PROGRAMME AND ABSTRACTS

4th International Conference on

Econometrics and Statistics (EcoSta 2021)

http://cmstatistics.org/EcoSta2021

Hosted virtually by The Hong Kong University of Science and Technology
24 – 26 June 2021

ISBN: 978-9925-7812-0-1

Co-chairs:

Tomohiro Ando, Lixing Zhu, Andreas Christmann and Mike K.P. So.

EcoSta Editors:

Ana Colubi, Erricos J. Kontoghiorghes and Manfred Deistler.

Scientific Programme Committee:

Alessandra Amendola, Anastassia Baxevani, Monica Billio, Jeng-Min Chiou, Yulia Gel, Richard Gerlach, Michele Guindani, Wenqing He, Hyokyoung Grace Hong, Raphael Huser, Wen-Han Hwang, Ching-Kang Ing, Ci-Ren Jiang, Sungkyu Jung, Jaeyoung Kim, Tae Yoon Kim, Yang-Jin Kim, Yongdai Kim, Sangyeol Lee, Xin Li, Tsung-I Lin, Koichi Maekawa, Rogemar Mamon, Geoffrey McLachlan, Marc Paolella, Limin Peng, Igor Pruenster, Yumou Qiu, Jeroen Rombouts, Xiaofeng Shao, Mike K.P. So, Zhihua Su, Wenguang Sun, Garth Tarr, Wolfgang Trutschnig, Toshiaki Watanabe, Gongjun Xu, Weixin Yao, Yiming Ying, Peter Zadrozny, Ding-Xuan Zhou and Ji Zhu.

Local Organizing Committee:

The Hong Kong University of Science and Technology Business School, EcoSta, CMStatistics and CFEnetwork.

Dear Colleagues,

It is a great pleasure to welcome you to the 4th International Conference on Econometrics and Statistics (Virtual EcoSta 2021). This year we are passing through extraordinary events that are significantly affecting our personal and professional lives. Until January 2021, we were still hoping to hold part of the conference in person. However, due to the Covid-19 pandemic, we were forced to have the whole conference hosted virtually by The Hong Kong University of Science and Technology (HKUST). Despite the organizational challenges caused by these changes in such a short time, we are happy to welcome warmly the over 450 participants. More than ever, we acknowledge the efforts of all those involved in the conference, especially the session organizers, who had to rebuild their sessions to adapt to the online requirements.

The conference is co-organized by the working group on Computational and Methodological Statistics (CMStatistics), the network of Computational and Financial Econometrics (CFEnetwork), the journal Econometrics and Statistics (EcoSta) and HKUST Business School.

Following the success of the last three editions, the aim is for the conference to become a leading meeting in econometrics, statistics and their applications.

The EcoSta 2021 consists of about 110 sessions, three keynote talks, three invited sessions, and over 425 virtual presentations. These numbers confirm the support of the involved research communities to this important initiative. It is indeed promising that the EcoSta conference will become a successful medium for disseminating high-quality research in Econometrics and Statistics and facilitating networking. In this conference edition, special emphasis will be given to fintech (financial technology).

The Co-chairs acknowledge the collective effort of the scientific program committee, session organizers, and local organizing committee, which has produced a programme that spans all the areas of econometrics and statistics. The local host and assistants have substantially contributed through their effort to the successful organization of the conference. We thank them all for their support.

It is hoped that the quality of the scientific programme will assist in providing the participants with productive and stimulating virtual networking.

The Elsevier journals of Econometrics and Statistics (EcoSta) and Computational Statistics & Data Analysis (CSDA) are associated with CFEnetwork, CMStatistics, and the EcoSta 2021 conference. The participants are encouraged to join the networks and submit their papers to special or regular peer-reviewed issues of EcoSta and the CSDA Annals of Statistical Data Science (SDS).

Finally, we are happy to announce that the 5th International Conference on Econometrics and Statistics (EcoSta 2022) will take place at the Ryukoku University, Kyoto, Japan, from Saturday the 4th to Monday the 6th of June 2022. Tutorials will take place on Friday the 3rd of June 2022. You are invited to participate actively in these events.

Ana Colubi, Erricos J. Kontoghiorghes and Mike K.P. So
on behalf of the Co-Chairs and EcoSta Editors

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 about 1950 members and the following specialized teams

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

You are encouraged to become a member of the WG. For further information, please contact the Chairs of the specialized groups (see the WG's website), or 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 network's activities by organizing sessions, tracks and tutorials during the annual CFEnetwork meetings and 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. Now, the CFEnetwork has over 1100 members.

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

SCHEDULE (Hong Kong time, GMT+8)

2021-06-24	2021-06-25	2021-06-26
A - Opening and Keynote talk EcoSta2021 08:00 - 09:00	**G** EcoSta2021 08:00 - 09:40	**M** EcoSta2021 08:00 - 10:05
B EcoSta2021 09:10 - 10:25	**H** EcoSta2021 09:50 - 11:05	**N** EcoSta2021 10:15 - 11:55
C EcoSta2021 10:35 - 12:15	**I** EcoSta2021 11:15 - 12:30	**Networking Lunch Break** 11:55 - 12:55
Networking Lunch Break 12:15 - 13:15	**Networking Lunch Break** 12:30 - 13:30	**O - Keynote** EcoSta2021 12:55 - 13:45
D EcoSta2021 13:15 - 15:20	**J** EcoSta2021 13:30 - 14:45	**P** EcoSta2021 13:55 - 15:35
E EcoSta2021 15:30 - 16:45	**K** EcoSta2021 14:55 - 16:35	**Q** EcoSta2021 15:45 - 17:25
F EcoSta2021 16:55 - 18:35	**L** EcoSta2021 16:45 - 18:25	**R - Keynote talk and closing** EcoSta2021 17:35 - 18:30
Virtual Welcome Reception 18:35 - 19:30		

VIRTUAL TUTORIAL, MEETINGS, SOCIAL EVENTS AND ACCESS TO VIRTUAL SPACES

VIRTUAL TUTORIAL

The tutorial “Robust methods for sample selection models and treatment effects” will take place on Wednesday, 23rd of June 2021, 15:00-19:30 (GMT+8). It will be delivered by Prof. Elvezio Ronchetti. Only participants who had subscribed for the tutorial can attend. Registered participants will be able to access the virtual tutorial through the website.

SPECIAL MEETINGS by invitation to group members

The EcoSta (Econometrics and Statistics) and CSDA (Computational Statistics and Data Analysis) Editorial Board meetings will take place on Wednesday the 23th of June 2021, 15:00-16:00 (GMT+8). Indications to attend the virtual Editorial Board meetings will be sent to the AEs attending the conference in due course.

ACCESS TO THE VIRTUAL CONFERENCE

- Your access to the virtual conference is personal and cannot be transferred to anybody else. If you share your credentials and any non-registered participant enters the meeting with them, your access will be banned.
- Keep at hand your registration number. You can find your registration number in the email confirming that you are registered or in the conference receipt that you can download from the registration tool. The conference staff may request this number for identification purposes.

Scientific programme and social events

- The conference is live streaming, and it will not be recorded. The oral presentations will take place through Zoom, while the social events and poster presentations will run in Gather Town.
- **Scientific programme:** The virtual sessions are accessible from the interactive schedule. The conference programme time is set in GMT+8. Indications to access the rooms can be found on the website.
- **Networking lunch breaks:** During lunchtime each day, the conference participants are invited to interact in the conference virtual networking space. Indications to access the networking space can be found on the website.
- **Welcome reception:** The virtual welcome reception for registered participants will take place on Thursday, 24th of June 2021, 18:35-19:30 (GMT+8). Indications to access the networking space can be found on the website.

Presentation instructions

The paper presentations will take place through Zoom. Speakers should install the application, have a stable internet connection, and make sure their video and audio work. They will share their slides when the chair requires it, present their talk, and be ready to answer the question after the presentation. Detailed indications for speakers can be found on the website. Each speaker has 20 minutes for the talk and 3-4 mins for discussion as a general rule. Strict timing must be observed.

Posters

The poster sessions will take place through Gather Town. The posters should be sent in **png format** to info@CMStatistics.org by the 22nd of June 2021. Landscape orientation is advisable. Detailed indications for the poster presentations can be found on the website.

Session chairs

The session chairs will be responsible for introducing the session, the speakers and coordinating the discussion time. A member of the conference staff, identified by the name *Angel* followed by the room number, will assist in giving the rights to participate as the chair requests it. If any speaker is missing or has a technical problem, the chair can pass to the next speaker and come back later to resume if possible. Detailed indications for the session chairs can be found on the website.

Test session

A test session will be set up for Sunday, 20th of June 2021, 14:00–15:00 (Hong Kong time, GMT+8). The participants will be able to enter any of the virtual rooms in the programme to test their presentations, video, micro and audio. Detailed indications for the test session can be found on the website.

PUBLICATION OUTLETS

The Elsevier journal Econometrics and Statistics (EcoSta) is the official journal of the conference. The CMStatistics network, co-organizer of the conference, also publishes the Annals of Statistical Data Science as a supplement to the journal Computational Statistics and Data Analysis (CSDA).

Econometrics and Statistics (EcoSta)

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

Econometrics and Statistics is the official journal of the networks Computational and Financial Econometrics and Computational and Methodological Statistics published by Elsevier (http://www.journals.elsevier.com/econometrics-and-statistics/). It publishes research papers in all aspects of econometrics and statistics and comprises of two sections:

- **Part A: Econometrics.** Emphasis will be 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 to be considered when they involve an original methodology. Innovative papers in financial econometrics and its applications will be considered. The topics to be covered include portfolio allocation, option pricing, quantitative risk management, systemic risk and market microstructure. Interest will be 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 will 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 will not be of interest to the journal.

- **Part B: Statistics.** Papers providing important original contributions to methodological statistics inspired in applications will be considered for this section. Papers dealing, directly or indirectly, with computational and technical elements will be particularly encouraged. These cover developments concerning issues of high-dimensionality, re-sampling, dependence, robustness, filtering, and, in general, the interaction of mathematical methods, numerical implementations and the extra burden of analysing large and/or complex datasets with such methods in different areas such as medicine, epidemiology, biology, psychology, climatology and communication. Innovative algorithmic developments are also of interest, as are the computer programs and the computational environments that implement them as a complement.

The journal consists, preponderantly, of original research. Occasionally, review and short papers from experts are published, which may be accompanied by discussions. Special issues and sections within important areas of research are occasionally published.

EcoSta Special Issues

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

The Econometrics and Statistics (EcoSta) is inviting submissions for the special issues (deadline 30th November 2021):

- Part A: Econometrics. Financial Technology (Fintech). Guest Editors: Monica Billio, Marc Paolella and Mike K.P. So.

- Part A: Econometrics. 2nd Special Issue on Time Series Econometrics. Guest Editors: Alessandra Amendola, Christian Francq, Marc Hallin and Zacharias Psaradakis.

- Part A: Econometrics and Part B: Statistics. 2nd Special Issue on Bayesian methods in statistics and econometrics. Guest Editors: Michele Guindani, Daniel Henderson, Maria Kalli and Yasuhiro Omori.

- Part A: Econometrics. Annals of Computational and Financial Econometrics

Further information can be found at `http://www.cmstatistics.org/EcoSta2021/EcoSta_specialissues.php`

CSDA Annals of SDS

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

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

Submissions for the Annals should contain a significant computational or statistical methodological component for data analytics. In particular, the Annals welcomes contributions at the interface of computing, statistics addressing problems involving large and/or complex data. Emphasis will be given to comprehensive and reproducible research, including data-driven methodology, algorithms and software. There is no deadline for submissions. Papers can be submitted at any time. When they have been received, they will enter the editorial system immediately. All submissions must contain original unpublished work not being considered for publication elsewhere.

Please submit your paper electronically using the Editorial Manager system (choose Article Type: Research paper, and then Select "Section IV. Annals of Statistical Data Science").

Contents

Thursday 24.06.2021 08:00 - 09:00 Room: R01 Chair: Mike So Opening and keynote talk 1

Testing for time stochastic dominance

Speaker: **Yoon-Jae Whang, Seoul National University, Korea, South**

Nonparametric tests are proposed for the null hypothesis of time stochastic dominance. Time stochastic dominance makes a partial order of different prospects over time based on the net present value criteria for general utility and time discount function classes. For example, time stochastic dominance can be used for ranking investment strategies or environmental policies based on the expected net present value of the future benefits. We consider an L_p integrated test statistic and derive its large sample distribution. We suggest path-wise bootstrap procedures that allow for time-dependence in a panel data structure. In addition to the least favorable case-based bootstrap method, we describe two approaches, the contact-set approach and the numerical delta method, to enhance the test's power. We prove the asymptotic validity of our testing procedures. We investigate the finite sample performance of the tests in simulation studies. As an illustration, we apply the proposed tests to evaluate the welfare improvement of Thailand's Million Baht Village Fund Program.

Saturday 26.06.2021 12:55 - 13:45 Room: R01 Chair: Ana Colubi Keynote talk 2

Joint modeling of complex data with latent variables

Speaker: **Xinyuan Song, Chinese University of Hong Kong, Hong Kong**

Several joint modeling approaches are introduced for analyzing complex data with latent variables. The models under consideration include a latent factor-on-image regression model to investigate the associations between imaging predictors and a latent outcome, a mediation analysis model to examine the causal effects of interest in the presence of latent mediators, and various survival models to reveal observed and latent risk factors for time-to-event outcomes. Statistical methods, including functional principal component analysis, the expectation-maximization algorithm, estimating equation method, spike-and-slab procedure, and Markov chain Monte Carlo sampling techniques, are used to conduct statistical inference. Applications to real-life studies regarding Alzheimer's disease and the complications of type 2 diabetes are presented.

Saturday 26.06.2021 17:35 - 18:30 Room: R01 Chair: Erricos John Kontoghiorghes Closing and keynote talk 3

Robust and consistent variable selection in high-dimensional generalized linear models

Speaker: **Elvezio Ronchetti, University of Geneva, Switzerland** Marco Avella-Medina

Generalized linear models are popular for modelling a large variety of data. We consider variable selection through penalized methods by focusing on resistance issues in the presence of outlying data and other deviations from assumptions in high dimensions. In particular, we discuss the connections between robustness, sparsity, and oracle properties and the extension of basic robustness concepts to the high-dimensional setting. Specifically, we highlight the weaknesses of widely used penalized M-estimators, propose a robust penalized quasi-likelihood estimator, and show that it enjoys oracle properties in high dimensions and is stable in a neighborhood of the model. We illustrate its finite sample performance on simulated and real data.

Thursday 24.06.2021 09:10 - 10:25 Parallel Session B – EcoSta2021

EO287 Room R01 ADVANCED INFERENTIAL TOOLS FOR MODERN DATA Chair: Daoji Li

E0238: **Statistical inference for the mean function of longitudinal imaging data over complicated domains**
Presenter: **Jie Li**, Center for Statistical Science and Department of Industrial Engineering, Tsinghua University, China

Motivated by longitudinal imaging data, which possesses inherent spatial and temporal correlation, a novel procedure is proposed to estimate its mean function. Functional moving average is applied to depict the dependence among temporally ordered images. Flexible bivariate splines over triangulations are used to handle the irregular domain of images common in imaging studies. The bivariate spline estimator's global and local asymptotic properties for mean function are established with simultaneous confidence corridors (SCCs) as a theoretical byproduct. Under some mild conditions, the proposed estimator and its accompanying SCCs are shown to be consistent and oracle efficient, as if all images were entirely observed without errors. The finite sample performance of the proposed method through Monte Carlo simulation experiments strongly corroborates the asymptotic theory. The proposed method is illustrated by analyzing two seawater potential temperature data sets.

E0382: **Testing cell-type-specific mediation effects in genome-wide epigenetic studies**
Presenter: **Zhonghua Liu**, The University of Hong Kong, Hong Kong

In current epigenome-wide mediation analysis, DNA methylation CpGsites that mediate a causal effect can only be identified. Because the methylation values are mixed from a heterogeneous population of cells, it is crucial to get fine-grained results by detecting mediation CpG sites in a cell-type-specific way. However, there is a lack of methods and software for testing cell-type-specific mediation effects. We propose a novel method, MICS (Methylation In a Cell-type-Specific fashion), to identify cell-type-specific mediation effects in genome-wide epigenetic studies. MICS first estimates the cellular compositions via a reference methylation matrix, then uses the estimated cell proportions to obtain the cell-type-specific p-values with respect to the exposure effects on the CpG sites as well as the methylation effects on the outcome, and finally combines the two p-value matrices using a joint-significance-followed by-squaring procedure. We conduct simulation studies to demonstrate that our method has correct type I error control and is powerful and robust under practical settings. We also apply our method to the Normative Aging Study and identify three CpG sites in the monocytes that might mediate the effects of smoking on lung function.

E0478: **Statistical inference for noisy matrix completion incorporating auxiliary information**
Presenter: **PoYao Niu**, University of California, Riverside, United States
Co-authors: Shujie Ma, Yinchu Zhu

Statistical inference is investigated for noisy matrix completion in a semi-supervised model when auxiliary covariates are available. The model consists of two parts. One part is a low-rank matrix induced by unobserved latent factors; the other part models the effects of the observed covariates through a coefficient matrix composed of high-dimensional column vectors. We propose an iterative least squares (LS) estimation approach that fully enjoys a low computational cost. We show that we only need to iterate the LS estimation a few times. The resulting entry-wise estimators of the target matrix and the coefficient matrix are guaranteed to have asymptotic normal distributions. As a result, a pointwise confidence interval and individual inference for each entry of the unknown matrices can be conducted. Moreover, we propose a simultaneous testing procedure with multiplier bootstrap for the high-dimensional coefficient matrix. This simultaneous inferential tool can help us further investigate the effects of auxiliary covariates to predict all missing entries.

EO133 Room R02 NOVEL DEVELOPMENTS IN LOW-RANK MODELING Chair: Raymond Ka Wai Wong

E0201: **A scalable algorithm for estimating a low-rank plus sparse matrix**
Presenter: **Eric Chi**, North Carolina State University, United States

The problem of estimating a low-rank plus sparse matrix from noisy and potentially incomplete measurements are considered. Examples of this problem include robust matrix recovery, compressive principal component pursuit, and LORS regression in analysing expression quantitative trait loci. Prior formulations of this problem involve solving a convex optimization problem, which requires repeatedly computing a singular value decomposition. We will discuss a non-convex formulation that admits more scalable algorithms that avoid computing an expensive singular value decomposition.

E0330: **Robust reduced rank regression in a distributed setting**
Presenter: **Xiaojun Mao**, Fudan University, China

This paper studies the reduced rank regression problem, which assumes a low-rank structure of the coefficient matrix, together with heavy-tailed noises. To address the heavy-tailed noise, we adopt the quantile loss function instead of the commonly used squared loss. However, the non-smooth quantile loss brings new challenges to both computation and the development of statistical properties, especially when the data is large in size and distributed across different machines. To this end, we first transform the response variable and reformulate the problem into a trace-norm regularized least-squares problem, which greatly facilitates the computation. Based on this formulation, we further develop a distributed algorithm. Theoretically, we establish the convergence rate of the obtained estimator and the theoretical guarantee for rank recovery. The simulation analysis is provided to demonstrate the effectiveness of our method.

E0509: **Bayesian spatial blind source separation via the thresholded Gaussian process**
Presenter: **Jian Kang**, University of Michigan, United States
Co-authors: Ben Wu, Ying Guo

Blind source separation (BSS) aims to separate latent source signals from their mixtures. For spatially dependent signals in high dimensional and large-scale data, such as neuroimaging, most existing BSS methods do not consider the spatial dependence and the sparsity of the latent source signals. We propose a Bayesian spatial blind source separation (BSP-BSS) approach for neuroimaging data analysis to address these major limitations. We assume the expectation of the observed images as a linear mixture of multiple sparse and piece-wise smooth latent source signals, for which we construct a new class of Bayesian nonparametric prior models by thresholding Gaussian processes. We assign the von Mises-Fisher priors to mixing coefficients in the model. Under some regularity conditions, we show that the proposed method has several desirable theoretical properties, including the large support for the priors, the consistency of joint posterior distribution of the latent source intensity functions and the mixing coefficients and the selection consistency on the number of latent sources. We use extensive simulation studies and an analysis of the resting-state fMRI data in the Autism Brain Imaging Data Exchange (ABIDE) study to demonstrate that BSP-BSS outperforms the existing alternatives for separating latent brain networks and detecting activated brain activation in the latent sources.

EO017 Room R03 EMERGING DEVELOPMENT IN THE ANALYSIS OF CENSORED DATA Chair: Wenqing He

E0417: Modeling time-dependent effects of covariates in the analysis of time-to-event data arising from family-based studies

Presenter: **Yun-hee Choi**, Western University, Canada
Co-authors: Seungwoo Lee, Laurent Briollais

Time-to-event data arising from family-based studies are often complex due to various factors such as multiple outcomes, interventions, competing risks and familial correlation. The aim is to estimate the cancer risks over time and evaluate several risk factors associated with cancer occurrence. Hereditary breast and ovarian cancer families suffer from high risks of both breast and ovarian cancers. They are recommended to undergo frequent screenings or prophylactic surgery for prevention or early detection. A shared frailty model with time-varying covariates is applied to evaluate the effects of mammographic screening and risk-reducing salpingo-oophorectomy on breast cancer risks. The evaluation of these interventions is usually complicated because of their effects changing over time and the presence of correlated competing risks. We propose a competing risks model that accounts for time-varying interventions and their time-dependent effects and provides cause-specific penetrance estimates for breast and ovarian cancers in BRCA1 families. We apply our approach to 498 BRCA1 mutation carrier families recruited through the Breast Cancer Family Registry and illustrate the importance of our approach accounted for both competing risks and time-varying effects when estimating cause-specific penetrance of breast cancer in the presence of ovarian cancer and death.

E0638: Semiparametric methods for left-truncated and right-censored survival data with covariate measurement error

Presenter: **Grace Yi**, University of Western Ontario, Canada

Biased samples caused by left-truncation (or length-biased sampling) and measurement error often accompany survival analysis. While such data frequently arise in practice, little work has been available to address these features simultaneously. We explore valid inference methods for handling left-truncated and right-censored survival data with measurement error under the widely used Cox model. We first exploit a flexible estimator for the survival model parameters, which does not require the baseline hazard function specification. To improve the efficiency, we further develop an augmented nonparametric maximum likelihood estimator. We establish asymptotic results and examine the efficiency and robustness issues for the proposed estimators. The proposed methods enjoy appealing features that the distributions of the covariates and truncation times are left unspecified. Numerical studies are reported to assess the finite sample performance of the proposed methods.

E0669: Length-biased and interval-censored data with a nonsusceptible fraction

Presenter: **Yingwei Peng**, Queenś University, Canada
Co-authors: Pao-sheng Shen, Hsin-Jen Chen, Chyong-Mei Chen

Left-truncated data are often encountered in epidemiological cohort studies, where individuals are recruited according to a certain cross-sectional sampling criterion. Length-biased data, a special case of left-truncated data, assume that the incidence of the initial event follows a homogeneous Poisson process. We will introduce an analysis of length-biased and interval-censored data with a nonsusceptible fraction in this talk. The Cox proportional hazards model for the survival time of the susceptible individuals and the logistic regression model for the probability of being susceptible are employed to model the data. We construct the full likelihood function and obtain the nonparametric maximum likelihood estimates of the regression parameters by employing the EM algorithm. The large sample properties of the estimates are established. The performance of the method is assessed by simulations. The proposed model and method are applied to data from an early-onset diabetes mellitus study.

EO181 Room R04 DEVELOPMENTS IN MARKOV CHAIN THEORY AND METHODOLOGY Chair: James Flegal

E0172: Bayesian model selection for ultrahigh-dimensional doubly-intractable distributions

Presenter: **Jaewoo Park**, Yonsei University, Korea, South
Co-authors: Ick Hoon Jin

Item response data are common, for example, cognitive-developmental stages data in educational sciences and measurements of depression of healthy controls. Although several item response theory (IRT) models have been developed for studying such data sets, each method depends on assumptions about dependence structure which are unrealistic in many settings. We propose inhomogeneous exponential random graph models (I-ERGMs) that can easily incorporate local dependence among items without any assumptions. However, in practice, I-ERGMs pose some inferential and computational challenges; likelihood functions involve intractable normalizing functions. An increasing number of items can lead to ultrahigh dimensionality in the model. To address such challenges, we develop novel Markov chain Monte Carlo methods using Bayesian variable selection methods to identify strong interactions automatically. We illustrate applying the approaches to challenging simulated, and real item response data examples for which studying local dependence is very difficult. The proposed algorithm shows significant inferential gains over existing methods in the presence of strong dependence among items.

E0230: Lugsail lag windows and their application to MCMC

Presenter: **Dootika Vats**, Indian Institute of Technology, Kanpur, India
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 family of lag windows specifically designed for improved finite sample performance. We use this lag window for weighted batch means and spectral variance estimators in Markov chain Monte Carlo simulations to obtain strongly consistent estimators that exhibit positive first-order bias and are asymptotically unbiased. This quality is beneficial when calculating effective sample size and using sequential stopping rules to help avoid premature termination. Further, we calculate the bias and variance of lugsail estimators and demonstrate little loss compared to other estimators. The finite sample properties of lugsail estimators are studied in various examples.

E0574: Visualizing simultaneous uncertainty in Monte Carlo experiments

Presenter: **James Flegal**, University of California - Riverside, United States
Co-authors: Dootika Vats, Galin Jones, Nathan Robertson

Monte Carlo experiments produce samples in order to estimate features of a given distribution. However, simultaneous estimation of means and quantiles has received little attention, despite being common practice. In this setting, we establish a multivariate central limit theorem for any finite combination of sample means and quantiles under the assumption of a strongly mixing process, which includes the standard Monte Carlo and Markov chain Monte Carlo settings. We build on this to provide a fast algorithm for constructing hyperrectangular confidence regions having the desired simultaneous coverage probability and a convenient marginal interpretation. The methods are incorporated into standard ways of visualizing the results of Monte Carlo experiments, enabling the practitioner to assess the reliability of the results more easily. We demonstrate the utility of this approach in various Monte Carlo settings, including simulation studies based on independent and identically distributed samples and Bayesian analyses using Markov chain Monte Carlo sampling.

