in the complete intervals (and not just on the two point values represented by the end points and/or the center-range values) to analyze interval time series data.

EO370 Room C1 MULTIPLE TESTING Chair: Sebastian Doehler

E0799: **Comparing several adaptive multiple testing methods for discrete uniform homogeneous p-values**
Presenter: **Marta Cousido Rocha**, University of Vigo, Spain
Co-authors: Jacobo de Una-Alvarez, Sebastian Doehler

Large-scale discrete uniform homogeneous p-values arise in many applications, for example, in genome wide association studies. Several multiple testing procedures for such p-values are compared through simulations. Specifically, we consider the q-value approach based on several estimators for the proportion of true null hypotheses π_0: the usual estimator for continuous and possibly heterogeneous p-values, and two estimators proposed recently for discrete p-values. One of these two proposals is focused on discrete uniform homogeneous p-values while we adapt the other one, originally valid for discrete heterogeneous p-values to uniform p-values. The simulated scenario is that of the two-sample problem with low sample size, along a large number of locations or genes. The considered test statistics are the standard student's t test, a permutation test based on the absolute deviation between the sample means, the Kolmogorov-Smirnov two-sample test, and a permutation test based on the L_2 distance between the empirical characteristic functions pertaining to the two samples. The main conclusion is that the specific estimator for π_0 influences the power a lot, and that the approaches for discrete p-values may or may not improve the q-value procedure based on the continuous estimator of π_0.

E0900: **DiscreteFDR: An R-package for controlling the false discovery rate for discrete tests**
Presenter: **Sebastian Doehler**, Darmstadt University of Applied Science, Germany
Co-authors: Etienne Roquain, Guillermo Durand, Florian Junge

The Benjamini-Hochberg procedure and related methods are classical methods for controlling the false discovery rate for multiple testing problems. These procedures were originally designed for continuous test statistics. However, in many applications, the test statistics are discretely distributed. While it is well known that e.g. the Benjamini-Hochberg procedure still controls the false discovery rate in the discrete paradigm, it may be unnecessarily conservative. Thus, there is interest in developing more powerful FDR procedures for discrete data. We present improved procedures that incorporate the discreteness of the p-value distributions and introduce an R package which implements these approaches.

E0939: **Optimal exact tests for multiple binary endpoints**
Presenter: **Robin Ristl**, Medical University of Vienna, Austria
Co-authors: Dong Xi, Ekkehard Glimm, Martin Posch

In confirmatory clinical trials with small sample sizes, hypothesis tests based on asymptotic distributions are often not valid and exact non-parametric procedures are applied instead. However, the latter are based on discrete test statistics and can become very conservative, even more so, if adjustments for multiple testing as the Bonferroni correction are applied. Improved exact multiple testing procedures are proposed for the setting where two parallel groups are compared in multiple binary endpoints. Based on the joint conditional distribution of test statistics of Fisher's exact tests, optimal rejection regions for intersection hypothesis tests are constructed utilizing different objective functions. Depending on the optimization objective, the optimal test yields maximal power under a specific alternative, maximal exhaustion of the nominal type I error rate, or the largest possible rejection region controlling the type I error rate. To efficiently search the large space of possible rejection regions, an optimization algorithm based on constrained optimization and an alternative greedy algorithm are proposed. Applying the closed testing principle, optimized multiple testing procedures with strong familywise error rate control are constructed. The proposed methods are implemented in the R package multfisher.

EO472 Room D1 DIMENSION REDUCTION UNDER HIGH DIMENSION Chair: Zhigen Zhao

E0750: **Model-free variable selection and screening with matrix-valued predictors**
Presenter: **Yuexiao Dong**, Temple University, United States
Co-authors: Zeda Li

A novel framework is introduced for model-free variable selection with matrix-valued predictors. To test the importance of rows, columns, and submatrices of the predictor matrix in terms of predicting the response, three types of hypotheses are formulated under a unified framework. In the fixed-dimensional setting, an asymptotic test as well as a permutation test are proposed to approximate the distribution of the test statistics under the null hypotheses. In the high-dimensional setting, the proposed test statistics can be used for marginal screening. The effectiveness of the proposed methods are evaluated through extensive numerical studies and an application to the electroencephalography (EEG) dataset.

E1670: **Dimension reduction for functional data based on weak conditional moments**
Presenter: **Bing Li**, The Pennsylvania State University, United States
Co-authors: Jun Song

The aim is to develop a general theory and estimation methods for functional linear sufficient dimension reduction, where both the predictor and the

response can be random functions, or even vectors of functions. Unlike the existing dimension reduction methods, this approach does not rely on the estimation of conditional mean and conditional variance. Instead, it is based on a new statistical construction — the weak conditional expectation, which is based on Carleman operators and their inducing functions. Weak conditional expectation is a generalization of conditional expectation. Its key advantage is to replace the projection on to an L_2-space — which defines conditional expectation — by projection on to an arbitrary Hilbert space, while still maintaining the unbiasedness of the related dimension reduction methods. This flexibility is particularly important for functional data, because attempting to estimate a full-fledged conditional mean or conditional variance by slicing or smoothing over the space of vector-valued functions may be inefficient due to the curse of dimensionality. We evaluated the performances of the new methods by simulation and in several applied settings.

E1625: **Ignoring the differences in model properties of sparse PCA and standard PCA can be dangerous and misguide practice**
Presenter: **Soogeun Park**, Tilburg University, Netherlands
Co-authors: Katrijn Van Deun, Eva Ceulemans

Principal component analysis (PCA) is a widely used data reduction technique which finds a weights matrix that orthogonally transforms variables into components with maximal variance. PCA has the special property that this weights matrix is equivalent to a loadings matrix which represents variable-component correlation. Furthermore, PCA is intrinsically linked to the eigenvalue decomposition with loadings and weights being equal to the eigenvectors of the correlation matrix. Sparse PCA, devised to improve interpretability of PCA, introduces sparsity to either the weights or loadings matrix, at the cost of this property: weights and loadings are no longer equivalent in sparse PCA. However, most researchers appear to have maintained an inattentive conception that sparse PCA has equivalent modeling characteristics as PCA. This had led to misguided practices in research such as generating data from simplistic PCA models comprised of sparse eigenvectors for simulation studies and naive use of PCA-based initial values. These mistakes are brought to light and suggestions are made to fix them. The aim is to contribute to shifting the research towards the necessary attention on the statistical models of sparse PCA.

EO174 Room F1 CHANGE POINTS ANALYSIS AND STATISTICAL INFERENCE FOR HIGH DIMENSIONAL DATA Chair: Ping-Shou Zhong

E1114: **Multiple changepoint estimation in high dimensional Gaussian graphical models**
Presenter: **Alex Gibberd**, Lancaster University, United Kingdom
Co-authors: Sandipan Roy

Many modern datasets exhibit a multivariate dependence structure that can be modelled using networks or graphs. For instance, in financial applications, one may study Markowitz minimum-variance portfolios based on sparse inverse covariance matrices. However, in reality, we expect that the underlying volatility and dependency structure between data-streams may change over time, we thus require a way of tracking these dynamic network structures. We will discuss consistency properties for a regularised M-estimator which simultaneously identifies both change points and graphical dependency structure in multivariate time-series. Specifically, we will study the Group-Fused Graphical Lasso (GFGL), which penalises partial-correlations with an l_1 penalty, while simultaneously inducing block-wise smoothness over time to detect multiple change points. Under mild conditions we present a proof of change-point consistency for this estimator. In particular, it is demonstrated that both the changepoint and graphical structure of the process can be consistently recovered, for which finite sample bounds are provided.