EO161 Room R05 RECENT ADVANCES IN STATISTICAL LEARNING Chair: Paromita Dubey

E0209: **Detecting voice spoofing attacks with residual network and max filter map with Grad-CAM activation**
Presenter: **Il-Youp Kwak**, Chung-Ang University, Korea, South

The 2019 automatic speaker verification spoofing and countermeasures challenge (ASVspoof) competition aims to facilitate the design of highly accurate voice spoofing attack detection systems. However, they do not emphasize model complexity and latency requirements. Such constraints are strict and integral in a real-world deployment. Hence, most of the top-performing solutions from the competition use an ensemble approach and combine multiple complex deep learning models to maximize detection accuracy. This kind of approach would sit uneasily with real-world deployment constraints. To design a lightweight system, we combine the notions of skip connection (from ResNet) and max filter map (from Light CNN), and evaluate its accuracy using the ASVspoof 2019 dataset by optimizing a well-known signal processing feature called constant Q transform (CQT), our single model achieved a spoofing attack detection equal error rate (EER) of 0.16%, outperforming the top ensemble system from the competition that achieved an EER of 0.39% Furthermore, we applied Grad-CAM for the better explanation of our deep learning models on sound data.

E0503: **The cost of privacy in generalized linear models: Algorithms and minimax lower bounds**
Presenter: **Linjun Zhang**, Rutgers University, United States

Differentially private algorithms are proposed for parameter estimation in both low-dimensional and high-dimensional sparse generalized linear models (GLMs) by constructing private versions of projected gradient descent. We show that the proposed algorithms are nearly rate-optimal by characterizing their statistical performance and establishing privacy-constrained minimax lower bounds for GLMs. The lower bounds are obtained via a novel technique based on Stein's Lemma that generalizes the tracing attack technique for privacy-constrained lower bounds. This lower bound argument can be of independent interest as it applies to general parametric models. Simulated and real data experiments are conducted to demonstrate the numerical performance of our algorithms.

E0558: **Functional models for time varying random objects**
Presenter: **Paromita Dubey**, Stanford University, United States
Co-authors: Hans-Georg Mueller

In recent years, samples of time-varying object data such as time-varying networks that are not in a vector space have been increasingly collected. These data can be viewed as elements of a general metric space. Therefore common approaches that have been used with great success for the analysis of functional data, such as functional principal component analysis, cannot be applied directly. We will discuss ways to obtain dominant modes of variations in time-varying object data. We will describe metric covariance, a novel association measure for paired object data lying in a metric space that we use to define a metric auto-covariance function for a sample of metric space valued curves, where space will not have a vector space or manifold structure. Then the eigenfunctions of the linear operator with the auto-covariance function as the kernel can be used as building blocks for an object functional principal component analysis for object functional data, including time-varying probability distributions covariance matrices and time-dynamic networks. Finally, I will describe how one can obtain analogues of functional principal components for time-varying objects by applying Frechet means and projections of distance functions of the random object trajectories in the directions of the eigenfunctions.

EO033 Room R06 NOVEL STATISTICAL MODELING AND COMPUTING METHODS FOR COMPLEX DATA Chair: Tsung-I Lin

E0476: **A Bayesian nonparametric approach to inference for orthogonal nonnegative matrix factorization**
Presenter: **Hiroyasu Abe**, Kyoto University, Japan

Orthogonal non-negative matrix factorization (ONMF) is a technique that is applied to capture a hidden structure within a given data matrix with non-negative entries; examples of data matrices include gene expression data and document-term data. Using ONMF, the data matrix is decomposed into two matrices. The entries of these two matrices are non-negative, and one of the matrices is column orthogonal. Because the column orthogonal matrix with non-negative constraints plays the role of a membership matrix in k-means clustering, ONMF is also referred to as weighted spherical k-means clustering. One of the challenges of ONMF is model order selection, i.e., the process of inferring the number of bases or clusters. Although several model order selection approaches have been proposed for matrix decomposition techniques, to the best of our knowledge, none have been proposed for ONMF. Our proposed algorithm for ONMF is deterministic and is based on variational inference for Bayesian nonparametric modeling. The number of bases or clusters, as well as two decomposed matrices, are estimated in one run of the algorithm.

E0199: **Mixtures of multivariate t nonlinear mixed models for multiple longitudinal data with heterogeneity and missingness**
Presenter: **Wan-Lun Wang**, Feng Chia University, Taiwan

The multivariate t nonlinear mixed-effects model (MtNLMM) has been shown to be effective for analyzing multi-outcome longitudinal data following nonlinear growth patterns with fat-tailed noises or potential outliers. The problem of clustering heterogeneous longitudinal profiles in a mixture framework of MtNLMM is considered. A finite mixture of multivariate t nonlinear mixed model is proposed. This new model allows accommodating complex features of longitudinal data. Intermittent missing values frequently occur in the data collection process of multiple repeated measures. Under a missing at random (MAR) mechanism, a pseudo-data version of the alternating expectation conditional maximization (AECM) algorithm is developed to carry out maximum likelihood estimation and impute missing values simultaneously. The techniques for clustering incomplete multiple trajectories, recovery of missing responses, and estimation of random effects are also investigated. The utility of the proposed methods is illustrated through a simulation study and a real-data example coming from a study of pregnant women.

E0500: **Scale mixture of skew-normal linear mixed models with within-subject serial dependence**
Presenter: **Fernanda Schumacher**, University of Campinas - UNICAMP, Brazil
Co-authors: Victor Hugo Lachos Davila, Larissa Avila Matos

In longitudinal studies, repeated measures are collected over time, and hence they tend to be serially correlated. These studies are commonly analyzed using linear mixed models (LMMs). However, the usual Gaussian assumptions may result in a lack of robustness against departures from the normal distribution and invalid statistical inferences, especially when the data show heavy tails and skewness. In this regard, an extension of the scale mixture of skew-normal (SMSN) LMM is considered. The error term has a dependence structure, such as damped exponential correlation or autoregressive correlation of order p. The proposed model is estimated via maximum likelihood using an EM-type algorithm and provides flexibility in capturing the effects of skewness and heavy tails simultaneously when continuous repeated measures are serially correlated. The methodology is illustrated through an application to schizophrenia data using the R package skewlmm, which enables estimation and prediction of SMSN-LMMs with a user-friendly interface. Furthermore, some tools for model evaluation are discussed, including the Mahalanobis distance and the empirical autocorrelation function.

EO221 Room R07 RECENT ADVANCES IN MULTIVARIATE AND HIGH-DIMENSIONAL TIME SERIES Chair: Sayar Karmakar

E0609: **Simultaneous prediction intervals for high-dimensional vector autoregression**
Presenter: **Sayar Karmakar**, University of Florida, United States

Simultaneous prediction intervals are studied for high-dimensional vector autoregressive processes. Motivated from an online change-point problem, we wish to construct prediction intervals for one-step-ahead predictions in a high-dimensional VAR process. A de-biased calibration is used to post-regularize the lasso estimation, and a novel Gaussian-multiplier bootstrap-based method is developed for one-step-ahead prediction. The asymptotic coverage consistency of the prediction interval is obtained. We also substantiate our theoretical result by some simulations for evaluating finite sample performance and show some real data analysis. The simulated results show considerably good performance even when p is much larger than n, while most of the previous literature only focused on situations where p grows but remains less than n.

E0716: **A new clustering approach for spatio-temporal data**
Presenter: **Soudeep Deb**, Indian Institue of Management Bangalore, India
Co-authors: Sayar Karmakar

Spatio-temporal clustering is useful in various fields, such as climatology, epidemiology, social sciences. We develop a new algorithm that leverages both the spatial proximity and the temporal closeness for the individual locations. At the core of our method, we use a weighted combination of a spatial distance matrix and a temporal distance matrix between the locations. The haversine formula is used for the spatial distance, while for the latter, we use the L_2 distance between the spectral densities for every pair of time series. Next, the partition around medoids and the gap statistic are used to develop the algorithm and determine the optimal number of clusters. The proposed algorithm is attractive in three aspects. First, the weight parameter gives a great leeway to balance between the spatial and temporal closeness among the locations. Second, using the L_2 distance between spectral density estimates is a new contribution to time-series clustering. Third, the algorithm is flexible to allow different lengths of time series as well. For real-life applications, we cluster the time series of COVID19 incidence rate at the county level in the USA. The method provides interesting insights into the epidemic progression in the country. We also apply the algorithm on the spatio-temporal data of air temperature (for 67 years) from India and find similarities between different locations, which can be useful to understand global warming.

E0305: **Bayesian time-aligned factor analysis of paired multivariate time series**
Presenter: **Arkaprava Roy**, University of Florida, United States

Many modern data sets require inference methods that can estimate the shared and individual-specific components of variability in collections of matrices that change over time. Promising methods have been developed to analyze these data types in static cases, but only a few approaches are available for dynamic settings. To address this gap, we consider novel models and inference methods for pairs of matrices in which the columns correspond to multivariate observations at different time points. To characterize common and individual features, we propose a Bayesian dynamic factor modeling framework called Time Aligned Common and Individual Factor Analysis (TACIFA) that includes uncertainty in time alignment through an unknown warping function. We provide theoretical support for the proposed model, showing identifiability and posterior concentration. The structure enables efficient computation through a Hamiltonian Monte Carlo (HMC) algorithm. We show excellent performance in simulations and illustrate the method through application to a social synchrony experiment.

EO205 Room R08 NONPARAMETRIC AND SEMIPARAMETRIC METHODS Chair: Yuedong Wang

E0226: **Low rank approximation for smoothing splinevia eigensystem truncation**
Presenter: **Yuedong Wang**, University of California - Santa Barbara, United States

Smoothing splines provide a powerful and flexible means for nonparametric estimation and inference. With a cubic time complexity, fitting smoothing spline models to large data is computationally prohibitive. We use the theoretical optimal eigenspace to derive a low-rank approximation of the smoothing spline estimates. We develop a method to approximate the eigensystem when it is unknown and derive error bounds for the approximate estimates. The proposed methods are easy to implement with existing software. Extensive simulations show that the new methods are accurate, fast, and compares favorably against existing methods.

E0436: **Joint asymptotics for smoothing spline semiparametric nonlinear models**
Presenter: **Jiahui Yu**, Boston University, United States
Co-authors: Yuedong Wang, Anna Liu, Jian Shi

Density estimation and regression play fundamental roles in many areas of statistics and machine learning. Existing literature on semiparametric models is scattered and lacks a systematic framework. We consider unified frameworks based on reproducing kernel Hilbert space for modeling, estimation, and theory. We focus on a general (nonlinear) semiparametric density model, which includes many existing models as special cases. We establish joint consistency and derive convergence rates of the proposed estimators as well as the overall density function. Lastly, we present analogous results in the regression setting.

E0488: **Statistical computing meets quantum computing**
Presenter: **Ping Ma**, University of Georgia, United States

With the rapid development of quantum computers, quantum computing has been studied extensively. Unlike electronic computers, a quantum computer operates on quantum processing units, or qubits, which can take values 0, 1, or both simultaneously due to the superposition property. The number of complex numbers required to characterize quantum states usually grows exponentially with the size of the system. For example, a quantum system with p qubits can be in any superposition of 2^p orthonormal states simultaneously, while a classical system can only be in one state at a time. Such a paradigm change has motivated significant developments of scalable quantum algorithms in many areas. However, quantum algorithms tackling statistical problems are still lacking. We will present challenges and opportunities for developing quantum algorithms. We will introduce a novel quantum algorithm for the variable selection problem.

Thursday 24.06.2021 10:35 - 12:15 Parallel Session C – EcoSta2021

EO255 Room R01 STATISTICAL METHODS OF CAUSAL INFERENCE Chair: Yen-Tsung Huang

E0258: **A novel approach to semi-competing risks with left truncation via causal mediation modeling**
Presenter: **Jih-Chang Yu**, Academia Sinica, Taiwan
Co-authors: Yen-Tsung Huang

A novel method via causal mediation inference is proposed to analyze the semi-competing risk data under the left truncation sampling by considering the intermediate event as a mediator and the terminal event as an outcome. While the existing methods such as the copula models and multistate models focus on the correlation between the terminal event and primary event or transition between their states, we are interested in the causal relationship from the exposure to the outcome. We study the direct effect, the effect of the exposure on the terminal event, not through the intermediate event, and the indirect effect. We propose a nonparametric method and a semiparametric method, where both estimators account for left truncation. The non-parametric estimator can be viewed as a model-free time-varying Nelson-Aalen type of estimator and thus is a robust estimator; the semi-parametric estimator constructed by the Cox proportional hazards model is an efficient method with the flexibility of adjusting for potential confounders as covariates. Asymptotic properties for both estimators, including uniform consistency and weak convergence, are established by the martingale theorem and the functional delta method. The finite sample performance of the proposed estimators is evaluated via extensive numerical studies investigating the influence of left truncation, confounding, and sample size.

E0365: **Causal mediation of disease natural history**
Presenter: **Yen-Tsung Huang**, Academia Sinica, Taiwan

The natural history of human diseases comprises several milestones of sequential events where these milestones are all time-to-event by nature. For example, from hepatitis C infection to death, patients may experience intermediate events such as liver cirrhosis and liver cancer. The events of hepatitis, cirrhosis, cancer and death have the sequential order and are subject to right censoring; moreover, the latter events may mask the former ones. By casting the natural history of human diseases in the framework of causal mediation modeling, we set up a mediation model with intermediate events and a terminal event, respectively, as the mediators and the outcome. We define the interventional analogue of path-specific effects (iPSEs) as the effect of the exposure on the terminal event mediated by any combination of the intermediate events, including not through any of the events. We derive the expression of the counterfactual hazard. We construct composite nonparametric likelihood and derive a Nelson-Aalen type of estimator for the counterfactual hazard. We establish the asymptotic unbiasedness, uniform consistency and weak convergence for the proposed estimators. Numerical studies, including simulation and data application, illustrate the finite sample performance and utility of the proposed method.

E0372: **Multimodal neuroimaging data integration and pathway analysis**
Presenter: **Yi Zhao**, Indiana University, United States
Co-authors: Lexin Li, Brian Caffo

With fast advancements in technologies, collecting multiple types of measurements on a common set of subjects is becoming routine in science. Some notable examples include multi-modal neuroimaging studies for the simultaneous investigation of brain structure and function and multi-omics studies for combining genetic and genomic information. Integrative analysis of multimodal data allows scientists to interrogate new mechanistic questions. However, the data collection and generation of integrative hypotheses is outpacing available methodology for joint analysis of multimodal measurements. We study high-dimensional multimodal data integration in the context of mediation analysis. We aim to understand the roles different data modalities play as possible mediators in the pathway between an exposure variable and an outcome. We propose a mediation model framework with two data types serving as separate sets of mediators and develop a penalized optimization approach for parameter estimation. We study both the theoretical properties of the estimator through an asymptotic analysis and its finite-sample performance through simulations. We illustrate our method with a multimodal brain pathway analysis with structural and functional connectivities as mediators in the association between sex and language processing.

E0195: **Generalized interventional approach for causal mediation analysis with causally ordered multiple mediators**
Presenter: **Sheng-Hsuan Lin**, Institute of Statistics, Taiwan

Causal mediation analysis has shown the advantage of mechanism investigation. Under conditions with causally ordered mediators, path-specific effect (PSE) is introduced for specifying the effect mediated by a certain combination of mediators. However, most of PSEs are unidentifiable. An alternative approach, called interventional analogue of PSE (iPSE), is widely applied to effect decomposition to address this issue. Previous literature for multiple mediators mainly focused on discussing the case of two mediators due to the complexity of the mediation formula. A generalized method under the settings with an arbitrary number of mediators is attractive to study causal parameter identification and statistical estimation. A generalized interventional approach is proposed to discuss the effect mediated by ordered multiple mediators. It provides a general definition of iPSE by a recursive formula, assumptions for non-parametric identification, a regression-based method and a G-computation algorithm to estimate all iPSEs. This approach is applied to a Taiwanese cohort study for exploring the mechanism among hepatitis C virus on mortality through hepatitis B Virus, liver function, and hepatocellular carcinoma.

EO219 Room R02 CAUSAL INFERENCE AND LARGE-SCALE DATA ANALYSIS Chair: Daoji Li

E0599: **Nonparametric inference of heterogeneous treatment effects with two-scale distributional nearest neighbors**
Presenter: **Lan Gao**, University of Southern California, United States

Understanding heterogeneous treatment effects (HTE) plays a key role in many contemporary causal inference applications arising from different areas. Most of the existing works have focused on the estimation of HTE. Yet, the statistical inference aspect of the problem remains relatively undeveloped. We investigate the inference of HTE in a nonparametric setting for randomized experiments. We formulate the problem as two separate nonparametric mean regressions, one for the control group and the other for the treatment group. For each mean regression, we extend the tool of k-nearest neighbors to the framework of distributional nearest neighbors (DNN). To reduce the finite sample bias of DNN, we further suggest a new method of two-scale distributional nearest neighbors (TDNN). Under some regularity conditions, we show through delicate higher-order asymptotic expansions that the TDNN heterogeneous treatment effect estimator is asymptotically normal. We further establish the consistency of the variance estimates of the TDNN estimator with both jackknife and bootstrap, enabling user-friendly inference tools for heterogeneous treatment effects. The theoretical results and appealing finite-sample performance of the suggested TDNN method are illustrated with several simulation examples and a children's birth weight application.

E0597: **Parallel integrative learning for large-scale multi-response regression with incomplete outcomes**
Presenter: **Ruipeng Dong**, University of Science and Technology of China, China

Multi-task learning is increasingly used to investigate the association structure between multiple responses and a single set of predictor variables in many applications. In the era of big data, the coexistence of incomplete outcomes, a large number of responses, and high dimensionality in predictors pose unprecedented challenges in estimation, prediction and computation. We present a scalable and computationally efficient procedure, called PEER, for large-scale multi-response regression with incomplete outcomes, where both the numbers of responses and predictors can be high-dimensional. Motivated by sparse factor regression, we convert the multi-response regression into a set of univariate-response regressions, which can be efficiently implemented in parallel. Under some mild regularity conditions, we show that PEER enjoys nice sampling properties, including consistency in estimation, prediction, and variable selection. Extensive simulation studies demonstrate the effectiveness and scalability of our approach.

E0608: **Estimating mode effects from a sequential mixed-modes experiment**
Presenter: **Yanchun Bao**, University of Essex, United Kingdom
Co-authors: Paul Clarke

The large-scale household panel study Understanding Society (The U.K. Household Longitudinal Study UKHLS) has, until recently, used interviewers to administer its questionnaires but is now in the process of allowing individuals to participate using the web. Survey data are known to be affected by survey mode, so a sequential mode-effects experiment was carried out to evaluate the impact of this change on the panel. We present a novel estimator and analysis strategy to quantify the impact of mode across a wide range of variables, with large mode effects on the covariance of a pair of variables used to indicate an increased risk that statistical analyses involving this pair will be affected.

E0674: **A new group variable screening approach for ultrahigh-dimensional data**
Presenter: **Daoji Li**, California State University Fullerton, United States

In many applications with ultrahigh-dimensional data, variables are naturally grouped. There is a rich literature on variable screening, but most of the existing work does not consider the grouping structure. We propose a new group screening approach that allows predefined groups of predictors to be identified to address this issue jointly. The proposed method enjoys the sure screening property and does not require any assumption on the distribution of variables. Various numerical studies confirm the superior empirical performance of the proposed methods.

EO189 Room R03 RECENT ADVANCES IN STATISTICAL METHODS FOR COMPLEX BIOMEDICAL DATA Chair: Yingwei Peng

E0302: **Improving marginal hazard ratio estimation using quadratic inference functions**
Presenter: **Yi Niu**, Dalian University of Technology, China
Co-authors: Yingwei Peng, Hongkai Liang, Xiaoguang Wang

Clustered and multivariate failure time data are commonly encountered in biomedical studies and a marginal regression approach is often employed to identify the potential factors on the risk of failure. We consider a semiparametric marginal Cox proportional hazards model for right-censored survival data with potential correlation. We propose to use a quadratic inference function method based on the generalized method of moments to obtain the optimal hazard ratio estimators. The inverse of the working correlation matrix is represented by the linear combination of basis matrices in the context of the estimating equation. We investigate the asymptotic properties of the regression estimators from the proposed method. Our simulation study shows that the regression estimators from the quadratic inference approach are more efficient than that from the estimating equation method whether or not the working correlation structure is correctly specified. Finally, we apply the model and the proposed method to a real data analysis for illustration.

E0525: **Causal mediation analysis based on partial linear models**
Presenter: **Yeying Zhu**, University of Waterloo, Canada
Co-authors: Xizhen Cai, Yuan Huang

A set of generalized structural equations is proposed to estimate the direct and indirect effects for mediation analysis. We allow a nonlinear relationship among the baseline covariates and the response variables in each model. Since we are only interested in estimating the coefficients for the treatment and the mediator in the structural models, we assume partial linear models where the baseline covariates are regarded as a nuisance. The estimates can be interpreted as causal effects without the linearity assumption. We also propose variable selection procedures when the set of mediators is high-dimensional. Simulation results show the superior performance of our proposed method, and a data application is conducted when the set of candidate mediators are high-dimensional methylations.

E0639: **Support vector machine with graphical network structures in features**
Presenter: **Wenqing He**, University of Western Ontario, Canada

Regardless of being supervised or unsupervised, machine learning techniques have attracted extensive research attention in handling data classification. Typically, among supervised machine learning algorithms, Support Vector Machine (SVM) and its extensions have been widely used in various areas due to their great prediction capability. These learning algorithms basically treat features of the instances independently when using them to do classification. However, in applications, features are commonly correlated with complex network structures. Ignoring such a characteristic and naively implementing the SVM algorithm may yield erroneous classification results. To address the limitation of the SVM algorithm, we propose new learning algorithms that accommodate network structures in the features of the instances. Our algorithms capitalize on graphical model theory and make use of the available R software package for SVM. The implementation of the proposed learning algorithms is computationally straightforward. We apply the new algorithms to analyze the data arising from a gene expression study.

E0661: **A comparison between GBM and the Cox PH approaches with a focus on the R-packages GBM and survival**
Presenter: **Peizhi Li**, Dongbei University of Finance and Economics, China

Gradient Boosting Machine (GBM) and Cox Proportional Hazards (PH) are two popular statistical tools used to analyze survival data. However, fewer studies focus on the situation in which the two models perform better. We carry out two simulation studies to compare the two models comprehensively: one is PH is correctly specified, and the other is PH is misspecified, while GBM is fitted in the usual way in both two studies, i.e., ignoring the nonlinear and interaction terms. We divide the results into two parts: risk scores and baseline survival estimates. From the results, we find that when PH is correctly specified, risk scores by PH are better, and baseline survival estimates in linear scenarios by GBM are better. With increasing covariates, more risk scores of lower baseline hazard values in nonlinear scenarios by GBM tend to be better, and more baseline survival estimates in linear scenarios by PH are better. When PH is misspecified, it is found that PH can achieve similar performance when misspecified nonlinear terms but achieve worse performance when misspecified interaction terms. Finally, we use three real datasets to test our conclusions.

EO241 Room R04 BAYESIAN MODELING IN BIOSTATISTICS Chair: Michele Guindani

E0237: **Integrative graphical modeling and network mediation analysis**
Presenter: **Min Jin Ha**, UT MD Anderson Cancer Center, United States
Co-authors: Francesco Stingo, Veerabhadran Baladandayuthapani, James Long

Integrative network modeling of data arising from multiple genomic platforms provides insight into the holistic picture of the interactive system and the flow of information across many disease domains, including cancer. The basic data structure consists of a sequence of hierarchically ordered datasets for each individual subject, facilitating the integration of diverse inputs, such as genomics, transcriptomic, and proteomic data. In such contexts, a primary analytical task is to model the layered architecture of networks where the vertices can be naturally partitioned into ordered layers, dictated by multiple platforms, and exhibit both undirected and directed relationships. Given the underlying multi-layered graphical structure of data, we will discuss the Bayesian node-wise selection (BANS) approach to recover the multi-layered graphical structure, and causal mediation analysis framework that quantifies the paths one variable in a layer causes changes in another in the downstream layers.

E0736: **Horseshoe pit: A unified framework for large-scale Bayesian inference**
Presenter: **Francesco Denti**, University of California Irvine, United States

Variable selection is a central topic in supervised learning models. In a Bayesian linear regression framework, variable selection is performed by adopting a regularizing prior for the regression coefficients to shrink towards zero the irrelevant parameters. There are two main types of priors to accomplish this goal: the spike-and-slab and the continuous scale mixtures of Gaussians. The former is a discrete mixture between two distributions characterized by low and high variance. The latter specifies a hierarchical structure, where a continuous prior is elicited on the scale of a zero-mean Gaussian distribution.In contrast to these existing methods, we propose a discrete mixture of continuous scale mixtures, providing a connection between the two alternatives. We substitute the observation-specific local shrinkage parameters (typical of continuous mixtures) with cluster shrinkage parameters. Our proposal drastically reduces the number of parameters needed in the model and allows sharing statistical strength across coefficients of similar magnitude, improving the shrinkage effect. From a practical perspective, we adopt half-Cauchy priors. This choice leads to a cluster-shrinkage version of the Horseshoe prior, the Horseshoe Pit (HSP). We investigate the performance of the HSP in a multiple testing framework, applying our model to neurological data to detect activated brain regions.

E0155: **A common atom model for the Bayesian nonparametric analysis of nested data**
Presenter: **Michele Guindani**, University of California, Irvine, United States
Co-authors: Francesco Denti, Federico Camerlenghi, Antonietta Mira

The use of large datasets for targeted therapeutic interventions requires new ways to characterize the heterogeneity observed across subgroups of a specific population. In particular, models for partially exchangeable data are needed for inference on nested datasets, where the observations are assumed to be organized in different units, and some sharing of information is required to learn distinctive features of the units. We propose a nested Common Atoms Model (CAM) that is particularly suited for the analysis of nested datasets where the distributions of the units are expected to differ only over a small fraction of the observations sampled from each unit. The proposed CAM allows a two-layered clustering at the distributional and observational level. It is amenable to scalable posterior inference through the use of a computationally efficient nested slice sampler algorithm. We further discuss how to extend the proposed modeling framework to handle discrete measurements. We conduct posterior inference on a real microbiome dataset from a diet swap study to investigate how the alterations in intestinal microbiota composition are associated with different eating habits. We further investigate the performance of our model in capturing true distributional structures in the population through a simulation study.

EO123 Room R05 HYBRID STATISTICAL APPROACH FOR COMPLEX DATA Chair: Wenyu Gao

E0163: **Fixed-effects model: The most convincing model for meta-analysis with few studies**
Presenter: **Tiejun Tong**, Hong Kong Baptist University, Hong Kong

According to the Cochrane Database of Systematic Reviews, the median number of studies included in each meta-analysis is only three. In other words, about a half or more of meta-analyses conducted in the literature contain only two or three studies. When a meta-analysis involves few studies, it is known that the common-effect model and the random-effects model are likely to yield misleading or unreliable results. As another alternative, the fixed-effects model is also available in the literature for meta-analysis; yet for some reasons, this model was often overlooked until recently. For meta-analysis with few studies, the fixed-effects model can provide a good balance between the common-effect model and the random-effects model. It avoids the unreasonable assumption on a common effect in the common-effect model when the heterogeneity exists and avoids the low accuracy of the between-study variance estimate in the random-effects model when there are only a few studies. We propose to further improve the estimation accuracy of the average effect in the fixed-effects model by assigning different weight for each study and fully utilizing the information in the within-study variances. We demonstrate that the fixed-effects model can serve as the most convincing model for meta-analysis with few studies through theory and simulation.

E0216: **Performing genetic association tests in arbitrarily structured populations**
Presenter: **Minsun Song**, Sookmyung Women's University, Korea, South

A new statistical test of association between a trait and genetic markers is proposed. We theoretically and practically prove it 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 shown to provide accurate association tests in populations with complex structures, manifested in both the genetic and non-genetic contributions to the trait. The proposed framework provides a substantially different approach to the problem from existing methods.

E0741: **Using an estimated sparse two level gene pathway network for QDA to predict breast cancer patients treatment responses**
Presenter: **Liang Shan**, UAB, United States

Thanks to the development of microarray technology, gene-expression data are becoming more available, and it has been shown that genomic predictors have a great clinical impact in predicting treatment responses. For instance, pathologic complete response (pCR) is a strong indicator of survival after neoadjuvant chemotherapy for breast cancer patients. On the other hand, given that genes are grouped into pathways for particular functions and that pathways are not isolated, we have proposed a method to estimate the two-level Gaussian graphical models across heterogeneous classes jointly. While the method has the advantage of controlling sparsity on both pathway and gene-level networks, individually, the estimated precision matrices can also be used in quadratic discriminant analysis and its variants. We use some real breast cancer data and explore how the proposed estimation method, after incorporating the pathway information into the sparse precision matrix estimation, contributes to predicting breast cancer patients treatment responses via quadratic discriminant analysis and its variants.

E0749: **Nonparametric Bayesian functional clustering for breast cancer disparities**
Presenter: **Wenyu Gao**, Harvard University, United States
Co-authors: Inyoung Kim, Wonil Nam, Wei Zhou

It has been found that different incidence and mortality rates for breast cancer exist among various racial populations. For instance, Caucasian women are more likely to develop breast cancer than African American women. To study these disparities, surface-enhanced Raman spectroscopy (SERS) has been conducted to provide biomolecular fingerprint information. Extracellular SERS signals from each cell type were measured by a practical high-performance SERS device. However, large intraclass variations exist due to cellular and additional cancerous heterogeneity. Therefore, we need to reduce the amount of noise information and make each group distinguishable. The noises exist in two directions: the large number of heterogeneously behaved signal curves, as well as the massive change points on each curve. To study the differences between two types of triple-negative breast cancer cell lines at the molecular level, we performed functional cluster analyses and change point selection methods on the massive nonlinear curves of signals versus Raman shifts. Thus, we propose a nonparametric Bayesian functional clustering and change point selection method via weighted Dirichlet process mixture (WDPM) modeling, together with conditional Laplace prior. The proposed method is named WDPM-VS for short, and it will greatly outperform its comparison methods. Based on this proposed method, we identified important wavelengths that will explain the racial disparities.

EO069 Room R06 DESIGN AND ANALYSIS OF COMPLEX EXPERIMENTS: THEORY AND APPLICATIONS Chair: MingHung Kao

E0615: **Factor selection in screening experiments**
Presenter: **John Stufken**, University of North Carolina at Greensboro, United States
Co-authors: Rakhi Singh

Screening designs are used in experiments when, with limited resources, a few important factors are to be identified from a potentially large pool of factors. A second experiment often follows such an experiment to study the effect of the selected factors in more detail. Therefore, a screening experiment should ideally screen out many factors to make the follow-up experiment manageable without screening out important factors. The Gauss-Dantzig Selector (GDS) is often the preferred analysis method for screening experiments. When fitting a main-effects model, this can lead to incorrect conclusions if there are interactions. But including two-factor interactions in the model increases the number of model terms dramatically and challenges the GDS analysis. We discuss a new analysis method, called Gauss Dantzig Selector-Aggregation over Random Models (GDS-ARM), which aggregates the effects from different iterations of the GDS analysis performed using different sets of randomly selected interactions columns each time. The GDS-ARM draws its motivation in part from random forests by building many models and by identifying important factors after running a GDS analysis on all of these models. We will discuss the proposed method and study its performance.

E0707: **A general approach for constrained optimal design**
Presenter: **Saumen Mandal**, University of Manitoba, Canada

A brief introduction to optimal design theory and an overview of constrained optimal design using a Lagrangian approach will be given. The approach is quite appealing and has some potential applications. One such application is the problem of determining maximum likelihood estimates of cell probabilities under a hypothesis of marginal homogeneity for data in a square contingency table. This is an optimization problem with respect to variables that are nonnegative and satisfy several linear constraints. A second application is the problem of constructing optimal designs subject to achieving zero correlation among certain parameter estimators. To estimate certain parameters independently of the others, we require that the correlations between relevant parameter estimates be zero. The methodology can be applied to a wide class of estimation problems where constraints are imposed on the parameters. We will conclude with another potential application of the approach in model discrimination in which we optimize one objective subject to achieving a given efficiency of a parameter.

E0650: **Fast approximation of the Shapley values**
Presenter: **Wei Zheng**, University of Tennessee, United States

The Shapley value has been widely used in game theory, local model explanation and sensitivity analysis. However, its calculation is an NP-complete problem. Specifically, calculating a d-player Shapley value requires evaluating $d!$ or 2^d marginal contribution values, each associated with permutations of the d players. Hence it becomes infeasible to calculate the Shapley value when d is too large. A common remedy is to take a random sample of the permutations as a surrogate for the complete list of permutations. An advanced sampling scheme can be designed to yield a much more accurate estimation of the Shapley value than simple random sampling (SRS). Our sampling scheme is based on combinatorial structures in the field of design of experiment (DOE), particularly the order-of-addition experimental designs, to study how the orderings of the components would affect the output. We show that the obtained estimates are unbiased and consistent, sometimes even deterministically recover the original Shapley value based on the full permutation. Both theoretical and simulations results show that our DOE based sampling scheme outperforms SRS.

E0653: **Optimal designs for mixed continuous and binary responses with quantitative and qualitative factors**
Presenter: **MingHung Kao**, Arizona State University, United States

The focus is on optimal designs for multivariate regression with mixed variable types responses (continuous and binary) on quantitative and qualitative factors. New complete class results with respect to the Loewner ordering and relevant Chebyshev systems are derived to identify a small class of designs, within which locally optimal designs can be found for a group of models and optimality criteria. The complete class results facilitate the search for optimal designs via some general-purpose optimization techniques. Extensions of some previous results for characterizing optimal designs are also provided.

EC378 Room R07 CONTRIBUTIONS IN METHODOLOGICAL ECONOMETRICS Chair: Christopher Parmeter

E0695: **Inferential theory for granular instrumental variables in high dimensions**
Presenter: **Saman Banafti**, UC Riverside, United States

The Granular Instrumental Variables (GIV) methodology takes advantage of panel data to construct instruments to estimate structural time series regression models that involve endogenous regressors. The GIVs are constructed based on panel data models with factor structures, where the idiosyncratic error terms may have extraordinarily useful information. We extend the GIV methodology in several dimensions. First, we develop the GIV identification procedure to a large N and large T framework by establishing and restricting the asymptotic behavior of the Herfindahl index for large N markets as a function of the tail index of the size distribution of the cross-sectional units. Second, we allow the unknown factor loadings in the first stage regression to be estimated non-parametrically in constructing the GIVs, which therefore become “feasible” GIV (FGIV). Third, we show that the FGIV procedure requires estimation of a high dimensional precision matrix for large N and it is shown the sampling error in the high dimensional precision matrix is also asymptotically negligible when considering the limiting distribution. Fourth, we find instruments in addition to the GIVs to overidentify the structural parameters of interest. Monte Carlo analysis shows FGIV converges towards the infeasible oracle estimator. Finally, an empirical application of the FGIV estimation algorithm to estimate demand and supply elasticities of the global crude oil markets are presented.

E0469: **Shape constrained kernel PDF and PMF estimation**
Presenter: **Christopher Parmeter**, University of Miami, United States
Co-authors: Jeffrey Racine, Pang Du

The focus is on shape constrained kernel-based probability density function (PDF) and probability mass function (PMF) estimation, the former for modeling the PDF of a continuous random covariate and the latter for modeling the PMF of a categorical covariate. The proposal is of widespread potential applicability and includes, separately or jointly, constraints on the PDF (PMF) function itself, its integral (sum), and derivatives (finite-differences) of any order. We also allow for pointwise upper and lower bounds (i.e., inequality constraints) on the PDF and PMF in addition to more popular equality constraints. The approach handles a range of transformations of the PDF and PMF, including, e.g., logarithmic transformations (which allows for the imposition of log-concave or log-convex constraints that are popular with practitioners). Theoretical underpinnings for the procedures are provided. A simulation-based comparison of our proposed approach with those obtained using Grenander-type methods is favourable to our approach when the DGP is itself smooth. As far as we know, ours is also the only smooth framework that handles PDFs and PMFs in the presence of inequality bounds, equality constraints, and other popular constraints such as those mentioned above. Implementation in R exists that incorporates constraints such as monotonicity (both increasing and decreasing), convexity and concavity, and log-convexity and log-concavity, among others while allowing for bounded support.

E0493: **Inference using nuclear-norm penalized estimator and its applications**
Presenter: **Jungjun Choi**, Rutgers University, United States
Co-authors: Hyukjun Kwon

The inference theory of the (debiased) nuclear norm penalized estimator of the latent approximate low-rank matrix is studied when the observation matrix is subject to missingness. The alpha test in empirical asset pricing and the average treatment effect estimator are provided as applications. Although the nuclear norm penalization causes shrinkage bias which makes inference infeasible, our debiasing procedure successfully removes it, and the resulting debiased estimator attains the asymptotic normality. Unlike other debiasing schemes for the inference using the nuclear norm penalized estimator, our debiasing method does not resort to sampling splitting. So our estimation step is simple, and we can avoid some undesirable properties of sample splitting. In addition, the heterogeneous missing probability is allowed, and the inverse probability weighting is used, which improves the estimation performance by treating units with different missing probabilities in an equal manner.

E0331: **Binary response model with many weak instrumental variables**
Presenter: **Dakyung Seong**, University of Sydney, Korea, South

An endogenous binary response model with many weak instruments is considered. In contrast with linear simultaneous equations models, binary response models with endogenous regressors and many weak instruments have received minimal attention from researchers despite their practical importance. Two consistent and asymptotically normally distributed estimators are proposed: a regularized conditional maximum likelihood estimator (RCMLE) and a regularized nonlinear least square estimator (RNLSE), using regularization in the first stage. Standard regularization schemes such as Tikhonov regularization and Spectral cut-off can be employed for the proposed estimators, and consistent estimators of their asymptotic variances are also provided. Monte Carlo simulations show that both the RCMLE and the RNLSE outperform existing estimators when many weak instruments are present. We apply our estimators to two empirical examples to demonstrate their empirical relevance.

EC343 Room R08 CONTRIBUTIONS IN FINANCIAL ECONOMETRICS I Chair: Ke Zhu

E0613: **A new valuation measure for the stock market**
Presenter: **Andrey Sarantsev**, University of Nevada in Reno, United States

The classic Shiller cyclically adjusted price-earnings ratio, used to predict future total returns of the American stock market, is generalized. We split total returns into three components: earnings growth, dividend yield, and valuation change. The first two components are fundamental, the third is speculative. We develop two time series models: one for valuation change, and the other for dividend yield plus valuation change. These models are simple auto-regressions of order 1. The second model shows that long-run real return is equal to earnings growth plus 4-5%. We verify the classic 4% withdrawal rate: A retiree should invest in stocks and withdraw 4% of initial wealth after adjusting for inflation.

E0591: **Predicting gold risk premium using machine learning methods**
Presenter: **Erwin Hansen**, University of Chile, Chile

Forecasts of the gold risk premium obtained using several machine learning methods are compared for an extensive set of 174 predictors. We perform an out-of-sample evaluation and consider both statistical and portfolio metrics. The results show that neither individual machine learning methods nor forecast combinations outperform the historical mean when predicting the gold risk premium. Nevertheless, superior forecasts are obtained when using predictors individually. More specifically, we find that several technical indicators (moving average and momentum series) have forecasting power during expansion periods, while several macroeconomic variables such as the default spread, housing starts, the unemployment rate, and the producer price index help to predict the gold risk premium during recessions.

E0202: **Return predictability from variance differences: A fractionally co-integrated analysis**
Presenter: **Xingzhi Yao**, Xián Jiaotong Liverpool University, China
Co-authors: Marwan Izzeldin, Zhenxiong Li

Decomposing the realized and implied variances into upside and downside semi-variances, this paper examines the fractional co-integration for each pair of variances is examined. A positive link is revealed between the strength of the co-fractional relation and the return predictability afforded by the corresponding variance differences. Such a linkage is attributable to the common long-memory component in the fractionally co-integrated system, representing the volatility-of-volatility factor driving the variance risk premium. Despite the evidence for the presence of fractional co-integration between the upside and downside realized (implied) semi-variances, their difference is only weakly associated with future returns at low frequencies, and the predictive power dissipates over long horizons. The empirical results are further verified in a simulation study where the issue of the limited number and range of option strikes is alleviated in the construction of implied variances.

E0723: **Reflexivity analysis of cryptocurrencies with a time-varying semi-parametric Hawkes process**
Presenter: **Alev Atak**, METU, Turkey

The self-excitability and price clustering properties of the cryptocurrency market are studied to investigate the main sources of volatility, in particular the reflexivity or the endogeneity issues. We apply our kernel estimation of the spectrum localized both in time and frequency to data sets of transaction times, revealing pertinent features in the data that had not been made visible by classical non-localized approaches based on models with constant fertility functions over time. We apply the empirical analysis to the three largest crypto assets, i.e. Bitcoin - Ethereum - Ripple, and provide a comparison with other financial assets such as SP500, Gold, and the volatility index VIX observed from January 2018 to December 2020. The results show high levels of endogeneity in the basket of cryptocurrencies under investigation, underlining the evidence of a significant role of endogenous feedback mechanisms in the price formation process. We also demonstrate that the level of the endogeneity of markets, quantified by the branching ratio of the Hawkes process, is overestimated if the time variation is not considered.

Thursday 24.06.2021 13:15 - 15:20 Parallel Session D – EcoSta2021

EO370 Room R01 TRENDS AND DEVELOPMENTS IN REGULARIZATION AND SHRINKAGE Chair: Keisuke Yano

E0287: **Modeling the missing mechanism in partially ranked data with adjacency-based regularization**
Presenter: **Kento Nakamura**, The Universeity of Tokyo, Japan
Co-authors: Keisuke Yano, Fumiyasu Komaki

Ranked data appear in a wide variety of social domains such as election and consumer survey. They often comprise partial rankings that represent only a part of preferences. Partial rankings can be regarded as the result of missing from complete rankings. Although several studies have proposed estimators for partially ranked data, they ignore the modeling of a missing mechanism, which generally leads to significant bias in estimation. However, the modeling of a missing mechanism requires a large number of parameters, which leads to over-fitting when the sample size is small. To solve this trade-off, we propose the regularized estimator for both complete rankings and missing mechanisms based on an adjacency structure inherent to partially ranked data. We introduce the implementation of the proposed estimator by using a graph regularization framework and the expectation-maximization (EM) algorithm. Simulation studies and real data application indicate that the proposed estimator performs better than existing estimators under the non-ignorable missing mechanism.

E0418: **Shrinkage priors on complex-valued Gaussian processes**
Presenter: **Hidemasa Oda**, The University of Tokyo, Japan
Co-authors: Fumiyasu Komaki

Shrinkage priors on power spectral densities are investigated for complex-valued Gaussian processes. We propose general constructions of objective priors for Kahler parameter spaces. We discuss the importance of a positive continuous eigenfunction of the Laplace-Beltrami operator with a negative eigenvalue.

E0466: **Estimation under matrix quadratic loss and matrix superharmonicity**
Presenter: **Takeru Matsuda**, RIKEN Center for Brain Science, Japan
Co-authors: William Strawderman

The estimation of a normal mean matrix under the matrix quadratic loss is investigated. Improved estimation under the matrix quadratic loss implies the improved estimation of any linear combination of the columns. First, an unbiased estimate of risk is derived, and the Efron-Morris estimator is shown to be minimax. Next, a notion of matrix superharmonicity for matrix-variate functions is introduced and shown to have analogous properties with usual superharmonic functions, which may be of independent interest. Then, we show that the generalized Bayes estimator with respect to a matrix superharmonic prior is minimax. We also provide a class of matrix superharmonic priors that includes the previously proposed generalization of Steins prior. Numerical results demonstrate that matrix superharmonic priors work well for low-rank matrices.

E0675: **An information-geometric method for parameter estimation on moving-average models**
Presenter: **Yoshihiro Hirose**, Meiji University, Japan
Co-authors: Ryota Hasegawa, Hideyuki Imai

A sparse estimation method is proposed for moving-average (MA) models. Our method makes some sparse estimates of the autocovariance matrix of an MA model. The MA model is set to be a sufficiently large-dimensional model, and our method outputs sparse matrices of the model, which means that our method narrows down candidate models. The dimension selection of MA models is included by our problem because a band matrix is a special case of sparse matrices. In the method, an MA model, a set of MA processes, is treated as a manifold. From the viewpoint of geometry, parameter estimation is a selection of a point in the manifold. The sparse estimation method has some iterations of estimation, and, in each iteration, it selects a point in a lower-dimensional manifold included by the original manifold. Each point is obtained by shrinking the point made in the previous iteration. In the presentation, we illustrate our method and show the results of numerical experiments.

E0694: **Capturing network effect via fused lasso penalty with application on shared-bike data**
Presenter: **Yunjin Choi**, University of Seoul, Korea, South
Co-authors: Haeran Cho, Hyelim Son

Given a dataset with network structures, one common research interest is to model nodal features accounting for network effects. We investigate shared-bike data in Seoul under a spatial network framework. Each station's number of rental counts is modelled via regularized generalized linear model with fused lasso penalty terms. Parameters are posed in a station-wise manner, and fused lasso terms are applied on the station-wise parameters that are locationally close to each other. This approach encodes a station as a node and a pair of nearby stations as a connected edge. The fused lasso penalty term facilitates accounting for network effect by enabling neighboring stations to have the same estimation of the station-wise parameters. The proposed method shows promising results.

EO376 Room R03 STATISTICAL METHODS IN PHYLOGENY, POPULATION GENETICS AND COVID RESEARCH Chair: Andreas Futschik

E0576: **Polymorphism-aware phylogenetic models for species tree inference**
Presenter: **Carolin Kosiol**, University of St Andrews, United Kingdom
Co-authors: Rui Borges

Polymorphism-aware phylogenetic models (PoMo) constitute an alternative approach for species tree estimation from genome-wide data. PoMo builds on the standard substitution models of DNA evolution but expands the classic alphabet of the four nucleotide bases to include polymorphic states. By doing so, PoMo accounts for ancestral and current intra-population variation while also accommodating population-level processes ruling the substitution process (e.g. genetic drift, mutations, allelic selection). We were able to prove that PoMo is identifiable and that the maximum a posteriori (MAP) tree estimator of PoMo is a consistent estimator of the species tree. We complement our theoretical results with a simulated data set mimicking the diversity observed in natural populations exhibiting incomplete lineage sorting. We implemented PoMo in a Bayesian framework and show that the MAP tree easily recovers the true tree for typical numbers of sites sampled in genome-wide analyses. We will present what can be learned by applying these methods to genome-wide data sites of great ape populations about the ancestral population history of this species.

E0626: **Multiple haplotype reconstruction from allele frequency data**
Presenter: **Marta Pelizzola**, Vetmeduni Vienna, Austria

Haplotype information is usually of interest in genetics and medical studies. One haplotype consists of alleles from an individual that have been inherited together. Often sequencing each individual separately can be hard or costly, and thus pool sequencing techniques are employed. By doing so, the haplotype information is lost. This is the case, for example, for many experimental evolution experiments where the adaptation of one or multiple populations is studied over time, for viral evolution or evolutionary phenomena at a spacial level (e.g. adaptive radiation). Our goal is to exploit the power of the allele frequency data obtained by pool sequencing experiments to characterize the haplotype information. We model our problem as a multivariate regression problem where both the design and coefficient matrices are unknown. We build a method to reconstruct the haplotype structure and frequency for the most common individuals given multiple samples of allele frequency data starting from the existing theory. Additionally, we provide more reliable estimates of the allele frequencies if compared to those from pool sequencing. Our method is also not restricted to time series data. Since we do not directly model time, the different sources of information can also be collected in other ways (e.g. different spatial locations). This extends the applicability of our proposed method to several different fields.

E0651: **Using mobility data in the design of optimal lockdown strategies for the COVID-19 pandemic**
Presenter: **Ritabrata Dutta**, Warwick University, United Kingdom
Co-authors: Lorenzo Pacchiardi, Susana Gomes, Dante Kalise

A mathematical model for the COVID-19 pandemic spread, which integrates age-structured Susceptible-Exposed-Infected-Recovered-Deceased dynamics with real mobile phone data accounting for the population mobility, is presented. The dynamical model adjustment is performed via Approximate Bayesian Computation. Optimal lockdown and exit strategies are determined based on nonlinear model predictive control, constrained to public-health and socio-economic factors. Through an extensive computational validation of the methodology, it is shown that it is possible to compute robust exit strategies with realistic reduced mobility values to inform public policymaking. We exemplify the applicability of the methodology using datasets from England and France.

E0684: **Estimation of the largest mean Gaussian mixture component with population genetic applications**
Presenter: **Andreas Futschik**, JKU Linz, Austria

In population genetics, the effective population size *Ne* is a key parameter determining the amount of genetic drift. When using genomic time series data, estimates of *Ne* often rely on the changes in allele frequency at a large number of SNP positions. If a considerable proportion of the genome is affected by selection, these estimates will be biased, however, as both neutral and selected SNPs contribute. Due to the central limit theorem, estimates of Ne will typically be approximately normally distributed, given they are computed from a sufficient number of SNPs. Their mean and variance will differ, however, between neutral and selected regions. Gaussian mixture models would seem to be a natural approach to model such data. Since the selection strength will differ between selected positions, the number of mixture components may be large, making parameter estimation using standard methods such as the EM algorithm a challenge. We, therefore, propose a completely new approach that estimates only the largest (neutral) mixture component and does not infer the full mixture model. We illustrate its application to neutral Ne estimation in our discussed framework.

E0708: **REDACS: Regional emergency driven adaptive cluster sampling for effective COVID-19 prevalence**
Presenter: **Milan Stehlik**, Johannes Kepler University Linz, Austria

As COVID-19 is spreading, national agencies need to monitor and track several metrics. Since we do not have perfect testing programs at hand, one needs to develop advanced sampling strategies for prevalence study. The recent importance of COVID-19 mitigation strategies motivates the necessity of scalable, interpretable and precise methodology, which has materialized as REDACS: Regional emergency driven adaptive cluster sampling for effective COVID-19 prevalence. We justify its usage as a COVID-19 mitigation strategy. We will discuss the feasibility of REDACS implementations. We show its advantages over classical massive individual testing sampling plans. We also point out how regional and spatial heterogeneity underlines proper sampling. The fundamental importance of adaptive control parameters from emergency health stations and the medical frontline is outlined. Since the Northern hemisphere entered the Autumn and Winter season, practical illustration from spatial heterogeneity of Chile (Southern hemisphere, which already experienced COVID-19 winter outbreak peak) is underlying the importance of proper regional heterogeneity of sampling plan. We explain the regional heterogeneity by microbiological backgrounds and link it to the behavior of Lyapunov exponents. We also discuss screening by antigen tests from the perspective of biomarker validation.

EO131 Room R04 ADVANCES IN EXTREMAL M-QUANTILE REGRESSION Chair: Abdelaati Daouia

E0480: **Extremiles: A new perspective on asymmetric least squares**
Presenter: **Gilles Stupfler**, ENSAI - CREST, France
Co-authors: Abdelaati Daouia, Irene Gijbels

Quantiles and expectiles of a distribution are 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 a better capability of fitting both the 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. Several 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.

E0474: **Extremile regression**
Presenter: **Abdelaati Daouia**, Fondation Jean-Jacques Laffont, France
Co-authors: Gilles Stupfler, Irene Gijbels

Regression extremiles define a least-squares analogue of regression quantiles. They are determined by weighted expectations rather than tail probabilities. Of special interest is their intuitive meaning in terms of expected minima and maxima. Their use appears naturally in risk management where, in contrast to quantiles, they fulfil the coherency axiom and take the severity of tail losses into account. In addition, they are comonotonically additive and belong to both the families of spectral risk measures and concave distortion risk measures. This paper provides the first detailed study exploring implications of the extremile terminology in a general setting of the presence of covariates. We rely on local linear (least squares) check function minimization for estimating conditional extremiles and deriving the asymptotic normality of their estimators. We also extend extremile regression far into the tails of heavy-tailed distributions. Extrapolated estimators are constructed, and their asymptotic theory is developed. Some applications to real data are provided.

E0750: **A unified approach for the estimation of some important risk measures**
Presenter: **Simone Padoan**, Bocconi University, Italy
Co-authors: Gilles Stupfler

A likelihood-based inferential approach is discussed for estimating extreme risk measures such as Value-at-Risk, Expected shortfall, Expectile and Extremile. In particular, we focus on a Bayesian framework that allows us to easily derive point and interval estimators of such risk measures. The performance of the proposed approach is illustrated and some extensions of the basic setting consisting of independent and identically distributed random variables are discussed.