E1137: **Asymptotically independent U-statistics for high dimensional adaptive testing**
Presenter: **Gongjun Xu**, University of Michigan, United States

Many high dimensional hypothesis tests examine the moments of the distributions that are of interest, such as testing of mean vectors and covariance matrices. We propose a general framework that constructs a family of U statistics as unbiased estimators of those moments. The usage of the framework is illustrated by testing off-diagonal elements of a covariance matrix. We show that under null hypothesis, the U statistics of different finite orders are asymptotically independent and normally distributed. Moreover, they are also asymptotically independent with the max-type test statistic. Based on the asymptotic independence property, we construct an adaptive testing procedure that maintains high power across a wide range of alternatives. Simulation and real data are further used to validate the proposed method.

E1567: **Computationally efficient detection of subset multivariate changepoints**
Presenter: **Sean Ryan**, Lancaster University, United Kingdom
Co-authors: Rebecca Killick

Due to the growing number of high dimensional datasets there is an increasing need for methods that can detect changepoints in multivariate time series. We focus on the problem of detecting changepoints where only a subset of the variables under observation undergo a change, so called subset multivariate changepoints. One approach to locating changepoints is to choose the segmentation that minimises a penalised cost function via a dynamic program. While this is possible in our setting, the computational complexity of the algorithm means it is infeasible even for small datasets. We propose a computationally efficient approximate dynamic program, SPOT. We demonstrate that SPOT always recovers a better segmentation, in terms of penalised cost, then other approaches which assume every variable changes. Furthermore, under mild assumptions the computational cost of SPOT is linear in the number of data points. As a result, we are able to apply our method to datasetst with millions of data points and thousands of series. In simulation studies we demonstrate that SPOT provides a good approximation to exact methods. We also demonstrate that our method compares favourably with other commonly used multivariate changepoint methods and achieves a substantial improvement in performance when compared with fully multivariate methods.

EO572 Room I1 COMPOSITE LIKELIHOOD ESTIMATION AND APPLICATIONS Chair: Davide Ferrari

E0266: **Quasi-ML estimation, marginal effects and asymptotics for spatial autoregressive nonlinear models**
Presenter: **Anna Gloria Bille**, Free University of Bozen, Italy
Co-authors: Samantha Leorato

The aim is to propose a pairwise-MLE for a general spatial nonlinear probit model, i.e. SARAR(1,1)-probit, defined through a SARAR(1,1) latent linear model. This model encompasses the SAE(1)-probit model and the more interesting SAR(1)-probit model. We perform a complete asymptotic analysis, and account for the possible finite sum approximation of the covariance matrix (Quasi-MLE) to speed the computation. Moreover, we address the issue of the choice of the groups (couples, in our case) by proposing an algorithm based on a minimum KL-divergence problem. Finally, we provide appropriate definitions of marginal effects for this setting. Finite sample properties of the estimator are studied through a simulation exercise and a real data application. In our simulations, we also consider both sparse and dense matrices for the specification of the true spatial models, and cases of model misspecification due to different assumed weighting matrices.

E0526: **Sparse and robust composite likelihood inference with application to parcel-based evoked brain activity analysis**
Presenter: **Zhendong Huang**, The University of Melbourne, Australia
Co-authors: Davide Ferrari

Analysing Blood-oxygen-level dependent (BOLD) signal in multiple active regions of brain is a popular and challenging problem in biological study. In an experiment, the brain region of interest is divided into voxels with BOLD signal observed in each voxel through time. Classical

methods for estimating time series suffer from loss of efficiency due to the high-dimension of the problem and the absence of knowledge on the correlation structure between voxels. An improved composite likelihood method will be introduced to give inference on BOLD signal. The new method seeks sparse composition rule to include only a small proportion of voxels, while obtaining efficiency to the largest extent in the final estimation. Performance of the new method will be illustrated through theoretical results, numerical studies and an application to real BOLD signal data.

E1025: **Lorelogram models for spatially clustered binary data**
Presenter: **Manuela Cattelan**, University of Padova, Italy
Co-authors: Cristiano Varin
Clustered data are often analysed under the assumption that observations from distinct clusters are independent. The assumption may not be correct when the clusters are associated with different locations within a study region, as, for example, in epidemiological studies involving subjects nested within larger units such as hospitals, districts or villages. In such cases, correct inferential conclusions critically depend on the amount of spatial dependence between locations. A modification of the method of generalized estimating equations is discussed to detect and account for spatial dependence between clusters in logistic regression for binary data. The approach proposed is based on parametric modelling of the lorelogram as a function of the distance between clusters. Model parameters are estimated by a two-step approach that combines optimal estimating equations for the regression parameters and pairwise likelihood for the lorelogram parameters. The methodology is illustrated with an analysis of a data set on the prevalence of malaria in children in the Gambia that was described previously.

EO024 Room L1 MEAN SHIFT AND LOCALIZATION TECHNIQUES Chair: Jochen Einbeck

E0472: **Hunting geometric features in the probability density function with direct density-derivative-ratio estimation**
Presenter: **Hiroaki Sasaki**, Nara Institute of Science and Technology, Japan
Geometric features in the probability density function underlying data is useful in statistical data analysis. For instance, the modes (i.e., local maxima) can be used for clustering, and the ridges unveil manifold structures hidden in data. A technical challenge to capture these geometric features is to estimate the ratio of the density derivatives to its density. A native approach to estimate the "density-derivative-ratios" is to first estimate the data density, then compute the derivatives of the estimated density, and finally take their ratios. However, this approach can be unreliable because a good density estimator does not necessarily mean a good density-derivative estimator. In addition, the division by the estimated density could magnify the estimation errors. To cope with this problem, a new estimator, which directly approximates the density-derivative-ratios without going through density estimation, is proposed. Then, with the developed estimator, we propose novel methods for mode-seeking clustering and density ridge estimation. The proposed methods are theoretically analysed. Finally, we numerically demonstrate that the proposed methods outperform existing methods especially for high-dimensional data.

E0645: **Learning to mean-shift in $O(1)$ for Bayesian image restoration**
Presenter: **Siavash Bigdeli**, EPFL, Switzerland
Finding strong oracle priors is an important topic for solving ill-posed problems. We show how denoising autoencoders (DAEs) learn to mean-shift in $O(1)$, and how we leverage this to employ DAEs as generic priors for the task of image restoration. We also discuss the case of Gaussian DAEs in a Bayesian framework, where the degradation parameters (e.g. noise and/or blur kernel) are unknown. Experimental results demonstrate state-of-the-art performance of the proposed DAE priors in image deblurring and super-resolution.