E0471: **Extreme conditional expectile estimation in heavy-tailed heteroscedastic regression models**
Presenter: **Antoine Usseglio-Carleve**, Fondation Jean-Jacques Laffont, France
Co-authors: Gilles Stupfler, Stephane Girard

Expectiles define a least-squares analogue of quantiles. They have been the focus of a substantial quantity of research in the context of actuarial and financial risk assessment over the last decade. The behaviour and estimation of unconditional extreme expectiles using independent and identically distributed heavy-tailed observations have been investigated in recent papers. We build a general theory for the estimation of extreme conditional expectiles in heteroscedastic regression models with heavy-tailed noise; the approach is supported by general results of independent interest on residual-based extreme value estimators in heavy-tailed regression models and is intended to cope with covariates having a large but fixed dimension. We demonstrate how the results can be applied to a wide class of important examples, among which linear models, single-index models, and ARMA and GARCH time series models. The estimators are showcased on a numerical simulation study and real sets of actuarial and financial data.

E0475: **Extremal expectile regression**
Presenter: **Yasser Abbas**, Fondation Jean-Jacques Laffont, France
Co-authors: Abdelaati Daouia, Gilles Stupfler

Studying rare events at the tails of heavy-tailed distributions is a burgeoning science and has many applications both in and out of finance. Most attempts to tackle the subject involve quantile regression, which usually offers a natural way of examining the impact of covariates at different levels of the dependent variable. We argue, however, that quantiles are not well equipped to deal with sparsity around the tails, especially in the active field of risk management, and motivate their least-square analogues, expectiles, as a more appropriate alternative. We introduce versatile estimators of tail conditional expectiles under an extremal additive regression model with heavy-tailed regression noise and derive their asymptotic properties in a general setting. We then tailor the discussion to the local linear estimation approach. We showcase the performance of our procedures in a detailed simulation study and apply them to a concrete dataset.

EO354 Room R05 TRENDS AND INNOVATIONS IN FINANCIAL MODELLING Chair: Rogemar Mamon

E0375: **The valuation of a guaranteed minimum maturity benefit under a regime-switching framework**
Presenter: **Heng Xiong**, Wuhan University, China
Co-authors: Rogemar Mamon, Yixing Zhao

Global insurance markets have become more sophisticated in recent times in response to the evolving needs of populations that tend to live longer. Policyholders desire the benefits of longevity/mortality protection while taking advantage of investment growth opportunities in equity markets. As a result, insurers incorporate payment guarantees in new insurance products, known as equity-linked contracts, whose values are dependent on the prices of risky assets. A guaranteed minimum maturity benefit (GMMB) is now common in many equity-linked contracts. We develop an integrated pricing framework for a GMMB focusing on segregated fund contracts. More specifically, we construct hidden Markov models (HMMs) for a stock index, interest rate, and mortality rate. The dependence between these risk factors is characterized explicitly. We assume that the stock index follows a Markov-modulated geometric Brownian motion, and the interest and mortality rates have Markov-modulated affine dynamics. A series of measure changes is employed to obtain a semi-closed-form solution for the GMMB price. A Fourier transform method is applied to approximate the prices more efficiently numerically. Recursive HMM filtering is used in our model calibration. Numerical investigations in our article demonstrate the accuracy of GMMB prices, and extensive analysis is included to systematically examine how risk factors affect the value of a GMMB.

E0407: **Embedding regime-switching in modelling of credit spreads through neural networks**
Presenter: **Christina Erlwein-Sayer**, University of Applied Sciences HTW Berlin, Germany
Co-authors: Stefanie Grimm

Corporate credit spreads are modelled through a hidden Markov model (HMM) based on a discretized Hull-White approach. The mean-reverting interest rate model parameters are governed by a discrete-time Markov chain enabling predictions of the time series through adaptive filtering. We forecast the credit spreads and filter out state-related information which is hidden in the observed spreads. We build an artificial neural network (ANN) that utilises regime-switching information to predict the credit spread and deduce the default probability of the corporate. The performance of the ANN is analysed and compared to the accuracy of an ANN without the regime-switching information.

E0623: **Time inconsistent optimal stopping under model ambiguity and financial applications**
Presenter: **Xiang Yu**, The Hong Kong Polytechnic University, Hong Kong
Co-authors: Yu-Jui Huang

An unconventional approach for optimal stopping under model ambiguity is introduced. Besides ambiguity itself, we take into account how ambiguity-averse an agent is. This inclusion of ambiguity attitude, via an alpha-maxmin nonlinear expectation, renders the stopping problem time-inconsistent. We look for subgame perfect equilibrium stopping policies, formulated as fixed points of an operator. We show that any initial stopping policy will converge to equilibrium through a fixed-point iteration for a one-dimensional diffusion with drift and volatility uncertainty. This allows us to capture much more diverse behavior, depending on an agent's ambiguous attitude, beyond the standard worst-case (or best-case) analysis. In a concrete example of real options valuation under model ambiguity, all equilibrium stopping policies and the best one among them are fully characterized under appropriate conditions. It explicitly demonstrates the effect of ambiguity attitude on decision-making: the more ambiguity-averse, the more eager to stop, to withdraw from the uncertain environment. The main result hinges on a delicate analysis of continuous sample paths in the canonical space and the capacity theory. To resolve measurability issues, a generalized measurable projection theorem, new to the literature, is also established.

E0630: **Adaptive online portfolio selection with transaction costs**
Presenter: **Sini Guo**, The University of Hong Kong, Hong Kong

As an application of machine learning techniques in financial fields, online portfolio selection has attracted great attention from practitioners and researchers, making timely sequential decision-making available when market information is constantly updated. For online portfolio selection, transaction costs incurred by changes of investment proportions on risky assets significantly impact the investment strategy and the return in the long-term investment horizon. However, in many online portfolio selection studies, transaction costs are usually neglected in the decision-making process. We consider an adaptive online portfolio selection problem with transaction costs. The adaptive online moving average method (AOLMA) is proposed to predict the future returns of risky assets by incorporating an adaptive decaying factor into the moving average method, which improves the accuracy of return prediction. The net profit maximization model (NPM) is then constructed where transaction costs are considered in each decision-making process. The adaptive online net profit maximization algorithm (AOLNPM) is designed to maximize the cumulative return by integrating AOLMA and NPM. Numerical experiments show that AOLNPM dominates several state-of-the-art online portfolio selection algorithms in terms of various performance metrics, i.e., cumulative return, mean excess return, Sharpe ratio, Information ratio and Calmar ratio.

E0655: **Calibrating with a smile the Mellin transform way**
Presenter: **Marianito Rodrigo**, University of Wollongong, Australia

It is well known that the volatility in the Black-Scholes framework is not a constant but a function of both the strike price (“smile/skew”) and the time to maturity. A popular approach to recovering the volatility surface is using deterministic volatility function models via Dupire's equation. A new method for volatility surface calibration based on the Mellin transform is proposed. An explicit formula for the volatility surface is obtained in terms of the Mellin transform of the call option price with respect to the strike price, and a numerical algorithm is provided. Results of numerical simulations are presented, and the stability of the method is numerically verified.

EO281 Room R06 RECENT ADVANCES IN APPLIED PROBABILITY AND STATISTICS Chair: Li-Hsien Sun

E0532: **Detection of threshold points in mean and variance forthreshold regression models**
Presenter: **ChihHao Chang**, National University of Kaohsiung, Taiwan

The threshold regression model is considered with one threshold point in the mean and in the variance, respectively, for dependent data with heteroscedasticity. We denote the threshold regression model in the mean without the continuity constraint at the threshold point by M2. We then provide an ordered iterative least squares (OiLS) method when estimating M2 and establish the consistency of the OiLS estimator under mild conditions. We denote the model by M1 when the continuity constraint is imposed on the threshold regression model. Further, we denote the model with no threshold points by M0 and apply a model selection procedure to select the three models. We establish the selection consistency under regularity conditions. The same estimation and selection procedures are further applied to detect the threshold point in the variance of the models.

E0573: **Bayesian clustering of spatial functional data based on random spanning trees**
Presenter: **Bohai Zhang**, Nankai University, China, China
Co-authors: Huiyan Sang, Hui Huang

A Bayesian wavelets model is proposed for modeling and clustering the spatial functional data, where domain partitioning is achieved by operating on the spanning trees. By imposing a proper prior on the spanning trees, the resulting clusters can have arbitrary shapes and are spatially contiguous in the input domain. The within-cluster parameters are updated through Gibbs samplers, and the between-cluster parameters are updated using a reversible jump Markov chain Monte Carlo (RJ-MCMC) algorithm. The numerical results show that the proposed model can identify the true clusters and yield reasonable parameter estimates. We then apply our model to the mobility dataset of Harris County in Houston during the COVID-19 pandemic.

E0578: **A unified framework for specification tests of continuous treatment effect models**
Presenter: **Zheng Zhang**, Renmin University of China, China
Co-authors: Oliver Linton

A general framework is proposed for the specification testing of continuous treatment effect models. We assume a general residual function, which includes the average and quantile treatment effect models as special cases. The null models are identified under the confoundedness condition and contain a nonparametric weighting function. We propose a test statistic for the null model in which the weighting function is estimated by solving an expanding set of moment equations. We establish the asymptotic distributions of our test statistic under the null hypothesis and under fixed and local alternatives. The proposed test statistic is shown to be more efficient than that constructed from the true weighting function and can detect local alternatives deviated from the null models at the rate of $\sqrt{N}$. A simulation method is provided to approximate the null distribution of the test statistic. Monte-Carlo simulations show that our test exhibits a satisfactory finite-sample performance, and an application shows its practical value.

E0619: **Randomized algorithms for functional data analysis**
Presenter: **Shiyuan He**, Renmin University of China, China

Functional data analysis is a well-established field of statistics, where continuous functions are viewed as a whole in infinite-dimensional Hilbert space. We are equipped with the ever-increasing ability to collect functional data too large for classical methods to handle. We adopt the randomized algorithms to tackle the large scale problems for common tasks such as covariance operator estimation, functional principal component analysis and functional linear regression. We propose a fast two-stage algorithm that reduces the sampling uncertainty. Both theoretical and empirical justification is provided for the proposed algorithm.

E0612: **Mean-field games with heterogeneous groups: Application to banking systems**
Presenter: **Li-Hsien Sun**, National Central University, Taiwan

The purpose is to study the system of heterogeneous interbank lending and borrowing based on the relative average of log-capitalization through the linear combination of the average within groups and the ensemble average and describe the evolution of log-capitalization by a system of coupled diffusions. The model incorporates a game feature with homogeneity within groups and heterogeneity between groups where banks search for the optimal lending or borrowing strategies and intend to minimize the heterogeneous linear quadratic costs in order to remain survival in the system. In addition, large banks intend to mean revert to the ensemble average in their own group through the term of the relative performance. In contrast, small banks intend to trace the overall ensemble average rather than the ensemble average in their own group. We obtain the Markov Nash equilibria governed by the mean-reverting term and the extra ensemble averages of individual groups given by heterogeneity. In addition, the corresponding heterogeneous mean-field game is also discussed.

E0757: **Estimating the covariance of fragmented and other related types of functional data**
Presenter: **Wei Huang**, University of Melbourne, Australia
Co-authors: Aurore Delaigle, Peter Hall, Alois Kneip

We consider the problem of estimating the covariance function of functional data which are only observed on a subset of their domain, such as fragments observed on small intervals or related types of functional data. We focus on situations where the data enable to compute the empirical covariance function or smooth versions of it only on a subset of its domain which contains a diagonal band. We show that estimating the covariance function consistently outside that subset is possible as long as the curves are sufficiently smooth. We establish conditions under which the covariance function is identifiable on its entire domain and propose a tensor product series approach for estimating it consistently. We derive asymptotic properties of our estimator and illustrate its finite sample properties on simulated and real data.

EO053 Room R08 ADVANCES IN FINANCIAL TIME SERIES ANALYSIS Chair: Toshiaki Watanabe

E0617: **Bayesian inference of temporal trends in daily financial news**
Presenter: **Mike So**, The Hong Kong University of Science and Technology, Hong Kong

Dynamic topic models offer a framework to model the temporal evolution of topics in news corpora through distributions of document-topic portions and topic-vocabulary portions. We consider a dynamic linear topic model which captures the topic evolution through a dynamic linear model structure. An advantage of dynamic linear modeling is to enable more complex trends (linear and quadratic) and seasonality in the news corpora. Bayesian data augmentation is implemented along with conditional independences to allow MCMC algorithms to be highly parallelized for inference on large corpora. We analyze daily financial news to estimate the topic evolutions to learn seasonal patterns and other local trends in this context. The analysis of the time series of new topics can provide interesting insights for financial risk assessment.

E0310: **Bayesian analysis of price discovery on time-varying partial adjustment model**
Presenter: **Kenji Hatakenaka**, Osaka University, Japan
Co-authors: Kosuke Oya

Price discovery is an important built-in function of financial markets and the central issue in microstructure research. Market participants need to know whether the price discovery has been achieved or how much progress has been made to trade at an appropriate price they consider. Since various economic events such as earning announcement affect price discovery, the intraday transition of price discovery varies date-by-date. We propose a statistical method to see when, how fast, and how accurate the intraday price discovery works using only the high-frequency price series in a single day. The method consists of estimating candidate models and selecting the most appropriate model based on a Bayesian approach. We conduct simulation studies to examine the performance of the proposed method and confirm the most reliable selection criteria. We will report how the selection criteria work and result from an empirical study using actual financial data.

E0314: **On the evaluation of intraday market quality in the limit-order book markets: A collaborative filtering approach**
Presenter: **Makoto Takahashi**, Hosei University, Japan
Co-authors: Takaki Hayashi

A methodology for evaluating the liquidity of individual stocks in the high-frequency domain is considered by utilizing a framework from recommender systems that have become ubiquitous in our daily lives. In general, it is not necessarily easy to evaluate the “true" liquidity of individual stocks. In particular, evaluating liquidity over a shorter term with high-frequency data can be challenging for many stocks due to the increasing sparsity of observations. Since stocks that have exhibited similar behavior in the past are expected to perform so in the future as well, one can expect that collaborative filtering, which is the main approach of recommender systems, can work effectively for the liquidity “estimation" problem. Specifically, we adopt a regression-based latent factor model (RLFM), hybrid-type collaborative filtering. It has a hierarchical structure designed to address the so-called “cold-start" problem in the recommender systems literature. As a result of the empirical analysis using high-frequency limit-order book data from the Tokyo Stock Exchange, various characteristics that characterize liquidity were identified from the estimated regression coefficients obtained by fitting the RLFM to the training dataset. In the meantime, there was room for improvement of the methodology regarding the accuracy of liquidity prediction.

E0326: **A time-varying jump tail risk measure using high-frequency options data**
Presenter: **Masato Ubukata**, Meiji Gakuin University, Japan

The aim is to propose a procedure for measuring daily jump tail risks obtained from high-frequency option data. The essentially nonparametric measure is difficult to construct if we use daily closing quotes and prices of the short-dated and deep-out-of-the-money options. Previous studies usually assume that the time-varying shape parameters of risk-neutral jump tails in asset returns change at a weekly frequency to mitigate the impact of noise. The high-frequency options data, which needs data cleaning, help relax the constancy assumption to more general cases such that the shape parameter can change at a daily frequency. In application to the high-frequency options data on the Nikkei 225 market index, we confirm that our daily tail risk measure is reasonably coherent with the existing measures from previous research and reveals relatively large spikes on particular days during the week associated with tail events. Our empirical analyses of the short-term predictability of variance risk premium (VRP), obtained as the difference between the option-based risk-neutral and the statistically expected future return variation, suggest that the daily tail risk measure, which is a jump tail risk component of VRP, has significant predictive power for future VRP, and that the inclusion of the diffusive and jump risk components of VRP as separate predictors yield the forecast improvement.

E0526: **Stochastic volatility models with time-varying leverage effect**
Presenter: **Jouchi Nakajima**, Bank of Japan, Japan
Co-authors: Toshiaki Watanabe

A stochastic volatility model with a time-varying leverage effect is proposed. The leverage effect, which is captured by a correlation coefficient between innovations to today’s return and tomorrow’s volatility in a standard stochastic volatility model, is assumed to evolve according to an autoregressive process. An efficient Bayesian method via Markov chain Monte Carlo is developed for the estimation of the proposed model. An empirical analysis using daily stock returns, foreign exchange rate, and cryptocurrency provides evidence that the leverage effect considerably changes over time.

EC341 Room R02 CONTRIBUTIONS IN TIME SERIES Chair: Matthieu Marbac

E0485: **Wilks theorem for semiparametric regressions with weakly dependent data**
Presenter: **Marie du Roy de Chaumaray**, CREST-ENSAI, France
Co-authors: Valentin Patilea, Matthieu Marbac

The empirical likelihood inference is extended to a class of semiparametric models for stationary, weakly dependent series. A partially linear single-index regression is used for the conditional mean of the series given its past and the present and past values of a vector of covariates. A parametric model for the conditional variance of the series is added to capture further nonlinear effects. We propose suitable moment equations which characterize the mean and variance model. We derive an empirical log-likelihood ratio which includes nonparametric estimators of several functions, and we show that this ratio behaves asymptotically as if the functions were given.

E0721: **Functional estimation and change detection for nonstationary time series**
Presenter: **Fabian Mies**, RWTH Aachen University, Germany

Tests for structural breaks in time series should ideally be sensitive to breaks in the parameter of interest while being robust to nuisance changes. Thus, the statistical analysis needs to allow for some form of nonstationarity under the null hypothesis of no change. We construct estimators for integrated parameters of locally stationary time series. A corresponding functional central limit theorem is established, enabling change-point inference for a broad class of parameters under mild assumptions. The proposed framework covers all parameters that may be expressed as nonlinear functions of moments, such as kurtosis, autocorrelation, and coefficients in a linear regression model. A bootstrap variant is proposed to perform feasible inference based on the derived limit distribution, and its consistency is established. The methodology is illustrated through a simulation study and by an application to high-frequency asset prices.

E0223: **Modified upper prediction limit of vector autoregressive model: the case for estimating value-at-risk**
Presenter: **Aniq Rohmawati**, Telkom University, Indonesia

An important issue in a modern society concerning on a stability system is people entrust risk assessment as safeguard. This paper proposes the appropriate modelling for upper limit prediction corresponding to risk/loss for future observation. In most events, each observation inherently allows a causal relationship to other observations, possibly as a linear function. A Vector Autoregressive (VAR) model handling the instantaneous interaction between response and predictors over time series horizon is proposed. It is further recognized that VAR allowed lagged values of multivariate time series, also outlined time-varying parameter to address essential drifts in coefficient. We shall examine a measure of upper prediction namely modified Value-at-Risk, addressed by involving parameters estimation of VAR model. Result shows that VAR model allows to predict accurately the coverage probability of Value-at-Risk.

E0711: **FOCuS: Online changepoint detection with a constant per-iteration computational cost**
Presenter: **Gaetano Romano**, Lancaster University, United Kingdom
Co-authors: Idris Eckley, Paul Fearnhead, Guillem Rigaill

Changepoint analysis has been of major interest in recent times, with an increasing number of applications demanding an online analysis of a data stream. And as one enters the real-time domain, several challenges appear that render most of the current methods infeasible. We will present the FOCuS procedure, a fast online changepoint detection algorithm based on the simple Page-CUSUM sequential likelihood ratio test, and show how it is possible to solve the online changepoint detection problem sequentially through an efficient dynamic programming recursion. The FOCuS procedure outperforms current state-of-the-art algorithms both in terms of efficiency and statistical power. Furthermore, the procedure was extended to allow for more general scenarios, such as the pre-change mean being unknown, or adding robustness to outliers via robust loss functions. We demonstrate FOCuS on the Amazon CPU utilization datasets from the Numenta Anomaly Benchmark, where the aim is to monitor and detect anomalous behaviors in the CPU utilization of various Amazon Cloudwatch instances in real-time.

E0691: **A modified sequential procedure to estimate the number of breaks in trend**
Presenter: **Daisuke Yamazaki**, Kyushu University, Japan

For univariate time series data with structural breaks in the deterministic trend, the number of breaks can be estimated by sequential testing. However, the existing method has two drawbacks. First, the procedure requires a consistent breakpoint estimator in order to apply the sequential tests, but the trend point estimator is inconsistent when the number of breaks is under-specified. This inconsistency results in the low power of the sequential tests. Second, the tests suffer from the non-monotonic power problem. This is because the long-run variance estimator of the error term is inconsistent when the number of breaks is under-specified. To solve these problems, we propose a modified procedure to estimate the number of breaks in trend. First, we develop a new breakpoint estimator that is consistent even when the number of breaks is under-specified. Second, in order to avoid the non-monotonic power problem, we construct a modified long-run variance estimator. Simulation results show that the power of the modified test is much higher than that of the existing test so that the proposed method has good finite sample properties.

EC345 Room R07 CONTRIBUTIONS IN HIGH DIMENSIONAL AND COMPLEX DATA ANALYSIS Chair: Chi Tim Ng

E0606: **A conditional randomization test for generalized additive models with bootstrap methods**
Presenter: **Mehmet Ali Kaygusuz**, Middle East technical university, Turkey
Co-authors: Vilda Purutcuoglu

Multiple testing procedures have received attention in recent years due to the necessity of model selection algorithms for the current high dimensional and massive amount of data. In order to detect the optimal model among alternatives, the knock-off filter method, also called Fixed-X, has been proposed. This approach uses the false discovery rate to control the underlying selection. Then, another version of the knockoff filter, known as Model-X, has been suggested. This approach is based on the conditional randomization test (CRT) under logistic regression and is applicable for dependent data by conducting more flexible testing procedures. Moreover, CRT can be used when the distribution of variables is unknown. On the other hand, it has been shown recently that bootstrapping before model selection improves the accuracy of results. Moreover, the consistent Akaike information criterion with the Fisher information matrix and information complexity criterion gives higher accuracy under the Gaussian graphical model (GGM) and multivariate adaptive regression splines (MARS), two well-known generalized additive models. As a novelty, we include both the knock-off filter and CRT procedures in GGM and MARS models while analyzing real and simulated biological network datasets.

E0683: **On the multiple comparison procedures among mean vectors for high-dimensional data under covariance heterogeneity**
Presenter: **Takahiro Nishiyama**, Senshu University, Japan
Co-authors: Masashi Hyodo, Hiromasa Hayashi

Two typical multivariate multiple comparisons procedures among mean vectors are discussed: pairwise comparisons and comparisons with a control. In traditional multivariate analysis, these multivariate multiple comparisons procedures are constructed based on Hotelling's T^2 statistic in multivariate normal populations. However, in high-dimensional settings, such as when the dimensions exceed total sample sizes, these methods cannot be applied. In such cases, asymptotically conservative simultaneous confidence intervals have been proposed under the assumption of homogeneity of variance-covariance matrices across groups. Unfortunately, these simultaneous confidence intervals are not asymptotically conservative when this assumption is violated. Motivated by this point, we newly obtain asymptotically conservative confidence intervals based on L^2-type statistic without assuming that the variance-covariance matrices are homogeneous across groups. Empirical results indicate that the proposed simultaneous confidence intervals outperform existing procedures.

E0548: **Directional estimation in l_0 constrained regression**
Presenter: **Mathieu Sauvenier**, Universite Catholique de Louvain, Belgium

In high dimensional sparse linear regression, a selection and an estimation of the parameters is studied based on an l_0 constraint on the direction of the vector of parameters. The direction of the parametric vector is a one-dimensional space that is identified through the leading generalized eigenspace of measurable matrices. This new connection with generalized eigenvalue problems also allows us to estimate the direction of the parameter vector in a high-dimensional setting consistently under an l_0 sparsity constraint. In particular, the truncated Rayleigh flow method (also called rifle) is an established method to estimate leading sparse generalized eigenvectors. We show that its use in high-dimensional linear regression leads to a non-linear estimator achieving the minimax rate of convergence for the L_2 loss in the sub-gaussian setting. The conference also shows the asymptotic normality of the estimator with an explicit asymptotic variance. Extensive simulations present situations where the proposed estimator outperforms its direct competitor given by a recent algorithm based on Support Detection And Root finding (so-called SDAR) that approximates the solution of an l_0 penalized least-square program.

E0588: **Outlyingness-oriented causality**
Presenter: **Jerzy Rydlewski**, AGH University of Science and Technology, Poland

Socio-economic datasets often do not fulfil restrictive assumptions of causal analysis procedures proposed in the literature. This proposition indicates certain empirical challenges and conceptual opportunities related to applications of data depth concept procedures into a process of causal inference related to socio-economic phenomena. Statistical functional depths are applied to indicate factual and counterfactual distributions commonly used within procedures of causal inference. Thus, a modification of the Rubin causality concept is proposed, which offers valuable possibilities for conducting robust causal inference in economics, especially in multivariate and functional cases. The concept of depth-based outlyingness applied to causality analysis is applied, as depths enable comparisons of control and treatment groups considered in causal analysis. Finally, an empirical example is shown to illustrate the method.

E0563: **Sparse common and distinctive covariates logistic regression: Classification method for high-dimensional multiblock data**
Presenter: **Soogeun Park**, Tilburg University, Netherlands
Co-authors: Eva Ceulemans, Katrijn Van Deun

Datasets comprised of large sets of variables from multiple sources concerning the same observation units are becoming more widespread today. Constructing a classification model in the context of such high-dimensional and multi-block datasets involves a multitude of challenges: variable selection, classification of the response variable and identification of processes at play underneath the predictors. These processes are of particular interest in the setting of multi-block data because they can either be associated individually with single data blocks or jointly with multiple blocks. Many methods have addressed the classification problem in high-dimensionality for a single block of data. However, the additional challenge of capturing and distinguishing distinctive and joint processes from multi-block data has not received sufficient attention. To this end, we propose Sparse Common and Distinctive Covariates Logistic Regression (SCD-Cov-logR). The method extends principal covariates regression to multi-block settings and combines with generalized linear modeling framework to allow classification of a categorical response while revealing predictive processes that involve single or multiple data blocks. In a simulation study, SCD-Cov-logR resulted in outperformance compared to related methods commonly used in behavioural sciences.

EP001 Room Poster POSTER SESSION Chair: Gil Gonzalez-Rodriguez

E0266: **Short-term forecasting for Korean coastal sea surface temperature and monitoring its levels**
Presenter: **Seunghwan Lee**, Inha University, Korea, South
Co-authors: Younsang Cho, Donghyeon Yu

Sea surface temperature (SST) is one of the main physical characteristics of the ocean and plays an important role to model other physical conditions within the ocean. In general, SST is obtained by either direct or indirect observation; a direct observation is a measurement by buoys and ships equipped with measurement instruments, whereas an indirect observation is derived from remote sensing by merging measurements from satellites. Although remote sensing can cover whole ocean areas, it could contain large measurement errors caused by climate conditions such as clouds or reflected and scattered solar radiation. Especially, observations near the coastal areas by remote sensing usually have large measurement errors. We focus on short-term forecasting for SST of Korean coastal areas measured at tidal observatories. We developed daily and hourly SST forecasting methods based on the autoregressive error model. The averages of root mean squared errors (RMSE) of one-day and 24-hours predictions using the daily and hourly forecasting models for 13 Korean coastal locations are 0.3038 and 0.3723, respectively. In addition, we developed a monitoring procedure based on the prediction of the SST levels, where for each time point in a year, the SST levels are pre-determined based on the estimated normal SST. We predict the levels of the SST for a targeted time using the ensemble of machine-learning algorithms such as random forest and artificial neural network models.

E0460: **Machine learning and big data in econometrics: A machine learning-based specification test**
Presenter: **Gilles Hacheme**, Aix Marseille Univ, CNRS, AMSE, France

Machine Learning (ML) and Big data become ubiquitous in many scientific fields. But their contribution to social sciences is not yet evident. Indeed, the massive flood of data generated by the growing use of the internet is revolutionizing social sciences and particularly Economics. There is an increasing number of research papers in Economics analyzing different web platforms to understand interactions inside those markets better. While the data sphere has been exponentially growing, ML methods have shown their relevance in extracting information from those massive data. The aim is to show the potential benefit of ML techniques for Econometrics clearly. We suggest some way ML can be used in Econometrics for better model specification. Indeed, the increasing popularity of ML models is related to their ability to give better forecasting/prediction results than structural (parametric) models. ML models are known for their ability to capture very complex interactions and non-linearities. The downside is often the poor explainability of their results. Nevertheless, instead of opposing the structural models to the ML ones, we can use the latter as a benchmark to improve the first ones (the structural models) that are far better in terms of explainability. So, we suggest the use of ML to specify better parametric models often used in Economics.

E0678: **Exact inference for an exponential parameter under generalized progressive type II hybrid censored competing risk data**
Presenter: **Subin Cho**, Daegu University, Korea, South
Co-authors: Insang Hwang, Kyeongjun Lee

The experimenter might not always obtain complete information on failure times for all experimental units. In many situations, the removal of units from the experiment is pre-planned and intentional to save time and cost or free up testing facilities for other experiments. Also, due to the complexity of internal structure and external environment, it is common that the failure of a unit results from several causes of failure. These causes of failure are known as the competing risks that compete with each other in the life cycle and can be frequently encountered in various application fields, such as reliability, engineering, and lifetime study. We discuss exact inference for competing risks model with generalized type II progressive hybrid censored exponential data. We derive the conditional moment generating functions of the scale parameters of exponential distribution and the resulting confidence interval under generalized type II progressive hybrid censored competing risks data.

E0679: **Estimation of the Weibull distribution based on generalized adaptive progressive hybrid censored competing risks data**
Presenter: **Kyeongjun Lee**, Daegu University, Korea, South
Co-authors: Yeongjae Seong, Hyein Koo

A competing risks model is considered under a generalized adaptive progressive hybrid censoring. When the latent failure times are Weibull distributed, maximum-likelihood estimates for the unknown distribution parameters are established where the associated existence and uniqueness are shown. The asymptotic distribution of the maximum-likelihood estimators is used to construct approximate confidence intervals. Moreover, Bayes estimates of the unknown parameters are obtained with respect to symmetric loss function and asymmetric loss function under the assumption of independent gamma priors. Lindley's approximation and the MCMC techniques have been utilized for Bayesian calculation. Simulation studies and real-life example are presented for illustration purpose.

E0732: **T-fold sequential-validation technique for out-of-distribution generalization with financial time series data**
Presenter: **Juan Francisco Munoz-Elguezabal**, Western Institue of Technology and Higher Education (ITESO), Mexico
Co-authors: Juan Diego Sanchez

The temporal structure in financial time series (FTS) data demands non-trivial considerations in the use of cross-validation (CV). Such frequently used technique is based on statistical learning theory, which is founded on the assumption that training samples are i.i.d. Although there is progress in studying fundamental phenomenons in certain learning methods such as feature selection imbalance during the learning stage, it is currently widely accepted that there will be no reason to expect good out of sample results from a learning process without such strong assumption. In FTS, there are conditions under which sub-sampling data leads to overshadow the effect of non-deterministic relationships between features and the target variable among different samples. Such effect remains unnoticed given the use of the additivity property in the decomposition of objective functions for the Learning Process. Moreover, it reduces to a particular operation the relationship among samples without information attribution. We present a technique that controls information leakage and decomposes the global probability distribution into local probability distributions, providing identification of each sample contribution to the learning process, maintaining information sparsity, therefore, relaxing the effects of the i.i.d. assumption. Parametric stability, as a result, is presented for exchange rate prediction using different predictive models.

Thursday 24.06.2021 15:30 - 16:45 Parallel Session E – EcoSta2021

EI007 Room R07 RECENT ADVANCEMENTS ON THE ANALYSIS OF PANEL DATA Chair: Tomohiro Ando

E0156: **Estimation of panel group structure models with structural breaks in group memberships and coefficients**
Presenter: **Ryo Okui**, Seoul National University, Korea, South
Co-authors: Wendun Wang, Robin Lumsdaine

The aim is to propose an estimation method for a change in group membership structure and/or the values of coefficients. We consider linear panel data models with a grouped pattern of heterogeneity. In this context, a structural break can arise when the group membership structure and/or the value of one or more coefficients change during the sample period. Considering both possibilities is important since failure to account for a change in the group membership structure may result in detecting a spurious structural break in the values of the coefficients. We propose a least-squares estimator for such models that simultaneously estimate the breakpoint, group membership structure, and coefficients. The estimator is consistent under a mild condition on the relative magnitude of the cross-sectional sample size and time series length. The asymptotic distribution of the coefficient estimator is identical to that under known breakpoint and known group membership structure. Monte Carlo simulations yield encouraging results. An empirical example illustrates the use of the approach and associated inference.

E0191: **Two-step instrumental variable estimation of linear panel data models with interactive effects**
Presenter: **Takashi Yamagata**, University of York & Osaka University, United Kingdom
Co-authors: Guowei Cui, Milda Norkute, Vasilis Sarafidis

Instrumental variable (IV) estimators are proposed for linear panel data models with interactive effects in the error term and regressors. The IVs are transformed regressors, and it is not necessary to search for external IVs. We consider the models with homogeneous slopes and heterogeneous slopes. The approach asymptotically eliminates interactive effects in the error term and the regressors *separately*. Asymptotic properties of the proposed estimators are investigated, which reveal that: (i) the $\sqrt{NT}$-consistent second-step estimator is free from asymptotic bias, which could arise due to the correlation between the regressors and estimation errors of interactive effects; (ii) under the same condition other existing estimators, which asymptotically eliminate interactive effects in the error term *only* or in the regressors and error term *jointly*, can suffer from asymptotic biases, and; (iii) the estimator is asymptotically as efficient as the latter estimator after the bias-correction, but the relative efficiency to the former estimator after the bias-correction is indeterminate.

EO247 Room R01 RECENT DEVELOPMENT IN HIGH-DIMENSIONAL STATISTICS Chair: Lizhu Tao

E0327: **Principal components analysis of correlated functional data**
Presenter: **Julian Austin**, Newcastle University, United Kingdom
Co-authors: Jian Qing Shi, Robin Henderson

A functional principal components analysis is considered where sparsely observed functional data exhibit correlation. We utilise penalised spline regression for estimation of both the mean and covariance surfaces. We account for between-curve correlation by correlating functional principal component scores using a Gaussian process to induce correlation. We consider the case of both stationary and non-stationary parametric covariance kernels for the Gaussian process. We propose a novel estimation procedure for the parametric covariance kernel using typical Gaussian process kernel estimation techniques. The performance of such a model is assessed on simulated sparsely observed functional data generated with various degrees of correlation from both stationary and non-stationary correlation structures. We show the proposed estimation procedure can reconstruct known hyperparameters of the various correlation structures well. We highlight the ability of such a model to identify dominant modes of variation and provide insight into the correlation structure between curves. We compare results to the principal components analysis assuming independent functional data and highlight the additional benefit for relaxing this assumption. Finally, we apply the method to spatio-temporal climate data.

E0399: **Linear discriminant analysis with high dimensional mixed variables**
Presenter: **Binyan Jiang**, The Hong Kong Polytechnic University, Hong Kong

With the rapid development of modern measurement technologies, datasets containing both discrete and continuous variables are more and more commonly seen in different areas. In particular, the dimensions of the discrete and continuous variables can often be very high. Though discriminant analysis for mixed variables under the traditional fixed dimension setting has been well studied since the 80s, promising approaches considering both the high dimensionality and the mixing nature of the data sets are still missing. We aim to develop a simple yet useful classification rule that addresses both the high dimensionality and the simultaneously mixing nature of the variables. Our framework is built on a location model, under which we further propose a semiparametric formulation for the optimal Bayes rule. We show that the optimal classification direction and the intercept in the optimal rule can be estimated separately. Efficient direct estimation schemes are then developed to obtain consistent estimators of the discriminant components. Asymptotic results on the estimation accuracy and the misclassification rates are established, and the competitive performance of the proposed classifier is illustrated by simulation and real data studies.

E0461: **A sparse Ising model with latent variables**
Presenter: **Lizhu Tao**, University of Warwick, United Kingdom
Co-authors: Yiyuan She, Chenlei Leng

A new method termed SIMPLE is proposed to estimate the graph structure of the sparse Ising model with latent variables under the high-dimensional setting. SIMPLE is built on a pseudo-likelihood function with an ℓ_1 penalty to ensure the sparsity of the graph and a nuclear norm to support the estimation of the structure of the latent variables. An efficient iterative block coordinate descent algorithm with algorithmic convergence guarantee is developed to solve the specific sparse plus low-rank optimization problem. Theoretical analysis shows that the global error of our estimators measured by the Bregman divergence achieves satisfactory accuracy. Theoretical guarantees on the consistent recovery of the graph structure are also provided. Extensive simulation studies with a wide range of settings show that our method significantly outperforms its counterparts. Two real-world data applications, including stock prices of oil companies in different markets and exchange rates to the US dollar of various currencies, demonstrate the efficacy of our new method.

EO079 Room R03 NONPARAMETRIC OR DEFAULT BAYESIAN METHODS WITH APPLICATIONS IN MEDICINE Chair: Yisheng Li

E0575: **Bayesian population finding in a randomized clinical trial**
Presenter: **Satoshi Morita**, Kyoto University Graduate School of Medicine, Japan
Co-authors: Peter Mueller, Hiroyasu Abe

A utility-based Bayesian approach is discussed to population finding in the context of randomized clinical trial (RCT). The approach is based on casting the population finding process as a formal decision problem together with a flexible probability model, Bayesian additive regression trees (BART), to summarize observed data. We define a utility function that addresses the competing aims of the desired report so that the decision is constrained to be parsimonious and interpretable. We illustrate the approach with a joint time-to-event and toxicity outcome from an RCT for locally advanced or metastatic breast cancer.

E0521: **Bayesian scalar-on-image regression for automatically detected of regions of interest**
Presenter: **Sara Wade**, University of Edinburgh, United Kingdom

In biomedical studies, vast amounts of imaging, biological and clinical data are increasingly collected to improve understanding of diseases or conditions. In this setting, we develop scalable Bayesian scalar-on-image regression models that allow for the integration of such data. Scalar-on-image regression models utilise the entire imaging data, making it is possible to capture the complex pattern of changes associated with the disease and improve accuracy; however, the massive dimension of the images, which is often in the millions, combined with the relatively small sample size, that at best is usually in the hundreds, pose serious challenges. We propose a novel class of Bayesian nonparametric scalar-on-image regression models based on the Potts-Dirichlet process model that group together voxels into spatially coherent clusters used as features in the regression model. This greatly eases the computational issues associated with the high-dimensional and highly-correlated inputs and allows for interpretable and reliable features that are automatically defined as the most discriminative. The posterior inference is based on a generalized Swendsen-Wang sampler, allowing efficient split-merge moves that take advantage of the spatial structure. Applications focus on early diagnosis and prognosis of Alzheimer's disease, irreversible brain disease and major international public health concern.

E0511: **A uniform shrinkage prior in spatiotemporal Poisson models for count data**
Presenter: **Yisheng Li**, The University of Texas MD Anderson Cancer Center, United States
Co-authors: Krisada Lekdee, Chao Yang, Lily Ingsrisawang

Default Bayesian inference is considered in a Poisson generalized linear mixed model for spatiotemporal data. Normal random effects are used to model the within-area correlation over time, and spatial effects represented with a proper conditional autoregressive model (CAR) are used to model the between-area correlations. We develop a uniform shrinkage prior (USP) for the variance components of the spatiotemporal random effects. We prove that the proposed USP and the resulting posterior are proper under the proposed USP, an independent flat prior for each fixed effect, and a uniform prior for a spatial parameter under suitable conditions. Posterior simulation is implemented, and inference made using the OpenBUGS, R2OpenBUGS and RStan software packages. We illustrate the proposed method by applying it to a leptospirosis count dataset with observations from 17 northern provinces of Thailand across four quarters in 2011 to construct the disease maps. We estimate the 10 highest leptospirosis morbidity rates across provinces and quarters. According to the deviance information criterion, the proposed USP for the variance components of the spatiotemporal effects yields better performance than the conventional inverse gamma priors. A simulation study suggests that the estimated fixed-effect parameters are accurate, based on a relative bias criterion.

EO013 Room R04 RECENT ADVANCES IN BAYESIAN INFERENCE I Chair: Igor Pruenster

E0387: **Post-processed posteriors for structured covariances**
Presenter: **Jaeyong Lee**, Seoul National University, Korea, South

Bayesian inference of structured covariance matrices are considered, and a post-processed posterior is proposed. The post-processing of the posterior consists of two steps. In the first step, posterior samples are obtained from the conjugate inverse-Wishart posterior, which does not satisfy any structural restrictions. In the second step, the posterior samples are transformed to satisfy the structural restriction through a post-processing function. The conceptually straightforward procedure of the post-processed posterior makes its computation efficient and can render interval estimators of any functional of covariance matrices. We also show that it has nearly optimal minimax rates for banded and bandable covariances among all possible pairs of priors and post-processing functions. The advantages of the post-processed posterior are demonstrated by a simulation study and a real data analysis.

E0492: **Gauss versus Laplace rates of contraction under Besov regularity**
Presenter: **Sergios Agapiou**, University of Cyprus, Cyprus
Co-authors: Masoumeh Dashti, Tapio Helin

The purpose is to discuss recent results on contraction rates with a family of priors with tails between Laplace and Gaussian, termed p-exponential priors. We will focus on the white noise model and discuss upper bounds on the rate of contraction under Besov regularity of the truth in L2-loss. We will use alpha-regular priors and see that Laplace priors achieve the same and often better rates than Gaussian ones. In particular, we will see that for spatially inhomogeneous unknown functions, that is, functions that are smooth in some areas but rough in other areas, Gaussian priors appear to be suboptimal. On the other hand, Laplace priors achieve better rates, which can be minimax when the prior is appropriately calibrated.

E0483: **Measuring dependence in the Wasserstein distance for Bayesian nonparametric models**
Presenter: **Marta Catalano**, University of Torino, Italy
Co-authors: Antonio Lijoi, Igor Pruenster

Bayesian nonparametric models are a prominent tool for performing flexible inference with a natural quantification of uncertainty. The main ingredient is discrete random measures, whose law acts as prior distribution for infinite-dimensional parameters in the models and, combined with the data, provides their posterior distribution. Recent works use dependent random measures to perform simultaneous inference across multiple samples. The borrowing of strength across different samples is regulated by the dependence structure of the random measures, with complete dependence corresponding to the maximal share of information and fully exchangeable observations. For a substantial prior elicitation, it is crucial to quantify the dependence in terms of the hyperparameters of the models. State-of-the-art methods partially achieve this through the expression of the pairwise linear correlation. We propose the first non-linear measure of dependence for random measures. Starting from the two samples case, dependence is characterized in terms of distance from exchangeability through a suitable metric on vectors of random measures based on the Wasserstein distance. This intuitive definition extends naturally to an arbitrary number of samples, and it is analytically tractable on noteworthy models in the literature.

EO257 Room R05 MODELLING ECONOMIC AND FINANCIAL TIME SERIES: ESTIMATION AND FORECASTING Chair: Seok Young Hong

E0303: **Stock return prediction: Stacking a variety of models**
Presenter: **Tingting Cheng**, Nankai University, China
Co-authors: Bo Zhao

An ensembling machine learning approach, Stacking, is employed to refine and combine a variety of linear and nonlinear individual stock return prediction models. In an application of forecasting U.S. market excess return, Stacking can outperform the traditional historical mean benchmark in terms of both in-sample and out-of-sample performances. Moreover, Stacking performs better than a simple combination forecast and the C-Enet forecast in terms of out-of-sample forecasting consistently over time. More importantly, we find that the out-of-sample gains of Stacking are especially evident during extreme downside market movements. Overall, Stacking can generate substantive improvements in out-of-sample market excess return predictability.

E0320: **Cojump anchoring**
Presenter: **Wenying Yao**, Deakin University, Australia
Co-authors: Lars Winkelmann

A two-step inference procedure is developed to test for a local one-for-one relation of contemporaneous jumps in high-frequency financial data corrupted by market microstructure noise. The first step develops a new bivariate Lee-Mykland jump test for pre-averaged, intra-day returns. If a jump is detected in at least one of the two assets, then the second step tests for equal jump sizes. We apply the test procedure to pairs of nominal and inflation-indexed government bond yields at monetary policy announcements in the U.S., U.K., and Euro Area. The analysis provides new high-frequency evidence about the anchoring of inflation expectations and central banks' ability to push a measure of inflation expectations towards their inflation target.

E0350: **Separate noise and jumps from tick data: An endogenous thresholding approach**
Presenter: **Seok Young Hong**, Lancaster University Management School, United Kingdom
Co-authors: Xiaolu Zhao, Oliver Linton

The problem of jump detection for ultra-high-frequency tick-by-tick data is studied. We propose a novel easy-to-implement procedure that can separate the contribution of microstructure noise and finite activity price jumps from the price process, which may have interesting implications on asset pricing and forecasting problems. We provide theoretical grounds for our approach and suggests practical guidelines for determining the tuning parameter. Making a comparison with the star performers in a recent comprehensive review for jump detection methods as well as a test on tick data, we show that the method performs admirably well via extensive simulations and rich empirical illustrations.

EO307 Room R06 BAYESIAN ANALYSIS OF BUSINESS ANALYTICS PROBLEMS Chair: Mike So

E0592: **Multivariate randomized response for binary and ordinal data**
Presenter: **Yasuhiro Omori**, University of Tokyo, Japan
Co-authors: Amanda Chu, Mike So, Hing-yu So

To identify and reduce various sources of serious drug administration errors in the hospitals, we recently conducted comprehensive online surveys on nurses in Hong Kong. The data consists of multiple binary and ordinal categorical responses to the sensitive questions concerning the administration errors for each nurse. To protect their data privacy from answering these sensitive questions honestly, the randomized response technique is applied when collecting these data. To investigate various effects of the covariates for the nurse, the joint modeling of the multiple binary and ordinal data is proposed under the restriction of randomizing the categorical responses to the sensitive questions. We develop the efficient Markov chain Monte Carlo estimation method of the proposed model using latent variables in the framework of the multivariate probit model. Our empirical study reveals the sources of drug administration errors and shows the appropriate direction for drug administration policies in the future.

E0392: **High-frequency realized stochastic volatility model**
Presenter: **Toshiaki Watanabe**, Hitotsubashi University, Japan
Co-authors: Jouchi Nakajima

A new high-frequency realized stochastic volatility model is proposed. Apart from the standard daily-frequency stochastic volatility model, the high-frequency stochastic volatility model is fit to intraday returns by extensively incorporating intraday volatility patterns. The daily realized volatility calculated using intraday returns is incorporated into the high-frequency stochastic volatility model by considering the bias in the daily realized volatility caused by microstructure noise. The volatility of intraday returns is assumed to consist of the autoregressive process, the seasonal component of the intraday volatility pattern, and the announcement component responding to macroeconomic announcements. A Bayesian method via Markov chain Monte Carlo is developed for the analysis of the proposed model. The empirical analysis using the 5-minute returns of E-mini S&P 500 futures provides evidence that our high-frequency realized stochastic volatility model improves in-sample model fit and volatility forecasting over the existing models.

E0600: **Structural learning in Bayesian networks and business applications**
Presenter: **Shun Hin Chan**, The Hong Kong University of Science and Technology, Hong Kong

A Bayesian network is a probabilistic graphical model that aims to model conditional dependence (causation) among variables. In most cases, the true underlying structure of a set of variables is unknown. The number of possible structures grows explosively when we have more variables in the networks. Recovering the structure from a given data set is challenging. Many existing algorithms can learn structures from data automatically, but those algorithms may not be able to handle networks with hundreds or thousands of nodes. The focus is a score-based approach. However, sampling the partial order from a graph using MCMC methods requires intensive computational time, and the sampling scheme is biased. We develop a new MCMC sampling scheme for structural learning. The proposed sampling scheme can potentially handle the problem of biasedness in existing methods. We demonstrate in simulations that our proposed method is more efficient to allow the graph samples to enter high score areas. We also illustrate our sampling scheme in a social science research study.

EO185 Room R08 ECONOMETRICS WITH APPLICATIONS TO CRYPTOCURRENCIES AND THE BLOCKCHAIN Chair: Jeffrey Chu

E0537: **Cryptocurrency market activity during extremely volatile periods**
Presenter: **Paraskevi Katsiampa**, The University of Sheffield, United Kingdom
Co-authors: Konstantinos Gkillas, Francois Longin

The tail dependence structure between price returns and trading volumes is studied for eight cryptocurrencies using bivariate extreme value theory. We show that for all cryptocurrencies considered, the extreme correlation between return and volume decreases as we move towards the distribution tails. We also find evidence of asymmetry in the return-volume relationship as the correlation between return and volume is significantly different for positive and negative return exceedances across all cryptocurrencies. We interpret the results in light of different economic models: misinterpretation of trades, market overreaction, market information, market manipulation, price bubble, imperfect information and multiple equilibria. A better understanding of the activity of market participants during extremely volatile periods in cryptocurrency markets is provided.

E0712: **Bitcoin versus high-performance technology stocks in diversifying against global stock market indices**
Presenter: **Stephen Chan**, American University of Sharjah, United Arab Emirates
Co-authors: Jeffrey Chu, Yuanyuan Zhang

Bitcoin is investigated in finance from the perspective of technology. By comparing the relationship between the returns of Bitcoin and high-performance technology stocks, and a range of global stock markets, we investigate whether there is any evidence of potential hedging and diversification properties and whether these are conditional on the states of the respective markets. We implement a quantile-on-quantile regression method to examine the relationship between both Bitcoin returns and technology stock returns and stock market returns at varying quantiles. The results show that although Bitcoin and high-performance technology stocks arguably share many similarities, it is clear that Bitcoin exhibits significant differences compared with high-performance technology stocks in terms of the diversification properties against global stock markets. From a financial perspective, this may suggest that individuals do not see and treat Bitcoin as a technology (or a technology company) but rather further supports the view of Bitcoin as a potential investment for financial gain (given its diversification properties).

E0722: **The maturity of crypto markets: A market efficiency test based on trading strategies**
Presenter: **Jeffrey Chu**, Renmin University of China, China

A rigorous research design is proposed to test trading strategies and apply it to the crypto market. The research design relies on three pillars: 1) Fundamental tests to gauge the information content in the trading strategies; 2) Predictability tests of trading strategies adjusted with risk; and 3) Profitability tests of trading strategies considering explicit and implicit transaction costs). We focus on technical analysis based on price patterns. Applied to the crypto market (Bitcoin, Ethereum and Litecoin), we find that such trading strategies do not convey relevant information and extra return. We obtain surprising results.

EG330 Room R02 CONTRIBUTIONS TO STATISTICAL MODELLING Chair: Gilles Stupfler

E0659: **The average conditional and partial Kendall's tau**
Presenter: **Margot Matterne**, KU Leuven, Belgium
Co-authors: Irene Gijbels

For investigating how other random variables influence the dependency between random variables, the concept of conditional association measures is useful. They measure the strength of conditional dependencies given a covariate. When studying conditional dependencies, some averaging over this covariate may be needed. Examples of average measures that can serve in this context are the average conditional and the partial Kendall's tau. It is known that these measures differ in general. This raises the question of how different they can be, an important question when using these concepts in statistical analysis. The aim is to provide a quantitative study of the possible differences between these two average measures and establish sufficient conditions under which they coincide.

E0662: **Information-criteria-based model selection for neural networks**
Presenter: **Andrew McInerney**, University of Limerick, Ireland
Co-authors: Kevin Burke

Neural network model selection is usually carried out using a 'trial-and-error' approach, with varying initial weights and network architecture. However, calculating an associated likelihood function opens the door to the information-criteria-based model and variable selection and likelihood-based confidence intervals for network weights. Novel 'bottom-up' and 'top-down' model selection methods are proposed using the Bayesian information criterion for feedforward multi-layer perceptions, whereby the optimal weights for one model are carried over to the next. Compared to the standard trial-and-error search through the space of models, this is both more computationally efficient and has an increased probability of recovering the true model. Simulation studies are used to evaluate the performance of the proposed methods and an application on real data is investigated.

E0690: **Smooth BIC variable selection procedure for heteroscedastic data**
Presenter: **Meadhbh ONeill**, University of Limerick, Ireland
Co-authors: James Gleeson, Kevin Burke

Modern variable selection procedures revolve around penalization methods to execute simultaneous model selection and estimation. A popular method is the lasso (least absolute shrinkage and selection operator), which contains a tuning parameter. This parameter is typically tuned by minimizing the cross-validation error or Bayesian information criterion (BIC), but this can be computationally intensive as it involves fitting an array of different models and selecting the best one. However, we have developed a procedure based on the so-called “smooth BIC" in which the tuning parameter is automatically selected in one step. We also extend this model selection procedure to the so-called “multi-parameter regression" framework more flexible than classical regression modelling. Multi-parameter regression introduces flexibility by taking account of the effect of covariates through multiple distributional parameters simultaneously, e.g., mean and variance. These models are useful in normal linear regression when the process under study exhibits heteroscedastic behavior. Reformulating the multi-parameter regression estimation problem in terms of penalized likelihood enables us to take advantage of the close relationship between model selection criteria and penalization. Utilizing the smooth BIC is extremely computationally advantageous as it obviates the issue of choosing multiple tuning parameters.