E1344: **The K-modes and Laplacian K-modes algorithms for clustering**
Presenter: **Miguel Carreira-Perpinan**, University of California, Merced, United States
Many clustering algorithms exist that estimate a cluster centroid, such as K-means, K-medoids or mean-shift, but no algorithm seems to exist that clusters data by returning exactly K meaningful modes. We propose a natural definition of a K-modes objective function by combining two powerful ideas in clustering: the explicit use of assignment variables (as in K-means), and the estimation of cluster centroids which are modes of each cluster's density estimate (as in mean-shift). The algorithm becomes K-means and K-medoids in the limit of very large and very small scales. Computationally, it is slightly slower than K-means but much faster than mean-shift or K-medoids. Unlike K-means, it is able to find centroids that are valid patterns, truly representative of a cluster, even with nonconvex clusters. Then, we extend this definition to the Laplacian K-modes objective function by regularizing K-modes with the graph Laplacian, which encourages similar assignments for nearby points (as in spectral clustering). The optimization alternates a convex assignment step and a mean-shift step. This finds meaningful estimates of the density for each cluster, even with challenging problems where the clusters have manifold structure, are highly nonconvex or in high dimension, as with images or text. It also provides an out-of-sample mapping that predicts soft assignments for a new point, in effect a nonparametric model for cluster posterior probabilities.

EO194 Room M1 HETEROGENEITY IN FUNCTIONAL DATA Chair: Pedro Galeano

E0832: **Recursive maxima hunting: Variable selection in FDA**
Presenter: **Jose Luis Torrecilla**, Universidad Autonoma de Madrid, Spain
In a world of big and complex data, the use of methodologies for dimensionality reduction is a commonplace. In this context, variable selection techniques have been proved to be very useful alternatives, since they provide interpretable reductions with important predictive power. We study variable selection for supervised classification, and we are interested in the case of having data that are functions. In this setting, one of the alternatives is the maxima hunting method (MH) which performs variable selection by identifying the maxima of a dependence function between the predictive functional variable and the class label. MH presents a good performance and some valuable properties, however, the relevance of a variable is assessed individually and it has some estimation issues. We present a recursive extension of MH which solves these limitations by subtracting the expectation of the process conditioned to the already selected variables. The new methodology entails some interesting properties and the improvement is illustrated and assessed with simulations and real examples.

E1000: **Functional variables selection in hyperspectral image classification**
Presenter: **Manuel Oviedo de la Fuente**, University of Santiago de Compostela, Spain
Co-authors: Manuel Febrero-Bande, Wenceslao Gonzalez-Manteiga
Different models classification algorithms are reviewed for the prediction of the future class of pixel in a hyperspectral image that have in common that make use of Functional Data Analysis (FDA). The advantage of FDA over classical model is that it is able to exploit this continuous nature of the information of spectral curves in a better way. We used a multiclass one-versus-one (majority voting) and one-versus-rest functional GAM model. In addition, functional non-parametric classification by mean of proximity measures (kNN and kernel classifiers) and by means of depth measures (depth-based classification) can also help to discriminate the pixel class through the shape of the spectral curve. The second part of the communication is devoted to the problem of variable selection. A selection method that is designed to mixed covariates of different nature: scalar, multivariate, functional, etc, is proposed. The proposal begins with a simple null model and sequentially selects a new variable (using distance

correlation) to be incorporated into the final prediction model. The algorithm has shown quite promising results in the regression framework and its extension to the classification problem is attempted.

E0757: **Estimation, imputation and prediction for the functional linear model with scalar response with missing responses**
Presenter: **Pedro Galeano**, Universidad Carlos III de Madrid, Spain
Co-authors: Manuel Febrero-Bande, Wenceslao Gonzalez-Manteiga

Two different methods for estimation, imputation and prediction for the functional linear model with scalar response when some of the responses are missing at random (MAR) are developed. On the one hand, the simplified method consists in estimating the model parameters using only the pairs of predictors and responses observed completely. On the other hand, the imputed method consists in estimating the model parameters using both the pairs of predictors and responses observed completely and the pairs of predictors and responses imputed with the parameters estimated with the simplified method. The two methodologies are compared in an extensive simulation study and the analysis of two real data examples. The comparison provides evidence that the imputed method might have better performance than the simplified method if the numbers of functional principal components used in the former strategy are selected appropriately.

EO318 Room O1 NEW DEVELOPMENTS IN VINE COPULAS AND THEIR APPLICATIONS Chair: Claudia Czado

E0300: **Probabilistic temperature forecasting using d-vine copula regression**
Presenter: **Annette Moeller**, Clausthal University of Technology, Germany
Co-authors: Claudia Czado, Daniel Kraus, Ludovica Spazzini

To account for forecast uncertainty in numerical weather prediction (NWP) models it has become common practice to employ ensemble prediction systems generating probabilistic forecast ensembles by multiple runs of the NWP model, each time with variations in the details of the numerical model and/or initial and boundary conditions. However, forecast ensembles typically exhibit biases and dispersion errors as they are not able to fully represent uncertainty in NWP models. Therefore, it is common practise to employ statistical postprocessing models to correct ensembles for biases and dispersion errors in conjunction with recently observed forecast errors. We propose a novel postprocessing approach for temperature forecasts based on d-vine copula quantile regression. The d-vine copula regression model is a multivariate regression approach predicting quantiles of a response (temperature observations) based on a set of predictor variables (ensemble forecasts). It exploits the dependence of observation and predictors, accounting for non-Gaussian dependencies in a flexible way. In a comparative study with temperature forecasts of different forecast horizons from the European Center for Medium Range Weather Forecast (ECMWF) the d-vine postprocessing approach shows to be highly competitive to the state-of-the-art EMOS model, clearly improving over standard EMOS for larger forecast horizons.

E0422: **Estimating dependence patterns in right-censored event time data using R-vine copula models**
Presenter: **Nicole Barthel**, Technische Universitaet Muenchen, Germany
Co-authors: Paul Janssen, Candida Geerdens, Claudia Czado

In many studies interest is in the time to a predefined event. Due to limited follow-up, instead of the true event times lower right-censoring times might be recorded for some sample units. The resulting lack of information has to be carefully taken into account by inference tools applied to right-censored data in order to arrive at a sound statistical analysis. If for the sample units multiple event times can be observed, the data might further exhibit complex association patterns, which claim elaborate dependence models. For this purpose, the flexible class of R-vine copulas was extended to right-censored event time data. We illustrate novel vine copula based estimation methods through several right-censored data examples: e.g. dependence between times until infection of the four udder quarters of cows is investigated. All four observation units are subject to right-censoring. To analyze data on recurrent asthma attacks in children, the subclass of D-vine copulas is used to capture the inherent temporal dependence. Additional challenges are unbalancedness of the data and dependent right-censoring induced by the serial data nature. Further, an outlook on ongoing research including R-vine copula based quantile prediction and quantile regression for right-censored data is given.

E0822: **Two-part D-vine copula models for insurance claim data**
Presenter: **Lu Yang**, University of Amsterdam, Netherlands
Co-authors: Claudia Czado

Insurance claim data usually follow a two-part mixed distribution: a point mass at zero corresponding to no claim and an otherwise positive claim from a skewed and long-tailed distribution. In addition, insurance companies usually keep track of policyholders' claim over time, resulting in longitudinal data. We study the longitudinal mixed claim data using a two-part D-vine copula model. We build two D-vine copulas, one is used to study the dependence of whether or not a claim is occurred over time, and the other is used to study the dependence in claim size given occurrence. We then use our model to investigate the time dependence of insurance claim using a dataset from the local government property insurance fund in the state of Wisconsin.