Thursday 24.06.2021 16:55 - 18:35 Parallel Session F – EcoSta2021

EI005 Room R06 ADVANCES IN FINANCIAL TECHNOLOGY AND RISK ANALYTICS Chair: Mike So

E0467: **Financing long-term care risks in a super-ageing society: A discrete choice experiment in Hong Kong**
Presenter: **Wai-Sum Chan**, The Chinese University of Hong Kong, Hong Kong

There is a stark contrast between rising long-term care (LTC) demands and limited financing capacity in many ageing societies. Despite the theoretical potential of private insurance in LTC financing reforms, the reality is that the market remains remarkably underdeveloped. This study adopts a novel two-phase approach to quantitatively examine the market demand for private long-term care insurance (LTCI) in Hong Kong, one of the worlds super-ageing societies. In order to examine peoples preferences regarding private LTCI in Hong Kong, which has been exploring alternative LTC financing mechanisms to relieve the overburdened public system, we conducted a discrete choice experiment (DCE) in 2019 to elicit the preferences of a representative sample of 410 middle-aged adults.

E0596: **Sparse vector error correction models with application to cointegration-based trading**
Presenter: **Philip Yu**, The Education University of Hong Kong, Hong Kong
Co-authors: Renjie Lu, Xiaohang Wang

Inspired by constructing large-size cointegrated portfolios, a vector error correction model is considered. The adaptive Lasso estimator of the cointegrating vectors is developed. The asymptotic properties of the estimators and the oracle property of the adaptive Lasso are derived. An optimization algorithm for estimating the model parameters is proposed. The simulation study shows the effectiveness of the parameter estimation procedures and the forecasting performance of our model. In the empirical study, we apply the proposed method to construct the sparse cointegrated portfolios with or without market-neutral property. The trading performances of different types of cointegrated portfolios are evaluated using the Dow Jones Industrial Average composite stocks. The empirical findings reveal that the sparse cointegrated market-neutral portfolios of a number of securities can benefit the investors who wish to construct statistical arbitrage portfolios that are market-neutral.

E0602: **Optimal liquidation with hidden orders under self-exciting market order dynamics**
Presenter: **Ge Zhang**, National University of Singapore, Singapore
Co-authors: Ying Chen, Zexin Wang, Chao Zhou

An optimal execution strategy with both hidden and limit orders in a continuous-time framework is developed where 1) market order arrivals are either Poisson process or Hawkes process, 2) limit order submissions is subject to the exposure cost, and 3) executions of limit and hidden order both incur the immediate execution cost. Under the homogeneous Poisson process, we derive a quasi-closed-form solution containing a switching time, at which the agent changes from a pure-hidden-order phase to a mixed-orders phase until termination. Under the Hawkes process with self-exciting orders, a numerical solution is derived with feedback controls. We show there is a similar two-phase strategy, except that the switching time becomes a function of the market order intensity. The theoretical model also suggests the different impact of time pressure on order size under the two phases. In particular, the hidden order size increases with the time pressure under the pure-hidden-order phase and decreases under the mixed-orders phase, while the sum of limit and hidden orders always increases. Given transaction-level data of 100 NASDAQ stocks, real data analysis shows that the Hawkes mixture strategy outperforms the pure limit order strategy with about 70% cost reduction and the Poisson-based mixture strategy with 27% cost reduction.

EO259 Room R01 TIME SERIES EXTREMES: MODELING AND FORECASTING Chair: Raphael Huser

E0321: **Copula models for time series extremes**
Presenter: **Pavel Krupskiy**, Melbourne University, Australia
Co-authors: Marc Genton, Raphael Huser

Two approaches for building flexible models for time series extremes are considered. The first approach uses Cauchy convolution processes with different kernel functions, and the second approach assumes that there exists a common factor that affects all the realizations of the time series process. We consider some interesting special cases and discuss inference methods for these two classes of models. We apply these models to analyze a wind data set.

E0359: **Asymmetric tail dependence modeling, with application to cryptocurrency market data**
Presenter: **Yan Gong**, KAUST, Saudi Arabia
Co-authors: Raphael Huser

Since the inception of Bitcoin in 2008, cryptocurrencies have played an increasing role in the world of e-commerce. Still, the recent turbulence in the cryptocurrency market in 2018 has raised some concerns about their stability and associated risks. For investors, it is crucial to uncover the dependence relationships between cryptocurrencies for more resilient portfolio diversification. Moreover, the stochastic behavior in both tails is important. In order to assess both risk types, we develop a flexible copula model which can distinctively capture asymptotic dependence or independence in its lower and upper tails. The proposed model is parsimonious and smoothly bridges (in each tail) both extremal dependence classes in the interior of the parameter space. The inference is performed using a full or censored likelihood approach. We also develop a local likelihood approach to capture the temporal dynamics of extremal dependences among five leading cryptocurrencies. The results of Bitcoin and Ethereum show that our proposed copula model outperforms alternative copula models and that the lower tail dependence level has become stronger over time, smoothly transitioning from an asymptotic independence regime to an asymptotic dependence regime in recent years, whilst the upper tail has been more stable. A full picture of the tail dependence structures between all pairs of cryptocurrencies would provide valuable information to investigators for risk mitigation.

E0432: **Bayesian modeling of time-varying extremal dependence in international stock markets**
Presenter: **Junho Lee**, University of Edinburgh, United Kingdom
Co-authors: Miguel de Carvalho, Antonio Rua

A Bayesian time-varying model is proposed to capture the dynamics of extreme joint losses in international stock markets over the last thirty years. The model relies on dual nonparametric time-varying extremal dependence measures, which can be used to assess the strength of dependence of extreme joint losses over time under the settings of asymptotic dependence and asymptotic independence. These measures are approximate by generalized additive models with a large threshold. The dynamics underlying the extremal dependence structure are tracked using Bayesian smoothing methods based on penalized B-splines. A simulation study is presented to assess the performance of the proposed methods. We analyse five international stock market indices and reveal complex extremal dependence behaviours among the indices, suggesting evidence for smooth transitions between regimes of asymptotic dependence and asymptotic independence.

E0490: **A bias-reduced approach for dynamic estimation of extreme risk measures in financial time series**
Presenter: **Hibiki Kaibuchi**, SOKENDAI The Graduate University of Advanced Studies, Japan
Co-authors: Gilles Stupfler, Yoshinori Kawasaki

The question of estimating risk measures at extreme levels is important in financial applications, both from operational and regulatory perspectives. We estimate alternative risk measures to the most widely-used Value-at-Risk that are extreme expectile, both expectile- and quantile-based forms of the expected shortfall in a time dynamic setting. This is because replacing quantiles with their least square analogues, called expectiles, has recently received increasing attention. For dynamic estimations of such risk measures, we: (i) filter the financial returns using an AR(1)-GARCH(1,1) model; (ii) apply an asymptotically bias-reduced estimator of extreme quantiles to the standardized residuals after filtering; (iii) use an asymptotic relationship between quantile and expectile (or expected shortfall). The results are illustrated on a financial real dataset.

EO021 Room R02 MULTIVARIATE DISTRIBUTIONS: ASYMMETRY, QUANTILES AND MORE Chair: Anneleen Verhasselt

E0643: **Inference for copulas with two-piece margins**
Presenter: **Jonas Baillien**, KU Leuven, Belgium
Co-authors: Irene Gijbels, Anneleen Verhasselt

Copulas provide a versatile tool in the modelling of multivariate distributions. With increased awareness of possible asymmetry in data, skewed copulas combined with classical margins have been employed to model these data appropriately. The reverse, skewed margins with a (classical) copula has also been considered, but mainly with skew-symmetrical margins. We focus on a different type of skewed margins, namely the two-piece distributions. More specifically, we use the recently proposed quantile-based asymmetric family of distributions in given copula structures. For this combination, we provide statistical inference results in consistency and asymptotic normality for the Inference Functions for Margins estimator. A simulation study complements the theoretical results, and the practical usefulness is shown through some real data examples.

E0633: **Optimal tests for elliptical symmetry: Specified and unspecified location**
Presenter: **Christophe Ley**, Ghent University, Belgium
Co-authors: Marc Hallin, Sladana Babic, Laetitia Gelbgras

Although the assumption of elliptical symmetry is quite common in multivariate analysis and widespread in a number of applications, the problem of testing the null hypothesis of ellipticity so far has not been addressed in a fully satisfactory way. Most of the literature in the area indeed addresses the null hypothesis of elliptical symmetry with specified location and actually addresses location rather than non-elliptical alternatives. We are proposing new classes of testing procedures, both for specified and unspecified location. The backbone of our construction is Le Cam's asymptotic theory of statistical experiments, and optimality is to be understood locally and asymptotically within the family of generalized skew-elliptical distributions. The tests we are proposing are meeting all the desired properties of a good test of elliptical symmetry: they have a simple asymptotic distribution under the entire null hypothesis of elliptical symmetry with unspecified radial density and shape parameter; they are affine-invariant, computationally fast, intuitively understandable, and not too demanding in terms of moments. While achieving optimality against generalized skew-elliptical alternatives, they remain quite powerful under a much broader class of non-elliptical distributions and significantly outperform the available competitors.

E0644: **Partially linear expectile regression using local polynomial fitting**
Presenter: **Cecile Adam**, KU Leuven, Belgium
Co-authors: Irene Gijbels

Among the main interests in regression analysis is to explore the influence that covariates have on a variable of interest, the response. There is extensive literature on flexible mean regression, in which the targeted quantity is the conditional mean of the response given the covariates. Quantile regression is another method that aims at estimating the conditional median or other quantiles of the response variable given the covariates. An alternative to quantiles is expectiles. Expectile regression estimates the conditional expectiles of the response variable given realized values of the predictor variables. After a brief introduction to expectiles and univariate nonparametric expectile regression, we discuss multivariate partially linear expectile regression. We will present the statistical methodology and discuss the issue of bandwidth choice. The finite-sample performance of the estimators is investigated in a simulation study, and the methodology is illustrated on real data.

E0632: **Quantile regression for longitudinal data via the multivariate generalized hyperbolic distribution**
Presenter: **Anneleen Verhasselt**, Hasselt University, Belgium

While extensive research has been devoted to univariate quantile regression, this is considerably less the case for the multivariate (longitudinal) version. Quantile functions are easier to interpret for a population of curves than mean functions of non-linear curves. While the connection between (multivariate) quantiles and the (multivariate) asymmetric Laplace distribution is known, it is less well known that its use for maximum likelihood estimation poses mathematical and computational challenges. Therefore, we study a broader family of multivariate generalized hyperbolic distributions, of which the multivariate asymmetric Laplace distribution is a limiting case. We offer an asymptotic treatment. Simulations and a data example supplement the modeling and theoretical considerations.

EO279 Room R03 INNOVATIVE SURVIVAL MODELLING APPROACHES FOR VARIOUS TIME-TO-EVENT DATA Chair: Il Do Ha

E0252: **Semi-parametric multi-parameter regression survival modelling**
Presenter: **Kevin Burke**, University of Limerick, Ireland

A log-linear model for survival data is considered, where both the location and scale parameters depend on covariates (i.e., a "Multi-Parameter Regression" [MPR] approach), and the baseline hazard function is completely unspecified. This model provides the flexibility needed to capture many interesting features of survival data at a relatively low cost in model complexity. Estimation procedures are developed, and asymptotic properties of the resulting estimators are derived using empirical process theory. Finally, a resampling procedure is developed to estimate the limiting variances of the estimators. A practical application to lung cancer data is illustrated.

E0284: **Quantile regression analysis of survival data with covariates subject to detection limits**
Presenter: **Liming Xiang**, Nanyang Technological University, Singapore

With advances in biomedical research, biomarkers are becoming increasingly important prognostic factors for predicting overall survival. In contrast, the measurement of biomarkers is often censored due to instruments' lower limits of detection. This leads to two types of censoring: random censoring in overall survival outcomes and fixed censoring in biomarker covariates, posing new statistical modelling and inference challenges. We propose quantile regression to analyse survival data with covariates subject to detection limits (DL). The proposed method provides a versatile tool for modeling the distribution of survival outcomes by allowing covariate effects to vary across conditional quantiles of overall survival. To estimate the quantile process of regression coefficients, we develop a novel multiple imputation approach based on quantile regression for covariates under DL, avoiding the stringent parametric restrictions on the censored covariates as often assumed in the existing works. We show that the estimation procedure yields uniformly consistent and asymptotically normal estimators under regularity conditions. Simulation results demonstrate the satisfactory finite-sample performance of the method. The proposed method is applied to the motivating data from a study of genetic and inflammatory markers of Sepsis.

E0396: **Semi-parametric copula survival models for clustered time-to-event data**
Presenter: **Sookhee Kwon**, Pukyoung national university, Korea, South

Copula models have been widely used for analyzing clustered or correlated survival data. However, their modelling approaches are mainly based on two-stage estimation procedures. We propose a semi-parametric Archimedian-copula modelling approach using a one-stage likelihood procedure. Here, marginal baseline hazards are non-parametrically modeled based on a cubic M-spline. The proposed method is demonstrated via a simulation study. The new method is illustrated using a well-known clinical data set, comparing two-stage estimation methods.

E0494: **A semiparametric AFT random-effect model**
Presenter: **Byungtae Seo**, Sungkyunkwan University, Korea, South
Co-authors: Il Do Ha

The Accelerated Failure Time (AFT) model with random effects, a useful alternative to the frailty model, has been widely used for analyzing clustered or correlated time-to-event data. In the AFT model, the distribution of the unobserved random effect is usually assumed to be parametric. However, the resulting regression coefficient estimates could be sensitive against misspecification of the random-effect distribution, especially when there is heavy censoring. We propose a semiparametric AFT random-effect model in which the random effect distribution is completely unspecified. The proposed method is demonstrated using simulation studies and practical data examples.

EO015 Room R04 RECENT ADVANCES IN BAYESIAN INFERENCE II Chair: Jaeyong Lee

E0437: **Beta-binomial stick-breaking non-parametric prior**
Presenter: **Ramses Mena**, UNAM, Mexico

A new class of nonparametric prior distributions, termed the Beta-Binomial stick-breaking process, is proposed. An appealing discrete random probability measure arises by allowing the underlying length random variables to be dependent through a Beta marginals Markov chain. The chain's dependence parameter controls the ordering of the stick-breaking weights and thus tunes the model's label-switching ability. Also, the resulting class contains the Dirichlet process and the Geometric process priors as particular cases, which is of interest for MCMC implementations by tuning this parameter. Some model properties are discussed, and a density estimation algorithm is proposed and tested with simulated datasets.

E0531: **On the non-i.i.d. misspecified Bernstein-Von Mises theorem**
Presenter: **Geerten Koers**, Leiden University, Netherlands

The asymptotic behaviour of misspecified posterior distributions is considered in a non-i.i.d. parametric setting. It is shown that a misspecified Bernstein-Von Mises theorem holds, and conditions on the distribution of the data and the likelihood functions of the model are relaxed compared to earlier results. The asymptotic behaviour of the well-specified posterior distribution is compared to that of the misspecified posterior distribution in a non-Gaussian model approximated by Gaussian likelihoods. Under regularity conditions, the misspecified posterior distribution will concentrate on the true parameter in these models. Natural examples in PDE-theory of models that were not covered by existing literature are analysed. The numerical analysis shows that the misspecified posterior distribution has an incorrect uncertainty quantification. It is observed that the resulting credible sets over-cover compared to the credible sets coming from the well-specified posterior distribution.

E0487: **Compositions of discrete random probabilities for inference on multiple samples**
Presenter: **Giovanni Rebaudo**, University of Texas at Austin, United States
Co-authors: Augusto Fasano, Antonio Lijoi, Igor Pruenster

Bayesian hierarchical models have proved to be an effective tool when observations are from different populations or studies since they naturally allow borrowing information across groups while allowing the subjects in the same group to share the same unknown distribution. We consider models induced by compositions of discrete random probabilities measures, most notably Pitman-Yor processes. Such compositions are well suited to account for both clustering of populations and clustering of observations. We identify an analytical expression of the induced random partition distribution, which allows us to gain a deeper insight into the theoretical properties of the model while deriving predictive distributions and urn schemes. The proposed models can be used as a building block for addressing density estimation problems, prediction with species sampling data and testing of distributional homogeneity. The theoretical results further lead us to devise novel MCMC sampling schemes whose effectiveness will be discussed through illustrative examples involving simulated and real data.

E0515: **Nonparametric priors with full-range borrowing of information**
Presenter: **Beatrice Franzolini**, Bocconi University, Italy
Co-authors: Filippo Ascolani, Antonio Lijoi, Igor Pruenster

When data are grouped into distinct samples, they typically are homogeneous within and heterogeneous across samples. In this case, the Bayesian paradigm requires a prior law over a collection of related unknown distributions. Such law characterizes the dependence among observations and thus controls how borrowing information across samples is performed. However, existing nonparametric priors can induce only non-negative correlation across samples, which may not always be appropriate. We propose a novel class of dependent nonparametric priors, which may induce either positive or negative correlation across samples based on the value of a hyperparameter. The proposal fills a gap in the literature of partially exchangeable models and introduces a new and more flexible idea of borrowing information across samples. Moreover, many of the models in the literature can be obtained as specific cases of the one proposed. We investigate prior and posterior theoretical properties of the model and develop algorithms to perform posterior inference. The merits of the proposal are further discussed through illustrative examples on simulated and real data, where our model outperforms competing ones.

EO245 Room R07 TOPICS IN SPATIAL STATISTICS Chair: Anastassia Baxevani

E0686: **Necessary and sufficient conditions for asymptotically optimal kriging of random fields on compact metric spaces**
Presenter: **David Bolin**, King Abdullah University of Science and Technology (KAUST), Saudi Arabia

Optimal linear prediction (also known as kriging) of a random field Z on a compact metric space can be obtained if the mean and covariance functions of Z are known. We consider the problem of predicting the value of Z at some location in the domain based on an increasing number of observations that accumulate at all locations in the domain, and more generally of predicting a linear functional $\phi(Z)$ of the random field (such as an integral) based on observations $\phi_1(Z), ..., \phi_n(Z)$. The main result characterizes the asymptotic performance of linear predictors (as the number of observations increases) based on an incorrect second-order structure, without any restrictive assumptions such as stationarity. For the first time, we provide necessary and sufficient conditions on the misspecified mean and covariance functions for asymptotic optimality of the corresponding linear predictor holding uniformly with respect to all possible linear functionals phi. We illustrate the general results for weakly stationary random fields on subsets of R^d, for isotropic random fields on the sphere, and non-stationary generalized Whittle-Matern random fields obtained as solutions to stochastic partial differential equations.

E0714: **Compound Poisson gamma process as a model for precipitation**
Presenter: **Anastassia Baxevani**, University of Cyprus, Cyprus
Co-authors: Christos Andreou

Many statistical models exist for modelling precipitation. The main challenge when modelling precipitation is that one needs to model both the presence of exact zeros that correspond to dry days and the amounts of precipitation on the wet days. The compound Poisson gamma distribution has the ability to model the exact zeros and the amounts of precipitation simultaneously. We present a methodology to construct a patio-temporal process with compound Poisson gamma marginals as a model for precipitation.

E0739: **Random fields with truncated polynomial spectral density**
Presenter: **Dionissios Hristopulos**, Technical University of Crete, Greece

Spatial random fields, with a specific spectral density given by truncated polynomials (TPSPD), are proposed. We show that the covariance functions for such random fields are given by combinations of Bessel and Lommel functions. We discuss the connection between random fields with truncated polynomial spectral density and continuous moving average processes and generalized random fields and the role of the spectral cutoff. The TPSPD random fields are shown to be mean-square differentiable for all orders by their finite spectral moments. We present simulated realizations of TPSPD random fields and discuss potential applications.

E0742: **Non-linear and joint models with the inlabru package**
Presenter: **Finn Lindgren**, University of Edinburgh, United Kingdom

The Integrated Nested Laplace Approximation (INLA) method was developed to handle latent Gaussian additive regression models. Combined with the stochastic partial differential equation method for constructing computationally efficient representations of Gaussian random fields, this has enabled fast Bayesian analysis of a wide range of models. The inlabru package extends this to a more general model class that allows more non-linearity, and a more user-friendly interface for specifying complex models, such as point process models and joint models for multiple response variables and spatial covariates. By using an iterated INLA approach, the computational power of the R-INLA implementation is extended to a wider range of models.

EC335 Room R05 CONTRIBUTIONS IN METHODS AND COMPUTATIONS Chair: Stefan Van Aelst

E0704: **Canonical correlation analysis for multimodal labeled data**
Presenter: **Mitsuhiro Hashiguchi**, Doshisha University, Japan
Co-authors: Masaaki Okabe, Hiroshi Yadohisa

Canonical correlation analysis (CCA) is a method of dimensionality reduction for two multivariate data. It is effective in reflecting the characteristics of each class in a low-dimensional space when one of the data represents a class label. On the other hand, CCA cannot reflect the structures of multimodal data in a low-dimensional space. The multimodal data has multiple unknown cluster structures within a class. This problem arises because CCA cannot take the local structure of the data into account. The entropic-regularized Wasserstein distance can capture the local structure. This distance corresponds to the squared Euclidean distance with a weighting. Therefore, to reflect the multiple structures within a class in multimodal data in a low-dimensional space, we incorporate the Wasserstein distance into CCA. Simulations and real data examples demonstrate the effectiveness of this method for data with multimodality in classes. Results suggest that the proposed method can reflect characteristics of multiclass data with multimodal classes in a low-dimensional space.

E0660: **Outlier testing in robust 2SLS models: An asymptotic study of the false outlier detection rate**
Presenter: **Jonas Kai Kurle**, University of Oxford, United Kingdom
Co-authors: Xiyu Jiao

A frequent concern in applied economics is that key empirical findings may be driven by a tiny set of outliers. To perform outlier robustness checks in instrumental variables regressions, the common practice is first to run ordinary two-stage least-squares (2SLS) and classify observations with residuals beyond a chosen cut-off value as outliers. Subsequently, 2SLS is re-calculated based on the non-outlying observations, and this procedure may be iterated until robust results are obtained. However, the above trimmed 2SLS has a positive probability of finding outliers even when the data generating process contains none. To answer the question of whether observations are correctly classified as outliers, this paper studies the concept of false outlier detection rate (FODR) asymptotically using empirical processes techniques. The established asymptotic theory of the FODR serves as the basis for the future construction of tests for the overall presence of outliers. Moreover, the asymptotic theory provides a guide for setting the cut-off value.

E0647: **Optimal moment-subset selection for the simulated method of moments using machine learning**
Presenter: **Jiri Kukacka**, UTIA AV CR, v.v.i., Czech Republic

Simulation-based estimation inference is expanded via machine learning techniques. The setup of the simulated method of moments (SMM) is extended with an automated selection of the optimal set of moments. To briefly demonstrate the importance of the issue, a relatively rich set of nine moments has been previously employed, while for a similar type of a financial agent-based model, four but also fifteen moments have been used. An insufficient set of moments will likely ignore some important dynamic properties of the model while overfilling the moment set will likely lead to estimation inefficiencies and problems with identifying parameters. Algorithmic subset selection methods generally developed for model construction are thus utilized. The methods evaluate subsets of features, moments in our case, in terms of their suitability for a given purpose and retain only the optimal ones. We conduct an extensive comparative study of the accuracy and computational complexity of the proposed machine learning extension of the SMM w.r.t. the original method and the simulated maximum likelihood method. As a laboratory, we take advantage of the New Keynesian macroeconomic model under rational expectations and various behavioral heuristics.

E0754: **Combining dimensionality reduction with neural networks for realized volatility forecasting**
Presenter: **Andrea Bucci**, universita degli studi g d annunzio di chieti pescara, Italy
Co-authors: Zhi Liu, Lidan He

The application of artificial neural networks to finance has received a great deal of attention from both investors and researchers, especially as a forecasting method. When the number of predictors is high, these methods suffer from the so-called "curse of dimensionality" and produce biased forecasts. We relied on dimensionality reduction methods to alleviate such issue when a wide set of financial and macroeconomic variables is considered in the prediction of stock market volatility. Specifically, we combined Bayesian Model Averaging (BMA), Principal Component Analysis (PCA), Non-negative Matrix Factorization (NMF) and Least Absolute Shrinkage and Selection Operator (LASSO) with hybrid artificial neural networks to forecast realized volatility. The results showed that reduced models could perform similarly or even outperform the full models in predictive accuracy.

Friday 25.06.2021 08:00 - 09:40 Parallel Session G – EcoSta2021

EO209 Room R01 STATISTICAL METHODS FOR HANDLING HIGH-DIMENSIONAL/LONGITUDINAL DATA Chair: MinJae Lee

E0553: **Estimation of particulate levels using deep dehazing network and temporal prior**
Presenter: **SungHwan Kim**, Konkuk University, Korea, South

Particulate matters (PM) have become one of the important pollutants that deteriorate public health. Since PM is ubiquitous in the atmosphere, it is closely related to life quality in many different ways. Thus, a system to accurately monitor PM in diverse environments is imperative. Previous studies using digital images have relied on individual atmospheric images, not benefiting from both spatial and temporal effects of image sequences. This weakness led to undermining predictive power. To address this drawback, we propose a predictive model using the deep dehazing cascaded CNN and temporal priors. The temporal prior accommodates instantaneous visual moves and estimates PM concentration from residuals between the original and dehazed images. The present method also provides, as by-product, high-quality dehazed image sequences superior to the nontemporal methods. The improvements are supported by various experiments under a range of simulation scenarios and assessments using standard metrics.

E0504: **Minimax powerful functional analysis of covariance tests for longitudinal genome-wide association studies**
Presenter: **Yehua Li**, University of California at Riverside, United States

The purpose is to model the Alzheimer's Disease (AD) related to phenotype response variables observed on irregular time points in longitudinal Genome-Wide Association Studies (GWAS) as sparse functional data and propose nonparametric test procedures to detect functional genotype effects while controlling the confounding effects of environmental covariates. Existing nonparametric tests do not consider within-subject correlations, suffer from low statistical power, and fail to reach the genome-wide significance level. We propose a new class of functional analysis of covariance (fANCOVA) tests based on a seemingly unrelated (SU) kernel smoother, which can incorporate the correlations. We show that the proposed SU-fANCOVA test combined with a uniformly consistent nonparametric covariance function estimator enjoys the Wilks phenomenon and is minimax most powerful. In an application to the Alzheimer's Disease Neuroimaging Initiative data, the proposed test leads to discovering new genes that may be related to AD.

E0362: **Evaluating causal effects of timing of treatment in marginal structural models for longitudinal data**
Presenter: **Hyunkeun Cho**, University of Iowa, United States
Co-authors: Seonjin Kim

A treatment that affects an outcome is often initiated during follow-up in observational studies. If it is of particular interest to assess the effects of timing of treatment on the outcome for a study population, modeling changes in the population mean outcome over time is useful to understand how times to treatment initiation (TTI) have different impacts on the outcome during the follow-up. We propose nonparametric marginal structural models for the time-varying potential outcome in longitudinal observational studies. TTI-varying coefficients in the proposed models describe the population mean outcome over time by accounting for the TTI that influences the changes in the outcome. Confounders on associations between the TTI and the time-varying outcome can occur in observational studies. A double-weighted estimation procedure is proposed based on a kernel function and an inverse of a propensity score. The resultant estimator of the TTI-varying coefficients asymptotically follows a multivariate normal distribution with a mean vector of the true coefficients under mild regularity conditions. In addition, we propose a hypothesis test that investigates whether the initiation of the treatment at a specific time is effective or not. This procedure is applied to evaluate causal effects of times to antiretroviral therapy initiation on an inflammation biomarker for HIV-infected adults who initiate the therapy following the World Health Organization guideline.

E0649: **Analyzing high-dimensional mediators by mixed integer optimization**
Presenter: **Peter Song**, University of Michigan, United States

The purpose is to introduce an extension of the best-subset regularization to perform a high-dimensional mediation analysis in the framework of directed acyclic graphs (DAGs). This new methodology allows a simultaneous operation of parameter clustering and estimation in structural equation models to search causal mediation pathways. The double regularization on homogeneity fusion and sparsity is formulated as a mixed-integer optimization (MIO) problem in the hope to minimize estimation bias and give rise to an appealing setting for post-variable selection inference. We develop a fast and reliable algorithm, Alternating Penalization Operator for L-zero Loss Optimization (APOLLO), to implement the MIO problem for numerical solutions, which is shown to be superior over existing commercial integer programming solver Gurobi. APOLLO algorithm begins with an upper bound search for a warm start, followed by a lower bound search via cutting-planes. The proposed MIO estimator is rigorously investigated for its key theoretical guarantees. Numerical examples illustrate the performance of the proposed MIO solver in simulation experiments and in motivating scientific studies.

EO085 Room R02 NOVEL STATISTICAL MODELING OF MULTIVARIATE DATA Chair: Anuradha Roy

E0298: **Mixtures of factor analyzers with covariates for clustering multiply censored dependent variables**
Presenter: **Tsung-I Lin**, National Chung Hsing University, Taiwan

Censored data frequently arise in diverse applications in which observations to be measured may be subject to some upper and lower detection limits due to the restriction of experimental apparatus. Thus, they are not exactly quantifiable. Mixtures of factor analyzers with censored data (MFAC) have been recently proposed for model-based density estimation and clustering of high-dimensional data under the presence of censored observations. We consider an extended version of MFAC with covariates to accommodate multiply censored dependent variables and develop two analytically feasible EM-type algorithms for computing maximum likelihood estimates of the parameters with closed-form expressions. Moreover, we provide an information-based method to compute asymptotic standard errors of mixing proportions and regression coefficients. The utility and performance of our proposed methodologies are illustrated through several simulated experiments and real data examples.

E0473: **Linear models for doubly multivariate data with block exchangeable covariance structure**
Presenter: **Anuradha Roy**, The University of Texas at San Antonio, United States
Co-authors: Timothy Opheim

The popularity of the classical general linear model (CGLM) is mostly due to the ease of modeling and authentication of the appropriateness of the model. However, CGLM is not appropriate and thus not applicable for correlated multivariate repeated measures data. We propose an extension of the linear model with exchangeably distributed errors for multivariate repeated measures data for multiple observations. Maximum likelihood estimates of the matrix parameters of the intercept, slope and the eigenblocks of the exchangeable error matrix are derived for the data. The practical implications of the methodological aspects of the proposed extended model are demonstrated using two medical datasets.

E0582: **Copula-based bivariate Poisson time series models**
Presenter: **Norou Diawara**, Old Dominion University, United States

The class of bivariate integer-valued time series models is gaining rapid popularity. However, its efficiency and adaptability are being challenged because of algorithm techniques. The computation will be proposed via copula theory. Each series follows a Markov chain with the serial dependence is captured using copula-based transition probabilities with Poisson and zero-inflated Poisson margins. The copula theory is also used to capture the dependence between the two series using either the bivariate gaussian or t copula functions. Likelihood-based inference is used to estimate the models' parameters with the bivariate integrals of the gaussian or t copula functions being evaluated using standard randomized Monte Carlo methods.

E0667: **Hierarchical multiclass discriminant analysis via cross-validation**
Presenter: **Kei Hirose**, Kyushu University, Japan
Co-authors: Kanta Miura

A novel cluster-based LDA is proposed that significantly improves both prediction accuracy and interpretability. We employ hierarchical clustering, and the dissimilarity measure of the two clusters is defined by the cross-validation (CV) value. Therefore, the clusters are constructed such that the error rate is minimized. Our proposed approach requires heavy computational loads because the CV value must be computed at each step of the hierarchical clustering algorithm. We construct an efficient algorithm that computes a consistent estimator of the CV to address the computational issue. The performance of our proposed method is investigated through the application to both artificial and real datasets.

EO097 Room R03 ADVANCES IN STATISTICAL METHODOLOGIES AND APPLICATIONS IN MEDICAL SCIENCES Chair: Hua Shen

E0379: **Alarm time quality, an metric for assessing epidemic detection models within a school absenteeism surveillance system**
Presenter: **Zeny Feng**, University of Guelph, Canada

Model-based school absenteeism surveillance systems have been proposed to raise seasonal influenza epidemic alarms. Previous studies used metrics such as False Alarm Rate, FAR, and accumulated days delayed for model evaluation and selection. However, they were unable to optimize both alarm accuracy and timeliness. We developed a metric, Alarm Time Quality, ATQ, that simultaneously evaluated both aspects by assessing alarms on a gradient, where alarms raised incrementally before or after an optimal time were informative but penalized. Summary statistics of ATQ, average alarm time quality (AATQ) and First Alarm Time Quality, FATQ, were used as a model selection criterion. Alarms raised by ATQ and FAR-selected logistic regression models were compared. Daily school absenteeism and laboratory-confirmed influenza data collected by Wellington-Dufferin-Guelph Public Health was used for demonstration. A simulation study representative of Wellington-Dufferin-Guelph was conducted for further evaluation. ATQ-selected models were found to raise alarms that were timelier than the FAR-selected model.

E0332: **A Bayesian 2D functional linear model: Application to GLCM matrices in LGG cancer radiomics data**
Presenter: **Thierry Chekouo**, University of Calgary, Canada

In cancer radiomics, textural-features evaluated from gray-level co-occurrence matrices (GLCM) have been studied to evaluate gray-level spatial dependence within regions of interest in the brain. Most of these analysis work with summary statistics (or texture-based features) and potentially overlook other structural properties in the GLCM. In the proposed Bayesian framework, we treat each GLCM as a realization of a 2D stochastic functional process observed with error at discrete time points. The latent process is then combined with the outcome model to evaluate the prediction performance. We use simulation studies to assess the performance of our method and apply it to data collected from individuals with lower-grade gliomas. We found that our approach outperforms competing methods that use only summary statistics to predict isocitrate dehydrogenase (IDH) mutation status.

E0311: **Some advances on test for stochastic dominance under density ratio model**
Presenter: **Weiwei Zhuang**, University of Science and Technology of China, China

Stochastic dominance is a partial order between univariate random variables or their cumulative distribution functions. The stochastic dominance relationship is important in finance, economics and many other disciplines. Numerous approaches have been developed to test the hypotheses regarding stochastic dominance. The traditional estimation method is to use empirical distribution functions to estimate it. Considering the populations under comparison are generally of the same nature, we can link the populations through a density ratio model under certain condition. Based on this model, we propose some new estimators for stochastic dominance, restricted stochastic dominance and high order stochastic dominance. We improve the power of the tests through two venues: the use of resampling procedures and the confidence interval approach. Finally, we apply our method to analyze fund performance and family income.

E0308: **Analysis of length-biased survival data with misclassified covariate and partially missing surrogates**
Presenter: **Hua Shen**, University of Calgary, Canada

Misclassification in a categorical variable can often arise in medical research where validation data is absent, and only surrogates subject to missingness are available. Moreover, studies of chronic disease often sample individuals subject to conditions on an event time of interest requiring subjects to have survived to the point of recruitment resulting in length-biased samples. In such studies designed to evaluate the covariate effect on event time, we need to deal with the fact that the distribution of the latent categorical variable is affected by the sampling mechanism. A latent variable model is proposed to deal with the unknown categorical covariate in such a setting and conduct the robust parameter estimation via an expectation-maximization algorithm. Simulation studies demonstrating the performances of the proposed method and an illustration based on the stimulating study on breast cancer are included.

EO215 Room R04 BAYESIAN METHODS WITH APPLICATIONS TO SPORTS, MEDICINE, AND TIME SERIES Chair: Zhihua Ma

E0505: **Zero-inflated Poisson model with clustered regression coefficients: an application to field goal attempts**
Presenter: **Hou-Cheng Yang**, Florida State University, United States
Co-authors: Yishu Xue, Guanyu Hu

Although basketball is a dynamic process sport, with 5 plus 5 players competing on both offense and defense simultaneously, learning some static information is predominant for professional players, coaches and team managers. In order to have a deep understanding of field goal attempts among different players, we propose a zero-inflated Poisson model with clustered regression coefficients to learn the shooting habits of different players over the court and the heterogeneity among them. Specifically, the zero-inflated model recovers the large proportion of the court with zero field goal attempts. The mixture of finite mixtures model learns the heterogeneity among different players based on clustered regression coefficients and inflated probabilities. Both theoretical and empirical justification through simulation studies validates the proposed method. We apply the model to the National Basketball Association (NBA) to learn players' shooting habits and heterogeneity among different players over the 20172018 regular season. This illustrates our model as a way of providing insights from different aspects.

E0528: **Nonparametric Bayesian changepoint model with signed beta process**
Presenter: **Zhihua Ma**, Shenzhen University, China
Co-authors: Catherine Liu, Junshan Shen

A model based on a global view of a stochastic process with changepoints is proposed. By viewing the regime as a time period where the stochastic process runs smoothly and the changepoints as jumping gaps, the modeling of a changepoints model can be separated to model a stochastic process and characterize discontinuities. An extension of the Beta process, called the signed Beta process, is introduced to model the discontinuities under a Bayesian framework. We demonstrate the methods through simulations and several real data examples.

E0552: **Non-negative matrix factorization on count data with concave pairwise group fusion**
Presenter: **Qingyang Liu**, University of Connecticut, United States
Co-authors: Guanyu Hu

Basketball shot location data have been widely investigated over these years. In order to have a profound understanding of shot selection patterns among different players, we propose a heterogeneity learning approach based on non-negative matrix factorization (NMF) of Poisson data. The Poisson NMF decomposes the shot count data into two parts, a few prototypes of shot patterns with smooth variation over the spatial grid and the corresponding scores of all players. We employ a concave pairwise group fusion approach to cluster the shooting habits of different players. The estimation of spatial prototypes and heterogeneity detection among players are conducted simultaneously. Our proposed method is further illustrated by simulation studies and an analysis of shot location data from selected players in the NBA's 2017-2018 regular season.

E0535: **Bayesian network meta-regression for ordinal outcomes with uncertain categories**
Presenter: **Yeongjin Gwon**, University of Nebraska Medical Center, United States

Crohn's Disease (CD) is an inflammatory bowel disease that causes chronic inflammation of the gastrointestinal tract. The study's endpoints to determine disease progression are accepted as the reduction of the CD Activity Index (CDAI) score from the baseline. As the endpoints evolve over time and vary across trials, score reductions (>=70 or >=100) are used to indicate a clinical response. However, this is not consistently used to determine remission, which is a status when the disease is no longer active, as it is defined as the absolute CDAI score (<=150) regardless of the score reduction. This may lead to a challenge in comparative effectiveness research because the number of subjects in remission is associated with that in clinical response. We present a network meta-regression model to deal with such ordinal outcomes in a Bayesian framework. The proposed approach provides an appropriate statistical method in the presence of trials with uncertain categories. We develop an efficient Markov chain Monte Carlo (MCMC) sampling algorithm to carry out Bayesian computation. Deviance information criterion is used for the assessment of goodness-of-fit. A case study demonstrating the usefulness of the proposed methodology is carried out using 10 clinical trials in treating CD.

EO317 Room R05 STATISTICAL METHODS FOR NETWORK DATA Chair: Yuan Zhang

E0451: **Fast network community detection with profile-pseudo likelihood methods**
Presenter: **Emma Jingfei Zhang**, University of Miami, United States

The stochastic block model is one of the most studied network models for community detection. It is well-known that most algorithms proposed for fitting the stochastic block model likelihood function cannot scale to large-scale networks. One prominent work that overcomes this computational challenge proposed a fast pseudo-likelihood approach for fitting stochastic block models to large sparse networks. However, this approach does not have a convergence guarantee and is not well suited for small- or medium-scale networks. We propose a novel likelihood-based approach that decouples row and column labels in the likelihood function, which enables a fast alternating maximization; the new method is computationally efficient, performs well for both small and large scale networks, and has a provable convergence guarantee. We show that our method provides strongly consistent estimates of the communities in a stochastic block model. As demonstrated in simulation studies, the proposed method outperforms the pseudo-likelihood approach in terms of both estimation accuracy and computation efficiency, especially for large sparse networks. We further consider extensions of our proposed method to handle networks with degree heterogeneity and bipartite properties.

E0671: **Spectral analysis of networks with latent space dynamics and signs**
Presenter: **Joshua Cape**, University of Pittsburgh, United States

The problem of modeling and analyzing latent space dynamics in collections of networks is pursued. Towards this end, we pose and study latent space generative models for signed networks that are amenable to inference via spectral methods. Permitting signs, rather than restricting to unsigned networks, enables richer latent space structure and permissible dynamic mechanisms that can be provably inferred via low-rank truncations of observed adjacency matrices. The treatment of and ability to recover latent space dynamics holds across different levels of granularity, namely at the overall graph level, for communities of nodes, and even at the individual node level. We provide synthetic and real data examples to illustrate methodologies' effectiveness and corroborate the accompanying theory. The contributions set forth complement an emerging statistical paradigm for random graph inference encompassing random dot product graphs and generalizations thereof.

E0720: **Learning organized patterns in voxel-wise genome-wide association from imaging-genetics data**
Presenter: **Qiong Wu**, University of Maryland, College Park, United States

The integrative analysis of imaging-genetics data facilitates the systematic evaluation of genetic effects on brain structures and functions with spatial specificity. We focus on voxel-wise genome-wide association analysis, which can involve billions of SNP-voxel pairs. We attempt to identify underlying organized patterns of SNP-voxel association pairs and understand the polygenic influence on brain imaging traits. We propose a new statistical inference framework to identify imaging-genetics association bi-clique, a set of alleles being systematically associated with a set of imaging features. We develop computational strategies to detect latent SNP-voxel bi-cliques and inference models for statistical testing. We further provide theoretical results to guarantee the performance of our computational algorithms and statistical inference. We validate our method by extensive simulation studies and then apply it to a voxel-wise genome-wide association analysis based on genetic data and white matter integrity data from the human connectome project (HCP).

E0715: **Accurate two-sample inference for network data**
Presenter: **Yuan Zhang**, Ohio State University, United States

A method is proposed to compare two networks' structures by the method of moments. The novel procedure is highly accurate and computationally efficient. The test achieves strong theoretical guarantees, namely, optimal separation condition under alternative hypothesis and higher-order accurate type I error rate, without needing resampling or making strong assumptions on knowledge about population distributions. We will showcase some simulations and real-life data applications.

EO297 Room R06 TOPOLOGICAL DATA ANALYSIS FOR SOCIO-ECONOMICS Chair: Dorcas Ofori-Boateng

E0598: **Forecasting COVID-19 spread with topological signatures of atmospheric conditions**
Presenter: **Ignacio Segovia-Dominguez**, University of Texas at Dallas, United States

Since mortality due to COVID-19 may be closely linked to a prior history of lung and other respiratory diseases, ambient air quality might shed important light on the expected severity of COVID-19 and associated survival rates. Understanding the impact of atmospheric conditions and air quality on COVID-19 progression and associated mortality is urgent and critical, not only in terms of efficiently responding to the current pandemic, but also in terms of forecasting impending hotspots and potential next-wave occurrences. We propose a new approach, TLife-LSTM, to investigate potential relationships between atmospheric conditions and air quality and COVID19 dynamics using deep learning models coupled with topological information on weather factors. We validate our framework using the number of confirmed cases and hospitalization rates recorded in the states of Washington and California in the USA. Our results demonstrate the predictive potential of TLife-LSTM in forecasting the dynamics of COVID-19 and modeling its complex spatio-temporal spread dynamics.

E0614: **Applications of topological data analysis ball mapper in economics**
Presenter: **Simon Rudkin**, University of Swansea, United Kingdom
Co-authors: Pawel Dlotko, Wanling Qiu

Topological data analysis ball mapper offers a lens through which to visualise complex multidimensional datasets. Representing each data point in a point cloud allows the topology of the dataset to be explored and captured through a coverage of equally sized balls. These balls provide direct information about the spatial concentration of the data, the connectivity of the dataset and the overall distribution. More powerfully, they may be overlaid with information from any other further variable associated with the data points to permit inference. We discuss how average values of outcome variables can consequentially be understood for their variability across space. Metrics to capture these outcomes are discussed and evaluated for their consistency in response to the methodologies only parameter, the ball radius. Applications in Economics and the Social Sciences are many, and we focus on constituency voting behaviours for the United Kingdom, regional productivity in Europe and self-reported health in Wales. In each case, topological data analysis ball mapper reveals new insights that can direct policy, spearhead recovery and help theorists to understand more about what is happening in the data. Critically, the approach is model-free, simply unlocking the information in data.

E0625: **Topological clustering of multilayer networks**
Presenter: **Asim Kumer Dey**, Princeton University and UT Dallas, United States
Co-authors: Monisha Yuvaraj, Vyacheslav Lyubchich, Yulia Gel, Vincent Poor

Multilayer networks continue to gain significant attention in many areas of study, particularly due to their high utility in modeling interdependent systems such as critical infrastructures, human brain connectome, and socio-environmental ecosystems. However, the clustering of multilayer networks, especially using the information on higher-order interactions of the system entities, remains in its infancy. In turn, higher-order connectivity is often the key in such multilayer network applications as developing optimal partitioning of critical infrastructures in order to isolate unhealthy system components under cyber-physical threats and simultaneous identification of multiple brain regions affected by trauma or mental illness. We introduce the concepts of Topological Data Analysis (TDA) to studies of complex multilayer networks and propose a new topological approach for network clustering. This new topological clustering approach allows for systematic accounting for the important heterogeneous higher-order properties of node interactions within and in-between network layers and integrating information from the node neighbors and their interactions. We illustrate the utility of the proposed clustering algorithm by applying it to an emerging problem of societal importance - vulnerability zoning of residential properties to weather- and climate-induced risks in the context of house insurance claim dynamics.

E0689: **Topology-based anomaly detection in dynamic multilayer networks**
Presenter: **Dorcas Ofori-Boateng**, Portland State University, United States
Co-authors: Yulia Gel, Ignacio Segovia-Dominguez, Murat Kantarcioglu, Cuneyt Akcora

Motivated by the recent surge of criminal activities involving cross-cryptocurrency trades, we introduce a new topological perspective to structural anomaly detection in dynamic multilayer networks. We postulate that anomalies in the underlying blockchain transaction graph that are composed of multiple layers are likely to be also manifested in anomalous patterns of the network shape properties. As such, we invoke the machinery of clique persistent homology on graphs to systematically and efficiently track the evolution of the network shape and, as a result, to detect changes in the underlying network topology and geometry. We develop a new persistence summary for multilayer networks, called the stacked persistence diagram, and prove its stability under input data perturbations. We validate our new topological anomaly detection framework in application to dynamic multilayer networks from the Ethereum Blockchain and the Ripple Credit Network and show that our stacked PD approach substantially outperforms the state-of-art techniques, yielding up to 40% gains in precision.

EO225 Room R07 NOVEL METHODS AND APPLICATIONS IN STATISTICS Chair: Yeonhee Park

E0577: **Sparse and efficient estimation with semiparametric models in meta-analysis**
Presenter: **Sunyoung Shin**, University of Texas at Dallas, United States

Semiparametric regression models in meta-analysis, where studies of heterogeneous designs may collect different types of data, are considered. With regression coefficients and their covariance estimates available for each study, we efficiently combine the study-specific estimates. Next, we employ penalized least-squares approximation technique for model selection, followed by final estimation under the selected model. The proposed approach, named sparse semiparametric meta-estimation, may improve statistical power and selection accuracy. We establish semiparametric efficiency and selection consistency of the sparse meta-estimator under reasonable assumptions. The superior performance and practical utility of the proposed method are demonstrated through numerical studies.

E0541: **A forward approach for sufficient dimension reduction in binary classification**
Presenter: **Seung Jun Shin**, Korea University, Korea, South
Co-authors: Jongkyeong Kang

Since the seminal sliced inverse regression (SIR) proposed, the inverse-type methods have been canonical in sufficient dimension reduction (SDR). However, they often suffer in binary classification since the binary response yields two slices at most. We develop a forward approach for SDR in binary classification based on weighted large-margin classifiers. We first show that the gradient of a large-margin classifier is unbiased for SDR as long as the corresponding loss function is Fisher consistent. This leads us to propose what we call the weighted outer-product of gradients (wOPG) method. The WOPG can recover the central subspace exhaustively without linearity or constant variance conditions routinely required for the inverse-type methods. We study the asymptotic behavior of the proposed estimator and demonstrate its promising finite-sample performance for both simulated and real data examples.

E0603: **Causal inference with hidden confounders: a comparison between two stage least squares and the causal Dantzig**
Presenter: **James Long**, University of Texas MD Anderson Cancer Center, United States
Co-authors: Min Jin Ha

The Causal Dantzig (CD) estimates causal effects by exploiting shifts in the exposure distribution across a set of data collection environments (e.g. experimental and observational). It is one of a small number of methods, along with instrumental variables techniques, which are consistent under hidden confounding. We propose a model for jointly analyzing the performance of the CD and the classical Two-Stage Least-Squares (TSLS) instrumental variable estimator. The model is appropriate for many settings, including genetic perturbation experiments and most standard applications of instrumental variables estimators. We derive the first analytic results comparing the CD and TSLS, including conditions under which TSLS has lower asymptotic variance than the CD. We compare regularized versions of the CD, TSLS, and other state-of-the-art methods in high dimensional genetic perturbation simulations and a real yeast perturbation data set. From a fitting perspective, the high dimensional CD is simpler than TSLS because the environment simplifies the selection of tuning parameters. Performance of the methods is assessed by accuracy in predicting the effect of test set knock out experiments. Regularized TSLS procedures obtain the best performance in many scenarios tested.

E0568: **Bayesian inference of the number of classes in restricted latent class models**
Presenter: **Yinghan Chen**, University of Nevada, Reno, United States
Co-authors: Yuguo Chen

Cognitive diagnosis models (CDMs) are structured latent class models widely used to classify a multidimensional collection of latent attributes. The applications of CDMs rely on the specification of the Q matrix, a binary matrix representing the requirement of each attribute in the test items. Estimation of the Q matrix is an important question for the correct classification of attribute profiles. Many existing exploratory methods for estimation of Q must pre-specify the number of attributes, K. We present a Bayesian framework for general CDMs to jointly infer the number of attributes K and the elements of Q. Using stick-breaking construction of priors and a Bayesian variable selection technique, we propose a constrained Gibbs sampling algorithm to estimate the underlying Q and model parameters of varying dimensions. The proposed method can also enforce model identifiability constraints.

EO227 Room R08 RECENT DEVELOPMENTS IN CLIMATE/ENVIRONMENTAL STATISTICS Chair: Whitney Huang

E0286: **A new computer model calibration framework based on deep neural network**
Presenter: **Won Chang**, University of Cincinnati, United States
Co-authors: Saumya Bhatnagar, Seonjin Kim, Jiali Wang

Computer model calibration is a statistical framework for combining information from computer model runs and the corresponding real-world observations to quantify and reduce parametric uncertainties. The existing calibration framework is subject to various issues, including non-identifiability between input parameter effects and data-model discrepancy effects and computational challenges due to large spatial or temporal model output. The aim is to develop a new class of statistical calibration framework based on deep neural network models that do not suffer from the major shortcomings of the existing approach. The central idea is to model the inverse relationship between the model output and input parameters directly using deep learning network models. By utilizing the feature extraction ability of deep neural networks, the proposed approach can filter out the signal from data-model discrepancy and accurately estimate the true input parameter values. The computational machinery currently available for deep neural network enables highly efficient computation even for large data with highly complex dependence structures. Our simulation study and real data application for the WRF-Hydro model show that our new method can provide accurate estimates for input parameter values along with sound uncertainty measures even under the presence of significant data-model discrepancies.

E0554: **Topological estimation of 2D image data via subsampling and applications**
Presenter: **Matthew Jester**, University of North Carolina at Greensboro, United States
Co-authors: Yu-Min Chung, Xiaoli Gao, Sarah Day, Kaitlin Keegan

A novel statistical approach is developed to estimate topological information from large, noisy images. The main motivation is to measure pore microstructure in 2-dimensional X-ray micro-computed tomography (micro-CT) images of ice cores at different depths. The pore space in these samples is where gas can move and get trapped within the ice column and is of interest to climate scientists. While the field of topological data analysis offers tools for estimating topological information in noisy images, direct application of these techniques to large images often leads to inaccuracies and proves infeasible as image size and noise levels grow. Our approach uses image subsampling to estimate the number of holes of a prescribed size range in a computationally feasible manner. We use a synthetic data set created by a Fourier discretization to validate our approach. In applications where holes naturally have a known size range on a smaller scale than the full image, this approach offers a means of estimating Betti numbers, or global counts of holes of various dimensions, via subsampling of the image.