EO248 Room O2 CSDA JOURNAL: BIOSTATISTICS Chair: Taesung Park

E0282: **On the sample mean after a group sequential trial**
Presenter: **Ben Berckmoes**, University of Antwerp, Belgium
Co-authors: Anna Ivanova, Geert Molenberghs

A popular setting in medical statistics is a group sequential trial with independent and identically distributed normal outcomes, in which interim analyses of the sum of the outcomes are performed. Based on a prescribed stopping rule, one decides after each interim analysis whether the trial is stopped or continued. Consequently, the actual length of the study is a random variable. It is reported in the literature that the interim analyses may cause bias if one uses the ordinary sample mean to estimate the location parameter. For a generic stopping rule, which contains many classical stopping rules as a special case, explicit formulas for the expected length of the trial, the bias, and the mean squared error (MSE) are provided. It is deduced that, for a fixed number of interim analyses, the bias and the MSE converge to zero if the first interim analysis is performed not too early. In addition, optimal rates for this convergence are provided. Furthermore, under a regularity condition, asymptotic normality in total variation distance for the sample mean is established. A conclusion for naive confidence intervals based on the sample mean is derived. It is also shown how the developed theory naturally fits in the broader framework of likelihood theory in a group sequential trial setting. A simulation study underpins the theoretical findings.

E0813: **Variance estimation for generalised pseudo-values**
Presenter: **Martina Mittlboeck**, Medical University of Vienna, Austria
Co-authors: Harald Heinzl, Ulrike Poetschger

Recently, a novel methodology based on generalised pseudo-values was suggested to compare survival of two cohorts, where cohort membership is a latent baseline variable. Patients in one cohort may undergo an intervention over time, dependent on an exogenous time-consuming search process. A typical example is stem cell transplantation, where identification of a suitable donor from existing databases takes time. Cohort membership becomes known if the search process is ended either successfully or unsuccessfully, yet it remains unknown if donor search is ceased due to patients' death or censoring. The calculation of the generalised pseudo-values for the cohort with time-dependent intervention consists of

two-parts: Firstly, the survival probability $S0(w)$ before the intervention at w, and secondly the survival probability $S1(t*|w)$ from intervention at w until time of interest $t*$. The survival probability $S0(w)$ before the intervention can easily be estimated by Kaplan-Meier. However, variability estimation for the calculation of proper confidence intervals and for testing group differences is not straightforward and often computationally intensive. Different approaches are investigated and compared with respect to coverage of 95 % confidence intervals.

E1708: **Dealing with a small number of large clusters using iterative bootstrap**
Presenter: **Stephane Heritier**, Monash University, Australia
Co-authors: Maria-Pia Victoria-Feser, Stephane Guerrier
Generalized estimating equations is commonly used in cluster randomized trials (CRTs) to account for within-cluster correlation. It is well known that the sandwich variance estimator is biased when the number of clusters is small (< 40), resulting in an inflated type I error rate. The problem is particularly acute with binary outcomes, a common situation in medicine. Various bias correction methods have been proposed in the statistical literature but are bound to fail due to their reliance on asymptotic formulas used beyond their validity domain. This situation is becoming alarming in multi-period CRTs such as stepped-wedge or cluster crossover designs where it is commonplace to have data with 10 to 20 large clusters, sometimes even less. We propose a radically new approach that does not rely on first-order asymptotics. The method starts with a simple estimator of an auxiliary parameter that is then corrected to estimate the main parameter of interest with virtually no bias. Inference is possible through a nearly exact distribution obtained by simulations using the iterative bootstrap. We illustrate the performance of this approach for binary clustered data with $n = 10$ to 20 clusters. The method is general enough to accommodate other models like generalised mixed models or different endpoints.

EO278 Room Q2 BAYESIAN QUANTILE REGRESSION Chair: Carolina Euan

E0719: **Quantile pyramids for quantile regression**
Presenter: **Yanan Fan**, University of New South Wales, Australia
Co-authors: Jean-Luc Dortet-Bernadet, Thais Rodrigues
Quantile regression models provide a wide picture of the conditional distributions of the response variable by capturing the effect of the covariates at different quantile levels. Fitting quantiles at multiple levels simultaneously allow for borrowing of information across the quantile levels, leading to an improvement in efficiency. At the same time, the long standing issue of quantiles crossing can be handled easily under this setup. We consider the use of Bayesian nonparametric prior known as the quantiles pyramid in the quantile regression setting. We show how to flexibly construct a base distribution in the context of quantile regression, and obtain inference at multiple quantile levels simultaneously. We discuss the implementation in both linear and nonlinear quantile regression cases. We discuss strategies for ensuring that the simultaneously fitted quantiles will not cross.

E0914: **Bayesian ensemble of quantile regression trees**
Presenter: **Mauro Bernardi**, University of Padova, Italy
Co-authors: Paola Stolfi
Decision trees and their population counterparts are becoming promising alternatives to classical linear regression techniques because of their superior ability to adapt to situations where the dependence structure between the response and the covariates is highly nonlinear. Despite their popularity, those methods have been developed for classification and regression, while often the conditional mean would not be enough when data strongly deviates from the Gaussian assumption. The proposed approach instead considers an ensemble of nonparametric regression trees to model the conditional quantile at level $\tau \in (0,1)$ of the response variable. Specifically, a flexible additive model is fitted to each partition of the data that corresponds to a given leaf of the tree. Unlike the most popular Bayesian approach (BART) that assumes a sum of regression trees, quantile estimates are obtained by averaging the ensemble trees, thereby reducing their variance. We develop a Bayesian procedure for fitting such models that effectively explores the space of B-spline functions of different orders that features the functional nonlinear relationship with the covariates. The approach is particularly valuable when skewness, fat-tails, outliers, truncated and censored data, and heteroskedasticity, can shadow the nature of the dependence between the variable of interest and the covariates.

E1313: **On consistency and inference for Bayesian quantile regression based on the asymmetric Laplace density**
Presenter: **Karthik Sriram**, Indian Institute of Management India, India
The 'misspecified' asymmetric Laplace density (ALD) is used as a working likelihood for Bayesian quantile regression (BQR). It has been previously shown posterior consistency for the true quantile regression parameters under this misspecification. It was further argued square-root-n consistency. In a recent correction note, it has been pointed out that the argument for square-root-n-rate was incorrect, but could not resolve the issue. We first show that square-root-n consistency can be achieved under some additional regularity conditions. We then discuss its connection to posterior inference with some potential extensions. In particular, we compare two previous works, both of which proposed an approach for posterior inference with BQR based on ALD.

EC635 Room G1 CONTRIBUTIONS IN METHODOLOGICAL STATISTICS AND APPLICATIONS II Chair: Anne Francoise Yao

E1736: **Bayesian new edge prediction and anomaly detection in large computer networks**
Presenter: **Silvia Metelli**, Imperial College London & The Alan Turing Institute, United Kingdom
Co-authors: Nicholas Heard
Statistical anomaly detection searches for outlying behaviour in a network with respect to a putative normal background. In this scenario, it is thus fundamental to build robust models describing the normal network background. This task becomes particularly challenging when considering cyber security applications, which require prompt evaluation on large sets of data. We will introduce a robust Bayesian model and anomaly detection method for simultaneously characterising network structure and modelling likely new edge formation in a large computer network graph. New edges represent connections between a client and server pair not previously observed, and can provide valuable evidence of anomalous activity. What constitutes normal behaviour for some hosts might be very unusual for some others and thus examining existing network structure (e.g. clusters of similar clients and servers) is key for accurately predicting likely future interactions. For this purpose, a notion of similarity between clients and servers is developed, first under hard-thresholding with a clustering model, and then extended to soft-thresholding in a flexible latent feature space. The model is then used to construct an anomaly detection method, which successfully identifies some of the machines known to be compromised when demonstrated on real computer network authentication data.

E1732: **A computational method for estimating the ratio of scale parameters in the two-sample problem**
Presenter: **Mona Alduailij**, Princess Noura Bint Abdulrahman University, Saudi Arabia
In statistical analysis testing, the variation of scale parameters between two samples plays an important role. A computational iterative method is proposed for finding an estimator and a confidence interval of the ratio of the scale parameters for the two-sample problem. A comparison is made between the existing parametric and non-parametric rank tests for the two-sample scale problem, which include linear rank tests and folded rank tests with different score functions, Lehmann test, jackknife test, Sukhatme test, placement tests, permutation tests and the classical Levene tests. A Monte Carlo simulation will be used to show the performance of our algorithm under symmetric and asymmetric distributions for different

sample sizes. The performance of the estimator will be compared with the available parametric method for estimation. Also, the performance of the proposed confidence interval will be analyzed by computing the length of the interval and its probability of coverage.

E1200: **A new modified Liu-type estimator for linear regression models with correlated regressors**
Presenter: **Aslam Muhammad**, Bahauddin Zakariya University, Pakistan

A new biased estimator is presented and its properties for linear regression models when regressors are correlated are discussed. This new estimator is a general class of biased estimators which includes some well-known biased estimators as a special case. Comparison in the sense of mean squared error matrix (MSEM) is made with popular estimators i.e., the ordinary least squares estimator, ordinary ridge regression estimator, Liu estimator (LE) and Liu-Type estimators. It has been shown that our proposed estimator outperforms the other estimators. For the empirical study, the Monte Carlo simulations are made which show some superior performance of the new proposed estimators. An illustrative example has also been provided.

EG006 Room E1 CONTRIBUTIONS IN MIXTURE MODELS Chair: Florence Forbes

E1468: **An EM type algorithm for maximum likelihood estimation of the negative binomial-gamma regression model**
Presenter: **George Tzougas**, London School of Economics and Political Science, United Kingdom

Mixed Poisson regression models have been massively overused for modelling heterogeneous count data in a wide range of areas, such as sociology, biology, biometrics, genetics, medicine, marketing, applied econometrics and insurance. The negative binomial - gamma regression model can be considered as an alternative to mixed Poisson models since it can adequately capture the stylized characteristics of highly dispersed count data. However, due to the complexity of its likelihood, direct maximization is difficult and has not been addressed in the literature so far. The main achievement is that we propose a simple expectation-maximization (EM) type algorithm for maximum likelihood estimation of the model which can overcome the numerical difficulties occurring when standard numerical techniques are used. Moreover, the proposed algorithm has the considerable mathematical flexibility for fitting other mixed negative binomial regression models stemming from several other mixing distributions. Additionally, the by-products of the algorithm can be useful for further inference. For example, the posterior expectations, which are readily available after the convergence of the EM algorithm, can be employed for empirical Bayes estimation and can be used to predict future outcomes. Finally, a real data application using motor insurance data is examined and some operating characteristics of the algorithm are discussed.

E1536: **Multiple-valued symbolic data clustering: A model-based approach**
Presenter: **Jose Dias**, ISCTE - Instituto Universitario de Lisboa, Portugal

Symbolic data analysis (SDA) has been developed as an extension to data analysis that handles more complex data structures. In this general framework, the pair observation/variable is characterized by more than one value: from two (e.g., interval-value data defined by minimum and maximum values) to multiple-valued variables (e.g., frequencies or proportions). Clustering of multiple-valued symbolic data is discussed. We propose a new model-based clustering framework based on Dirichlet distributions that includes mixture of regression/expert models. Results are illustrated with synthetic and demographic (population pyramids) data.

E1442: **A new Dirichlet-multinomial mixture model for count data**
Presenter: **Roberto Ascari**, University of Milano-Bicocca, Italy
Co-authors: Sonia Migliorati, Andrea Ongaro

The Dirichlet-multinomial is one of the most known compound distributions for multivariate count data. Let $\mathbf{X}|\mathbf{p} \sim \text{multinomial}(n, \mathbf{p})$ and $\mathbf{P} \sim \text{Dirichlet}(\alpha)$, then the marginal distribution of $\mathbf{X}$ is the Dirichlet-multinomial distribution. Because of the severe covariance structure imposed by the Dirichlet prior, covariance among distinct elements of $\mathbf{X}$ assumes only negative values and this could be unrealistic in some particular scenarios. In the literature there exist several other distributions defined on the simplex: a recent proposal is the Extended Flexible Dirichlet (EFD), a generalization of the Dirichlet with a less strict dependence structure. A new distribution for count data, called EFD-multinomial, can be obtained by compounding the multinomial model with an EFD prior on the parameters $\mathbf{P}$. Due to the covariance structure of the EFD, it allows for positive dependence for some pairs of count categories. Furthermore, thanks to its finite mixture representation, an EM-based estimation procedure can be derived. Some theoretical properties of the EFD-multinomial distribution are shown, and a preliminary simulation study is performed to evaluate the behavior of the EM-based MLE under several scenarios, including positively correlated counts.

EG263 Room H1 CONTRIBUTIONS IN SURVIVAL ANALYSIS Chair: Takeshi Emura

E1345: **Smooth backfitting of additively structured hazard rates for in-sample forecasting**
Presenter: **Stephan Bischofberger**, Cass Business School, United Kingdom
Co-authors: Jens Perch Nielsen, Munir Hiabu, Enno Mammen

Smooth backfitting has been established in nonparametric regression and in density estimation as a very promising alternative to the classic backfitting method. We apply the concept to a survival model with additively structured nonparametric hazard. The model allows for very general censoring and truncation patterns occurring in many forecasting applications such as medical studies or actuarial reserving. A crucial point is that - in contrast to classical backfitting - we do not assume independence between the covariates. Our estimators are shown to be a projection of the data into the space of multivariate hazard functions with additive components. Hence, our hazard estimator is the closest additive fit even if the actual hazard rate is not additive. Another big advantage of our additive model is that our estimators are straight forward to derive in theory including excellent properties as well as their simple implementation in practice even for high dimensional covariates. We provide full asymptotic theory for our estimators as well as a simulation study.

E1363: **Behaviour of tests in the Cox proportional hazards model**
Presenter: **Aneta Andrasikova**, Palacky University Olomouc, Czech Republic
Co-authors: Eva Fiserova

Survival analysis, or so called time-to-event analysis, is applied in a wide range of research fields, especially in medicine, where it allows evaluation of certain medical procedures. It is based on evaluating the time until the occurrence of the event of interest, e.g. relapse of certain disease or death. Typical features of survival data is censoring, i.e. incomplete information about the time of occurrence of the event. This incompleteness may be caused by the end of the study before occurrence of the event or loss of contact with subject of observation. These observations are called as right-censored. The effect of some specific covariates on survival time is usually analysed by the Cox proportional hazards model and the statistical significance of the effect of covariates is then verified by the likelihood ratio test, the Wald test, or the score test. These tests are asymptotically equivalent; however, they give numerically different results in applications in dependence on available data. The author deals with the behaviour of the power of these tests for small sample sizes under various proportion of right censored data and distributions of baseline hazard functions. The aim is to familiarize the audience with the results of performed simulations.

E1700: **On the analysis of discrete time competing risks data**
Presenter: **Minjung Lee**, Kangwon National University, Korea, South

Regression methodology has been well developed for competing risks data with continuous event times, both for the cause-specific hazard and cumulative incidence functions. However, in many applications, the event times may be observed discretely. Naive application of continuous time regression methods to such data is not appropriate. We propose maximum likelihood inferences for estimation of model parameters for the discrete time cause-specific hazard functions, develop predictions for the associated cumulative incidence functions, and derive consistent variance estimators for the predicted cumulative incidence functions. The methods are readily implemented using standard software for generalized estimating equations, where models for different event types may be fitted separately. Simulation studies demonstrate that the methods perform well in realistic set-ups. The methodology is illustrated with stage III colon cancer data from SEER.

EG027 Room N1 CONTRIBUTIONS IN FUNCTIONAL TIME SERIES ANALYSIS Chair: Gregory Rice

E0411: **Minimum entropy forecast for functional time series**
Presenter: **Nicolas Hernandez**, Universidad Carlos III de Madrid, Spain
Co-authors: Alberto Munoz, Gabriel Martos

Consider a functional time series (FTS) data set $\{X_k\}_{k\in\{1,\cdots,n\}}$, where each X_k is a random function $X_k(t), t \in [a,b]$. We introduce a novel approach to forecast functional time series, based on the truncated multivariate representation of the time series, obtained by projecting them onto an appropriate Reproducing Kernel Hilbert Space (RKHS). We propose a modelization scheme to obtain the h steps ahead prediction, $X_{(k+1)}(t),\cdots,X_{(k+h)}(t)$, from the multivariate RKHS representation of the FTS data set. A bootstrap method for dependent data and a minimum entropy criterion is then applied to obtain the point forecast and the confidence bands. As an interesting application, we conduct an analysis of fertility and mortality rate curves of functional time series forecast.

E1674: **Unit-root test for functional data based on records**
Presenter: **Israel Martinez Hernandez**, KAUST, Saudi Arabia
Co-authors: Marc Genton

A unit-root test for functional time series is proposed which based on records, where a record is a temporary maximum or minimum in the sequence of curves. The problem of unit roots in univariate time series has been deeply studied because of its importance in econometrics. However, economic and financial data can often be considered as a collection of curves over time. We extend the concept of records to functional data by using depth notions for curves, and we propose a new unit root test based on the ranks of the curves. We study the growth rate of the number of records over time and derive the asymptotic distribution of the test statistic. We present a Monte Carlo study to evaluate the performance of the proposed test and compare our test with the existing ones in the literature. We apply our method to a dataset of annual mortality rates in France.

E1589: **A bootstrap-based KPSS test for functional time series**
Presenter: **Yichao Chen**, Nanyang Technological University, Singapore
Co-authors: Chi Seng Pun

The bootstrap method is applied to the KPSS test of functional time series to estimate the limit distribution of the test statistic when the unobserved noises of original sample are independent. We find that bootstrap method makes the testing process faster and more efficient than the methods have found, especially when sample size N are not very big (no more than 70). The convergence of the bootstrap test statistic is established in a general probability space and then use simulation study to present the efficiency of our methods in KPSS test of functional time series.

EG179 Room P1 CONTRIBUTIONS IN MARKOV SWITCHING REGRESSION AND HIDDEN MARKOV MODELS Chair: Fulvia Pennoni

E0996: **Nonparametric estimation in hidden Markov models using the EM algorithm**
Presenter: **Sina Mews**, Bielefeld University, Germany
Co-authors: Roland Langrock, Timo Adam

Hidden Markov models (HMMs) constitute a flexible class of models for time series data, in which the observations are generated by conditional distributions as selected by an underlying Markov chain. While the state-dependent distributions are typically assumed to be a member of some parametric family, misspecifications in this regard can lead to biased parameter estimates, to a high misclassification rate when decoding the hidden states, and to invalid inference on the number of states, to name just a few undesirable consequences. To overcome this restrictive assumption, the state-dependent distributions can be modelled nonparametrically based on penalized splines (P-splines) in order to obtain density estimates sufficiently flexible to capture any distributional shape, with a wiggliness penalty to avoid overfitting. However, parameter estimation based on numerical maximisation of the likelihood requires a computationally intensive determination of the smoothing parameters via grid search. We suggest to instead use the EM algorithm, leading to the main advantage that one can iteratively update the smoothing parameters within each M step. A simulation study as well as a real data example are used to assess the performance of the EM-based nonparametric estimation approach. Its results are compared to the numerical ML equivalent as well as to a parametric model formulation, indicating that the EM-based estimation approach is a suitable alternative to the numerical ML one.

E1349: **Comparing behavioral dynamics between groups using hierarchical hidden semi Markov models**
Presenter: **Emmeke Aarts**, Utrecht University, Netherlands

Technological advances make it increasingly easy to collect intensive longitudinal data on multiple subjects or animals. Hidden Markov models (HMM) are becoming an increasingly popular method to summarize these behavioral data over time. However, when using conventional HMMs on behavioral data, there are two drawbacks. First, HMMs are not well suited to simultaneously model sequences of data of multiple subjects, or to formally compare parameters between groups of subjects. Second, a HMM assumes that the amount of time spent within a hidden state is a function of a memoryless process. However, when investigating behavior over time, this is biologically implausible. We develop and implement a hierarchical hidden semi Markov model (HHSMM) to describe - and formally compare - the temporal organization of behavior over groups. In our model, the state durations are explicitly modeled (i.e., an explicit duration HMM), and a Bayesian framework is used for parameter estimation. We illustrate our proposed model using a real data example, comparing the behavioral pattern of young adult and aged C57BL/6J mice. Our proposed framework is one of the first that models the behavior of multiple animals simultaneously, taking a HSMM approach while allowing for heterogeneity in - and formal group comparisons on - all model parameters.

E1598: **Efficient estimation for non-linear state space models of population survey data**
Presenter: **Takis Besbeas**, Athens University of Economics and Business, Greece

Time series data of population abundances are often described using population dynamics state-space models involving Gompertz, Moran-Ricker or Beverton-Holt latent processes. We show how hidden Markov model methodology provides a flexible framework for fitting a wide range of models to such data. The proposed method avoids any Kalman filter approximations or Monte Carlo simulation that might be employed, and allows model comparison and goodness-of-fit using standard likelihood tools. The method is illustrated using two real data sets of mammal populations from Europe and Australia. There is little difference between the three latent models for the two case studies, which suggests ecological time series

may not be sufficiently informative on latent structure when observation error is unknown in general.

EG421 Room P2 CONTRIBUTIONS IN BAYESIAN MODELLING AND COMPUTATION Chair: Sergio Bacallado

E1417: **Probabilistic Bayesian updating of IOTs**
Presenter: **Vladimir Potashnikov**, RANEPA, Russia
Co-authors: Oleg Lugovoy, Andrey Polbin
Efforts on developing and application of probabilistic method(s) for updating IO tables are summarized. The core of the methodology is the Bayesian framework which combines an information from observed data, additional beliefs (priors), and related uncertainties into the posterior joint distribution of input-output table (IOT) coefficients. The framework can be applied to various IOT problems, including updating, disaggregation, evaluation of uncertainties in the data, and addressing incomplete/missing observations. The flexibility of the methodology is partially based on sampling techniques. We apply modern Markov Chains Monte Carlo (MCMC) methods to explore the posterior distribution of IOT coefficients. Estimating IO tables by Bayes method is a computationally complex problem. The aim is to propose a modification of the algorithm, allowed to partially solve the problem of the curse of dimension. We also compare results with mainstream methods of updating IOT to investigate its performance. Various indicators of performance and application to various data suggest different results. The overall performance of the method is similar or comparable with mainstream techniques. The main advantage of the proposed methodology is an estimation of the full profile of joint probability distribution of unknown IOT matrices. The method can be also combined with any other techniques through prior information.

E1661: **Bayesian probit classification trees**
Presenter: **Paola Stolfi**, CNR - Institute for Applied Mathematics, Italy
Co-authors: Mauro Bernardi, Daniele Durante
Ensemble of decision trees are popular techniques for regression and classification either because of their forecasting performances and their ability to account for complex nonlinear dependence structures among predictors. Leveraging on the Bayesian Additive Regression Trees (BART) approach, we propose new methods to deal with binary classification for CART and BART. Specifically, we introduce a new representation for the probit classification model that avoid the data augmentation scheme previously used. The proposed approach is illustrated and validated through comparison with alternative methods on simulated and real datasets.

E1515: **On the use of scoring rules for Bayesian model selection with improper priors**
Presenter: **Erlis Ruli**, University of Padova, Italy
Co-authors: Laura Ventura, Monica Musio
The Bayes factor (BF) is the standard model selection tool. However, it is well known that BF tends to be sensitive to the prior distributions of the models under comparison and therefore requires careful elicitation. Furthermore, the Bayes factor cannot be used with objective improper priors, because of the dependence of the marginal likelihood on the arbitrary scaling constants of the model prior densities. It has been proposed to solve this problem by replacing marginal log-likelihood by a homogeneous proper scoring rule, which is insensitive to the scaling constants. We apply and study this methodology in the context of continuous exponential family. A couple of examples are provided.

CO454 Room D2 FINANCIAL NETWORKS Chair: Monica Billio

C1038: **Dependence structure in international bond returns**
Presenter: **Domenico Sartore**, Ca Foscari University of Venice, Italy
Co-authors: Andrea Berardi, Monica Billio, Roberto Casarin
Bond return fluctuations across different currency areas are generally highly interrelated. However, both the contemporaneous causal relationships and the temporal dependence structure vary over time. We consider government bond returns from Australia, Canada, Germany, Japan, Switzerland, the UK and the US, and document the time-varying behaviour of the degree of connectedness among the seven currency areas. We find that the dependence structure of bond returns can be significantly different for short and long maturities. The empirical analysis is based on a Bayesian graphical VAR model, where the contemporaneous and temporal causal structures of the structural VAR are represented by two different graphs and an efficient Markov chain Monte Carlo algorithm is used to estimate jointly the two causal structures and the parameters of the reduced-form VAR model.

C1592: **Linkage of contrarian and momentum traders in a stock market: Complex network approach**
Presenter: **Kestutis Baltakys**, Tampere University of Technology, Finland
Co-authors: Juho Kanniainen, Frank Emmert-Streib
While executing individual trading strategies in the stock markets, investors form unobservable indirect relationships in terms of behavioral similarities, which we call investor networks. The structure of these networks is important not only from the perspective of individual members, but also from that of the whole market. By undergoing a complex network analysis of the stock market, we want to shed light on the relation between individual investor behavior patterns to the emergence of collective market equilibrium. We investigate investor collective behavior in stock markets as investors taking different sides in transactions form strategy based trading structures. We will leverage our previously introduced multilayer aggregation framework to determine the investor groups that balance the scales of supply and demand in the stock exchange under different market conditions. We observed the existence of two clearly defined opposing counter-parties that exchange their roles as liquidity takers and providers. These trading structures emerge at the times of crisis, during extraordinary bad and good days, and stabilize the markets. We aim to show the power and usefulness of network methodologies to understand and visualize complex financial concepts.

C1005: **Bayesian Markov switching tensor regression for time-varying networks**
Presenter: **Matteo Iacopini**, Ca Foscari University of Venice, Italy
Co-authors: Monica Billio, Roberto Casarin
A new Bayesian Markov switching regression model is proposed for multi-dimensional arrays (tensors) of binary time series. We assume a zero-inflated logit dynamics with timevarying parameters and apply it to multi-layer temporal networks. The aim is threefold. First, in order to avoid over-fitting we propose a parsimonious parametrization of the model, based on a low-rank decomposition of the tensor of regression coefficients. Second, the parameters of the tensor model are driven by a hidden Markov chain, thus allowing for structural changes. The regimes are identified through prior constraints on the mixing probability of the zero-inflated model. Finally, we model the jointly dynamics of the network and of a set of variables of interest. We follow a Bayesian approach to inference, exploiting the Polya-Gamma data augmentation scheme for logit models in order to provide an efficient Gibbs sampler for posterior approximation. We show the effectiveness of the sampler on simulated datasets of medium-big sizes, finally we apply the methodology to a real dataset of financial networks.

CO544 Room E2 CONTRIBUTIONS IN INTEREST RATES Chair: Luca Benzoni

C0570: **The position of the hump as a predictor of the treasury yield curve: A cointegration approach**
Presenter: **Arturo Leccadito**, Universita della Calabria, Italy

The study proposes using a new factor to describe the temporal relationship between interest rates on Treasury securities. This new factor is described by the position of the hump in the yield curve along the maturity axis and it is affected only when movements in the slope parameter are not matched by moves in the curvature. As it is common in the literature, we find that Treasury rates are non-stationary and form a cointegrated system. Using daily data for the period 2006-2018, we first calculate the new parameter using a non-parametric method and subsequently show that changes in the position of the hump are significant in explaining the short-run dynamics of Treasury rates. Furthermore, we document that adding the variable to short-run dynamics of the vector error correction model (VECM) increases its forecasting ability. Finally, as a robustness check, the study employs data at a different frequency (monthly) in order to make it possible the use of additional macroeconomic variables (like inflation and the Federal Funds Rate) in the short-run dynamics of the VECM. Information criteria confirm the superiority of VECMs that include changes in the position of the hump as an exogenous variable.

C0432: **Forecasting the term structure of interest rates with potentially misspecified models**
Presenter: **Yunjong Eo**, University of Sydney, Australia
Co-authors: Kyu Ho Kang

The predictive gains of a Markov-switching mixture of three individual bond yield predictions is assessed: namely, the dynamic Nelson-Siegel model (DNS), the arbitrage-free Nelson-Siegel model, and the random walk (RW) model as a benchmark. Despite the popularity of these three frameworks, none of them dominates the others across all maturities and forecast horizons. This fact indicates that the models are potentially misspecified. We investigate whether combining the possibly misspecified models in a linear form helps improve predictive accuracy. To do this, we evaluate the out-of-sample forecasts of the mixture models compared to the individual models. Our findings provide strong evidence that model combination can be a better option than selecting one of the alternative models.

C1540: **Forecasting and trading monetary policy effects on the riskless yield curve with regime switching Nelson-Siegel models**
Presenter: **Massimo Guidolin**, Bocconi University, Italy
Co-authors: Manuela Pedio

The aim is to investigate whether accounting for regime shifts enhances the forecasting power of a Dynamic Nelson-Siegel (DNS) model in which exogenous factors that interact with the dynamics of regime shifts are used to capture different aspects of the monetary policy conducted by the Federal Reserve, with special emphasis to the emergency measures adopted in response to the 2008-2009 financial crisis. We also implement a set of systematic trading strategies relying on the forecasts of our models to evaluate the economic values of both a baseline DNS with and without regimes and of an extended econometric framework in which DNS, regimes, and factors capturing the monetary policy stance are interacted. The empirical results suggest that the Markov switching structure generally improves the out-of-sample forecasts of the single-state model, particularly during the crisis. Moreover, we find that the models augmented with a variety of macroeconomic variables significantly improve the predictability of different features of the yield curve. This is especially true during the crisis, coherently with the stated purposes of unconventional monetary policies. Finally, we obtain evidence that switching DNS models augmented to include monetary policy variables is economically beneficial in risk-adjusted terms.

CO356 Room M2 MACROECONOMICS AND FINANCE APPLICATIONS WITH LINEAR AND NONLINEAR FILTERS Chair: Huyen Nguyen

C0471: **State-dependent monetary policy regimes**
Presenter: **Shayan Zakipour-Saber**, Queen Mary University, United Kingdom

The aim is to uncover the determinants of monetary policy regime shifts by estimating a general equilibrium model in which parameters of a Taylor-type policy rule follow a Markov process. Typically, in this class of model commonly referred to as MSDSGE, the variable that determines the current regime in place is assumed to follow an exogenous Markov process. In contrast to the existing literature, we estimate an MSDSGE model allowing endogenous variables such as the first lag of inflation, output and contemporaneous economic shocks to directly influence which regime will be in place. We then apply Bayesian model comparison techniques to test if the data accepts this modification and to determine which variable or combination of variables influences monetary policy regime shifts.

C0613: **Financial factors and the natural rate of interest puzzle**
Presenter: **Josselin Roman**, Paris Dauphine and PSL Universities, France
Co-authors: Gauthier Vermandel

The aim is to develop and estimate a DSGE model for the US economy to study the natural rate of interest and its drivers under financial frictions. We get three main results. First, the analysis shows that permanent shocks, that capture a secular stagnation effect, are not a critical driver of the natural rate. Second, the persistent low level of the natural rate after the financial crisis finds its roots in the very long-lasting nature of financial shocks through a super debt-cycle mechanism. Third, we find that the effectiveness of the unconventional monetary policy in stabilizing the natural rate is conditional on the type of shocks. In particular, a credit policy is effective in offsetting financial and supply shocks.

C1090: **Dynamic risk-taking behavior of mutual funds**
Presenter: **Huyen Nguyen**, Le Mans University, France

The aim is to study the dynamics of risk exposures of different strategies of US domestic equity mutual funds (DEMF) as a group over the business cycle. Following previous works, we apply the Kalman filter to estimate the time-varying exposures of 6 DEMF indexes to various sources of risk over the period 1994-2015. The results show that for all strategies, risk coefficients vary strongly over time, indicating that these DEMF portfolios are actively managed and not simply buy-and-hold ones. In terms of market risk, different DEMF strategies display quite different betas before 2000 but their betas become more or more similar after. Since 2008, all DEMF strategies concomitantly reduce market risk in a significant proportion, certainly due to the fly-to-quality pressure. This finding implies a reduced diversification benefit for mutual funds investors and a serious potential threat of instability for the financial system during market turbulences.

CG071 Room F2 CONTRIBUTIONS IN HIGH-FREQUENCY Chair: Toshiaki Watanabe

C0812: **Using high-frequency exchange rate to identify direct and information effects of monetary policy shocks**
Presenter: **Boreum Kwak**, Martin Luther University Halle-Wittenberg and Halle Institute for Economic Research, Germany
Co-authors: Alexander Kriwoluzky, Oliver Holtemoeller

The impact of central bank announcements on the macroeconomy in the US is studied. Monetary policy announcements contain information about current and future interest rate policies and the economic outlook. We disentangle the surprises caused by direct changes in interest rate and information in policy announcements using changes in volatility of two shocks in high-frequency exchange rate. The information shock in high-frequency surprises becomes quantitatively significant after the recent financial crisis. We investigate the impact of identified direct and information shocks on macro variables using proxy SVAR: a positive information shock is perceived by private agents as a positive signal related to a future economic status, and induces an increase in output and easing financial conditions. Finally, we observe that a monetary policy uncertainty responds to the direct policy shock immediately, but to the information shock slowly with fluctuations.

C1487: **The timing of the flight to gold: An intra-day analysis of gold and the S&P500**
Presenter: **Konstantin Kuck**, University of Hohenheim, Germany
Co-authors: Dirk Baur

Intra-day gold and S&P500 data covering the period from 2007 to 2018 are used to investigate when and how fast gold prices react to extreme negative shocks in the equity market. The empirical analysis reveals three interesting features of gold: First, negative 5-min S&P500 returns smaller than -0.75% are associated with significant increases of the gold price. That is, we observe a fast reaction of the gold price to extreme equity price declines. Second, gold and equity prices do not co-move on days with extreme open-to-close price declines in the stock market. On these days, the gold price continues to increase after the end of stock trading, suggesting spillovers between the stock and gold markets in presence of extreme conditions. Moreover, equity prices tend to decline gradually, implying that there is time to get out of the equity market and time to get into the gold market. In essence, these findings confirm the safe haven property of gold with respect to equity, also at higher than daily data frequencies. Third, these features found for gold spot price returns are also present in gold futures price returns, which indicates that immediate tangibility seems not to be relevant under extreme conditions.

C1603: **Liquidity fluctuations and the latent dynamics of price impact**
Presenter: **Giulia Livieri**, Scuola Normale Superiore, Italy
Co-authors: Fabrizio Lillo , Luca Philippe Mertens, Alberto Ciacci

Market liquidity is a latent and dynamic variable. Building on the stylized limit order book (LOB) model, we propose a dynamical price impact model at high frequency in which price impact is determined by the product of three components: a daily price impact component, a deterministic intraday pattern, and a stochastic autoregressive component. The resulting model has a linear Gaussian state-space representation which can be estimated using a Kalman filter. We provide empirical evidence in support of our model by analyzing six months of order book data for eight liquid stocks traded on the NASDAQ in 2016. We show that the price change conditional on order flow imbalance predicted by our model explains on average 82% of price change variance. This represents a 16% increase with respect to a previous improved model and a 27% increase with respect to the stylized LOB model. Finally, we perform an out-of-sample analysis of real-time estimates of price impact, and show that our model provides a superior out-of-sample forecast of price impact with respect to historical estimates.

Authors Index