E0567: **Bayesian nonparametric multivariate spatial mixture mixed effects models**
Presenter: **Scott Holan**, University of Missouri, United States
Co-authors: Ryan Janicki, Andrew Raim, Jerry Maples

Leveraging multivariate spatial dependence to improve the precision of estimates using American Community Survey data and other sample survey data has been a topic of recent interest among data users and federal statistical agencies. One strategy is to use a multivariate spatial mixed-effects model with a Gaussian observation model and latent Gaussian process model. In practice, this works well for a wide range of tabulations. Nevertheless, in situations that exhibit heterogeneity among geographies and/or sparsity in the data, the Gaussian assumptions may be problematic and lead to underperformance. To remedy these situations, we propose a multivariate hierarchical Bayesian nonparametric mixed-effects spatial mixture model to increase model flexibility. The number of clusters is chosen automatically in a data-driven manner. The effectiveness of our approach is demonstrated through a simulation study and motivating application of special tabulations for American Community Survey data.

E0725: **Modeling the impacts of climate factors across the distribution of wildfires**
Presenter: **Adam Diaz**, Clemson University, United States

Climate change is widely expected to affect the behavior of wildland fires. However, the question of which climate factors are the most meaningful in the spread and sustenance of fires is not very straightforward to answer. The behavior of fire is dependent on terrain and geographic location, among many other non-climate factors. Additionally, the climate conditions that are ideal for ignition do not necessarily coincide with those that allow for a fire to grow completely out of control. We use a historical data set of California wildfires and ERA5 climate reanalysis data to model how this relationship differs between run-of-the-mill fires and extensive wildfires. Using a Bayesian Hierarchical framework, we attempt to relate wildfire characteristics with a set of climate covariates. Applying Integrated Nested Laplace Approximation (INLA) allows us to efficiently fit highly complex SPDE models and perform inference on our large data set. We utilize a spatial mixture model to account for the rich spatial dependence structure of the data, employ techniques from extreme value theory to analyze the most extreme fire events effectively.

Friday 25.06.2021 09:50 - 11:05 Parallel Session H – EcoSta2021

EI009 Room R08 MACHINE LEARNING AND APPROXIMATION THEORY Chair: Andreas Christmann

E0153: **Theory of deep convolutional neural networks and distributed learning**
Presenter: **Ding-Xuan Zhou**, City University of Hong Kong, Hong Kong

Deep learning has been widely applied and brought breakthroughs in speech recognition, computer vision, and many other domains. The involved deep neural network architectures and computational issues have been well studied in machine learning. But there lacks a theoretical foundation for understanding the modelling, approximation or generalization ability of deep learning models with network architectures. An important family of structured networks is deep convolutional neural networks (CNNs) with convolutional structures. The convolutional architecture gives essential differences between the deep CNNs and fully-connected deep neural networks, and the classical theory for fully-connected networks developed around 30 years ago does not apply. This talk describes a mathematical theory of deep CNNs associated with the rectified linear unit (ReLU) activation function. In particular, we give the first proof for the universality of deep CNNs, meaning that a deep CNN can be used to approximate any continuous function to an arbitrary accuracy when the depth of the neural network is large enough. We also give explicit approximation rates and show that the approximation ability of deep CNNs is at least as good as that of fully-connected multi-layer neural networks. Distributed learning with deep CNNs will also be discussed.

E0154: **Stochastic gradient descent with non-smooth loss: Stability, generalization and differential privacy**
Presenter: **Yiming Ying**, State University of New York at Albany, United States

Recently there is a large amount of work devoted to the study of algorithmic stability, generalization and differential privacy for stochastic gradient descent (SGD). However, most of them require to impose restrictive assumptions such as the boundedness of gradients, strongly smoothness on the loss functions. We show that these restrictive assumptions are not required for SGD and show that stability, generalization, and differential privacy for SGD still hold for non-smooth loss functions (e.g., hinge loss, Huber loss and absolute loss) even in unbounded domains. In particular, we first establish stability and generalization for SGD by removing the existing bounded gradient assumptions. Secondly, the smoothness assumption is relaxed by considering loss functions with Holder continuous gradients, for which we show that optimal bounds are still achieved by balancing computation and stability. Finally, we establish differential privacy guarantees for SGD with non-smooth loss functions. To the best of our knowledge, these are the first-ever known results for stability, generalization and differential privacy of SGD with non-smooth losses where no extra smoothing techniques are employed in its implementation.

EO374 Room R01 ADVANCES IN HIGH-DIMENSIONAL AND COMPLEX DATA ANALYSIS Chair: Debashis Paul

E0324: **High-dimensional classifiers under the strongly spiked eigenvalue model**
Presenter: **Aki Ishii**, Tokyo University of Science, Japan
Co-authors: Kazuyoshi Yata, Makoto Aoshima

One of the features of modern data is that the data dimension is extremely high. However, the sample size is relatively small. We call such data HDLSS data. In HDLSS situations, new theories and methodologies are required to develop for statistical inferences. We note that eigenvalues of high-dimensional data grow very rapidly depending on the dimension. There are two types of high-dimensional eigenvalue models: the strongly spiked eigenvalue (SSE) and non-SSE (NSSE) models. We consider high-dimensional classification under the SSE model. We give new classifiers by using a data transformation technique. We show that our classifiers have preferable properties in theory. Finally, we check the performances of our classifiers in simulations.

E0584: **Spectral estimation for high-dimensional linear processes**
Presenter: **Jamshid Namdari**, University of California Davis, United States
Co-authors: Alexander Aue, Debashis Paul

An estimation procedure for a class of high dimensional linear time series is proposed by estimating the joint eigenvalue distribution of the coefficient matrices of the process. The process being considered is of the form $X_t = \sum_{\ell=0}^{\infty} \mathbf{A}_\ell Z_{t-\ell}$, where $\{Z_t\}$ are i.i.d., p-dimensional random vectors with zero mean, and the coefficient matrices $\{\mathbf{A}_\ell\}$ and $\mathrm{Var}(Z_t)$ are digaonalizable in a common orthonormal basis. The proposed estimators rely on the asymptotic behavior of weighted integrals of the sample periodogram. Under the asymptotic regime where $p,n \to \infty$ such that $p/n \to c \in (0,\infty)$, a Marčenko-Pastur type limiting distribution for the aforementioned weighted sample periodograms is established. The limiting Stieltjes transforms of the respective empirical spectral distributions is utilized to develop a class of estimators by minimizing an L^κ discrepancy measure (for $\kappa > 0$) between the empirical and limiting Stiltjes transforms of the joint spectral distribution of the coefficient matrices of the linear process, by assuming that the latter is a discrete mixture of point masses. Finally, the methodology is illustrated through simulations and analysis of stock prices from the S&P 500 series.

E0696: **Estimation of spectra of high-dimensional separable covariance matrices**
Presenter: **Debashis Paul**, University of California, Davis, United States
Co-authors: Lili Wang

The aim is to estimate the joint spectra of high-dimensional time series for which the observed data matrix is assumed to have a separable covariance structure. The primary interest is in estimating the distribution of the eigenvalues of the marginal covariance of the observation vectors under partial information – such as stationarity or sparsity – on the temporal covariance structure. We develop a method that utilizes random matrix theory to estimate the unknown population spectra by repressing the spectrum of the dimensional covariance matrix on a simplex. We prove the consistency of the proposed estimator under the dimension proportional to the sample size setting. Furthermore, we develop a resampling based method for statistical inference on low-dimensional functionals of the joint spectrum of the population covariance matrix.

EO071 Room R02 ADVANCES IN CAUSAL INFERENCE Chair: Luke Keele

E0206: **Bracketing methods for difference-in-differences based on monotone trend assumption**
Presenter: **Luke Keele**, University of Pennsylvania, United States

The method of differences-in-differences (DID) is widely used to estimate the causal effect of policy interventions with observational data. DID exploits a before and after comparison of the treated and control group to remove time-invariant additive bias. However, estimates from DID will be biased if an event besides the treatment occurs at the same time and affects the treated group in a differential fashion. Recent work outlined a method of DID bracketing bounds to account for bias from an unmeasured confounder that has a differential effect in the post-treatment time period. These DID bracketing bounds require partitioning the control units into two separate groups based on past levels of the outcomes. We develop DID bracketing bounds based on a partition of control units based on changes in the outcomes in past time periods. We develop the identification conditions for the bounds, and we show that we can better account for bias from time-varying confounders. We then derive a key falsification test to probe a necessary assumption. Next, we outline a method of sensitivity analysis that adjusts the bounds for possible bias based on differences between the treated and control units from the pretreatment period. We apply these new methods to an application on the effect of voter identification laws on turnout. Specifically, we focus on estimating whether the enactment of voter identification laws in Georgia and Indiana affected voter turnout.

E0444: **Bounding local average treatment effects under exclusion-restriction violations in mobile health interventions**
Presenter: **Andrew Spieker**, Vanderbilt University Medical Center, United States

In the modern era, mobile health interventions are of increasing interest. In the context of randomized trials based on text-message interventions, emphasis has been placed on providing patients with opportunities for engagement and responses to the text-messages. One goal in such settings is to evaluate the effect of engagement with the intervention on the outcome of interest. As engagement is a post-randomization exposure that can be thought of as analogous to compliance in a randomized drug study, instrumental variable approaches appear at first to be a reasonable way to address this goal. However, in pragmatic text-message based interventions, violations to the exclusion-restriction assumption are almost certainly to be expected. In this sense, the goal is truly to tease apart the effect of the intervention itself and the effect of engagement with the intervention. We develop a sensitivity analysis approach to place sharp bounds on the local average treatment effect under different levels of engagement. We further illustrate this approach to accommodate a wide range of scenarios and various levels of departures from the exclusion-restriction assumption.

E0517: **Causal graphical views of fixed effects and random effects models**
Presenter: **Yongnam Kim**, Seoul National University, Korea, South
Co-authors: Peter Steiner

Despite the long-standing discussion on fixed effects (FE) and random effects (RE) models, how and under which conditions both methods can eliminate unmeasured confounding bias have not yet been widely understood in practice. Using a simple pretest-posttest design in a linear setting, we translate the conventional algebraic formalization of FE and RE models into causal graphs and provide intuitively accessible graphical explanations about their data-generating and bias-removing processes. The proposed causal graphs highlight that FE and RE models consider different data-generating models. RE models presume a data-generating model identical to a randomized controlled trial, while FE models allow for unobserved time-invariant treatment-outcome confounding. Augmenting regular causal graphs that describe data-generating processes by adding the computational structures of FE and RE estimators, we visualize how FE estimators and RE estimators offset unmeasured confounding bias. In contrast to standard regression or matching estimators that reduce confounding bias by blocking non-causal paths via conditioning, FE and RE estimators offset confounding bias by deliberately creating new non-causal paths and associations of opposite sign. Though FE and RE estimators are similar in their bias-offsetting mechanisms, the augmented graphs reveal their subtle differences that can result in different biases in observational studies.

EO121 Room R03 RECENT DEVELOPMENT IN BIOSTATISTICS Chair: Kyuhyun Kim

E0292: **A two-phase approach to account for unmeasured confounding and censoring of a survival endpoint**
Presenter: **Jaeun Choi**, Albert Einstein College of Medicine, United States

Consistent estimation of the effect of a treatment in the presence of unmeasured confounding is a common objective in observational studies. The Two-Stage Least-Squares (2SLS) Instrumental Variables (IV) procedure is frequently used but not applicable to time-to-event data with some observations censored. We develop a statistical method to account for unmeasured confounding of the effect of treatment on survival endpoints subject to censoring by considering censoring and confounding in sequence. We first jointly model survival time and treatment using a simultaneous equations model (SEM) under a specific bivariate distribution for the underlying data generating process. The joint model is used for the sole purpose of imputing the censored survival times. Then we apply an IV procedure to the completed dataset. This two-phase approach allows censoring to be accounted for while preserving the IV method's robustness to distributional miss-specifications in the joint model. The approach can be applied to any type of survival outcome, including continuous and fixed-time endpoints. The methodology is illustrated on two examples of a vascular surgery study and a mental health study.

E0268: **Bayesian causal inference with some invalid instrumental variables**
Presenter: **Gyuhyeong Goh**, Kansas State University, United States
Co-authors: Jisang Yu

In observational studies, instrumental variables estimation is greatly utilized to identify causal effects. The instrumental variables estimator is consistent only if instruments are not correlated with the error term of the estimation equation of interest, often referred to as the exclusion restriction condition. We aim to propose a Bayesian generalized method of moments approach to make consistent inferences about the causal effect when there are some invalid instruments in a way that they violate the exclusion restriction condition. Asymptotic properties of the proposed Bayes estimator, including consistency and normality, are established. A simulation study demonstrates that the proposed Bayesian method produces consistent point estimators and valid, credible intervals with correct coverage rates for Gaussian and non-Gaussian data with some invalid instruments. The proposed method is also examined through a real data application.

E0593: **Smoothed quantile regression for censored residual lifetime**
Presenter: **Kyuhyun Kim**, Yonsei University, Korea, South
Co-authors: Sangwook Kang

A regression modeling of the quantiles of the residual lifetime at a specific time given a set of covariates is considered. We propose an induced smoothed version of the existing non-smooth estimating equations approaches to estimate regression parameters. The proposed estimating equations are smooth in regression parameters so that solutions can be readily obtained via standard numerical algorithms. Moreover, smoothness in the proposed estimating equations enables one to obtain a closed-form expression of the robust sandwich-type covariance estimator of regression estimators. To handle data under right censoring, inverse probabilities of censoring are incorporated as weights. Consistency and asymptotic normality of the proposed estimator are established. Extensive simulation studies are conducted to verify the performances of the proposed estimator under various finite samples settings. We apply the proposed method to dental study data evaluating the longevity of dental restorations.

EO211 Room R04 ADVANCES IN FINITE MIXTURE MODELS: METHODOLOGY AND APPLICATIONS I Chair: Abbas Khalili

E0445: **Finite mixture regression models for treatment effect parameters when high-dimensional nuisance parameters are present**
Presenter: **Hiroyuki Kasahara**, University of British Columbia, Canada

The problem of estimating a low-dimensional parameter of interest-treatment parameter- is studied in the presence of high-dimensional parameter in finite mixture regression models. When the number of covariates diverges with the sample size, regularization methods such as lasso or ridge are employed to reduce variance by trading off between regularization bias and overfitting. However, the regularization induces substantial biases in the estimator of component-specific treatment parameter. We develop a procedure for estimating the low-dimensional component-specific treatment parameter based on Neyman-orthogonalization and sample splitting by extending the double/debiased estimation method of other researchers to finite mixture regression models. We randomly partition the sample into, say, $K = 2$ subsamples. Using the first sample, we employ a variant of the penalized likelihood approach for variable selection while keeping the low dimensional treatment variable always selected. Given the estimated mixture models and the selected variables from the first subsample, we re-estimate the low-dimensional treatment parameter based on the Neyman orthogonal score (efficient score) moment conditions by employing a variant of the EM algorithm using the second sample. Simulation shows the proposed procedure substantially improves the regularization bias and leads to good performance in inference.

E0611: **Identifiability of hierarchical latent attribute models**
Presenter: **Gongjun Xu**, University of Michigan, United States
Co-authors: Yuqi Gu

Hierarchical Latent Attribute Models (HLAMs) are a family of discrete latent variable models attracting increasing attention in educational, psychological, and behavioral sciences. The key ingredients of an HLAM include a binary structural matrix and a directed acyclic graph specifying hierarchical constraints on the configurations of latent attributes. These components encode practitioners' design information and carry important scientific meanings. Despite the popularity of HLAMs, the fundamental identifiability issue remains unaddressed. The existence of the attribute hierarchy graph leads to degenerate parameter space, and the potentially unknown structural matrix further complicates the identifiability problem. The purpose is to address the issue of identifying the latent structure and model parameters underlying an HLAM. We develop sufficient and necessary identifiability conditions. These results directly and sharply characterize the different impacts on identifiability cast by different attribute types in the graph. The proposed conditions provide insights into diagnostic test designs under the attribute hierarchy and serve as tools to assess the validity of an estimated HLAM.

E0699: **Statistical efficiency of parameter estimation in contaminated models**
Presenter: **Nhat Ho**, University of Texas, Austin, United States

Theoretical treatment of contaminated models is provided. The theory is based on a novel notion of distinguishability between a known distribution and a contaminated distribution. Drawing on optimal transport, we establish a connection between the distinguishability and the convergence rates of parameter estimation. Finally, we demonstrate that these convergence rates are minimax optimal under certain settings of the contaminated models.

EO125 Room R05 INFERENCE IN STATISTICAL NETWORKS AND GRAPHICAL MODELS Chair: Kuang-Yao Lee

E0235: **Learning healthcare delivery network with longitudinal electronic health records data**
Presenter: **Jiehuan Sun**, University of Illinois at Chicago, United States
Co-authors: Katherine Liao, Tianxi Cai

A few approaches have been proposed to learn healthcare delivery network using electronic health records (EHR) data, that is, to infer the relationship among different types of medical encounters (e.g. disease diagnosis, treatments, and procedures), which might, in turn, improve the healthcare system. However, these approaches do not fully account for the properties of the EHR data, including the longitudinal nature and patient heterogeneity. We propose a flexible model built upon the multivariate Hawkes process for HDN construction. Our model allows for patient-specific time-varying background intensity functions via random effects, which can also adjust for effects of important covariates. We adopt a penalized approach to select fixed effects, yielding a sparse network structure and removing unnecessary random effects from the model. Through extensive simulation studies and a dataset of type 2 diabetes, we show that the proposed method performs well in recovering the network structure and that it is essential to account for patient heterogeneities.

E0587: **On sufficient graphical models**
Presenter: **Kyongwon Kim**, Ewha Womans University, Korea, South

A sufficient graphical model is introduced by applying the recently developed nonlinear sufficient dimension reduction techniques to evaluate conditional independence. The graphical model is nonparametric in nature, as it does not make distributional assumptions such as the Gaussian or copula Gaussian assumptions. However, unlike a fully nonparametric graphical model, which relies on the high-dimensional kernel to characterize conditional independence, the proposed graphical model is based on conditional independence given a set of sufficient predictors with a substantially reduced dimension. In this way, we avoid the curse of dimensionality that comes with a high-dimensional kernel. We develop the population-level properties, convergence rate, and variable selection consistency of our estimate. By simulation comparisons and an analysis of the DREAM 4 Challenge data set, we demonstrate that our method outperforms the existing methods when the Gaussian or copula Gaussian assumptions are violated. Its performance remains excellent in the high-dimensional setting.

E0672: **Estimating finite mixtures of ordinal graphical models**
Presenter: **Kevin Lee**, Western Michigan University, United States
Co-authors: Qian Chen, Wayne DeSarbo, Lingzhou Xue

Graphical models have received increasing attention in network psychometrics as a promising probabilistic approach to studying conditional relations among variables using graph theory. Despite recent advances, existing methods on graphical models usually assume a homogeneous population and focus on binary or continuous variables. However, ordinal variables are very popular in many areas of psychological science, and the population often consists of several different groups based on the heterogeneity in ordinal data. Driven by these needs, we introduce the finite mixture of ordinal graphical models to study the heterogeneous conditional dependence relationships of ordinal data effectively. We develop a penalized likelihood approach for model estimation and design a generalized expectation-maximization (EM) algorithm to solve the computational challenges. We examine the performance of the proposed method and algorithm in simulation studies. Moreover, we demonstrate the potential usefulness of the proposed method in psychological science through a real application concerning the interests and attitudes related to fan avidity for students in a large public university in the United States.

EO177 Room R06 MULTIPLE FACETS OF DIMENSION REDUCTION Chair: Lo-Bin Chang

E0343: Nonparametric Bayesian latent factor model for multivariate functional clustering

Presenter: **Yeonseung Chung**, Korea Advanced Institute of Science and Technology, Korea, South
Co-authors: Taeryon Choi, Daewon Yang

Nowadays, multivariate functional data are frequently encountered in many fields of science. While there exist a variety of methodologies for univariate functional clustering, the approaches for multivariate functional clustering are less studied. Moreover, there is little research for functional clustering methods incorporating additional covariate information. We propose a Bayesian nonparametric sparse latent factor model for covariate-dependent multivariate functional clustering. Multiple functional curves are represented by basis coefficients for splines, which are reduced to latent factors through a Bayesian sparse latent factor model. Then, the factors and covariates are jointly modeled using a Dirichlet process (DP) mixture of Gaussians to facilitate a model-based covariate dependent multivariate functional clustering. The method is further extended to dynamic multivariate functional clustering to handle sequential multivariate functional data. The proposed methods are illustrated through a simulation study and applications to Canadian weather and air pollution data.

E0347: Two-stage approach to address the over-fitting problem in partial least squares regression

Presenter: **Seunggeun Lee**, Seoul National University, Korea, South
Co-authors: Rounak Dey, Kwangsik Nho

Partial least squares (PLS) is a widely used multivariate analysis method, which finds latent structures using both outcomes and predictors. In high-dimensional data with large numbers of predictors, however, PLS can over-fit the model and produce fitted outcomes that are surprisingly similar to the observed outcomes, even when there is little relationship between the outcomes and the predictors. Sparse PLS (SPLS) can partially address this problem. However, if the sparsity assumption is violated, it can still produce biased results. We systematically investigate the over-fitting problem and propose a simple two-stage PLS (TPLS) method, which first removes dimensions unlikely related to the latent structures and then performs PLS with the reduced dimension. To reduce the dimensions and to improve prediction accuracy, our approach utilizes the recent theoretical development of high-dimensional principal component analysis. To infer the proportion of the variability of outcomes explained by high dimensional predictors, we further derive a novel estimator of R2. We have shown that our method is robust and enables accurate inference and prediction through extensive simulation studies and the application to the Alzheimer's Disease Neuroimaging Initiative (ADNI) data.

E0447: Dimensionality reduction through conditional modeling

Presenter: **Lo-Bin Chang**, Ohio State University, United States
Co-authors: Eran Borenstein, Wei Zhang, Stuart Geman

Modeling high dimensional data often encounters issues of data sparsity and computational complexity. Among the most common approaches to addressing these difficulties is through dimensionality reduction, which amounts to searching for low-dimensional features or statistics in place of the high-dimensional raw data. However, proper selection of the features is critical. We provide a mathematically coherent approach to learning features and specifying the model through a conditional modeling technique. To demonstrate our approach, a fully generative image model will be developed where the resulting probability distribution is on image pixel intensities rather than on the features per se. Parameter estimation and applications on image classification and detection will be discussed to illustrate the practical usage of the model.

EO265 Room R07 RECENT ADVANCES IN ANALYSIS OF DEPENDENT DATA Chair: Yi Shen

E0452: Random topology in soft-thresholded Gaussian models

Presenter: **Yi Shen**, University of Waterloo, Canada
Co-authors: Jian Kang, Paul Marriott, Weinan Qi

The soft-thresholded Gaussian model has been developed in biostatistics with applications in brain imaging. It has a Bayesian structure and hence requires a rule to choose an appropriate prior distribution. This often means choosing the height of the threshold according to known information, for example, the number of active areas, which corresponds to the number of connected components of the excursion set above the threshold. We discuss the recent results that we obtained concerning the distribution of such a number. More precisely, we show that for certain Gaussian random fields, when the threshold tends to infinity and the searching area expands with a matching speed, both the location of the excursion sets and the location of the local maxima above the threshold will converge weakly to a Poisson point process. Moreover, when the threshold is high but not tending to infinity, the distribution of these locations can be satisfactorily approximated by a Poisson process plus a correction term. We provide theoretical support to predict the number of connecting components when performing topological data analysis to the extreme values of a random field model.

E0748: Disjoint and sliding blocks estimators for heavy tailed time series

Presenter: **Rafal Kulik**, University of Ottawa, Canada

Multivariate regularly varying time series with both weak and strong dependence are considered. Using probabilistic tools developed recently, we will study estimators of so-called cluster functionals. These cluster functionals quantify extremal clustering. We will present the asymptotic theory for disjoint blocks and sliding blocks estimators in the peak-over-threshold framework. In particular, we will show that both classes of estimators have the same asymptotic variance. This is in contrast to the situation when the block maxima method is implemented. Some open questions, especially in the context of long memory, will be presented.

E0652: Optimal sampling designs for online estimation of streaming multi-dimensional time series

Presenter: **Rui Xie**, Universify of Central Florida, United States
Co-authors: Shuyang Bai, Ping Ma

Online analysis of streaming time-series data often faces a trade-off between statistical efficiency and computational cost. One important approach to balance this trade-off is sampling, where only a small portion of the sample is selected for the model fitting and update. We study the data-dependent sample selection and online analysis problem for a multi-dimensional streaming time series. Motivated by the D-optimality criterion in the design of experiments, we propose a class of online data reduction methods that achieve an optimal sampling criterion and improve the computational efficiency of the online analysis. We show that the optimal solution amounts to a strategy that is a mixture of Bernoulli sampling and leverage score sampling. The leverage score sampling involves an auxiliary estimate of an inverse covariance matrix updated sparsely to gain the computational advantage. Theoretical properties of the auxiliary estimations involved are established. The performance of the sampling-assisted online estimation method is assessed via simulation studies and a real data example.

Friday 25.06.2021 11:15 - 12:30 Parallel Session I – EcoSta2021

EO285 Room R01 HIGH-DIMENSIONAL METHODS IN OBSERVATIONAL STUDIES Chair: Daoji Li

E0185: **Selective inference with robust Q-learning**
Presenter: **Ashkan Ertefaie**, University of Rochester, United States
Co-authors: Robert Strawderman

Constructing an optimal dynamic treatment regime becomes complex when there are many prognostic factors, such as patients' genetic information, demographic characteristics, and medical history over time. Existing methods only focus on selecting the important variables for the decision-making process and fall short in providing inference for the selected model. We fill this gap by leveraging the conditional selective inference methodology. We show that the proposed method is asymptotically valid given certain rate assumptions in semiparametric regression.

E0737: **Deconfounding scores: Feature representations for causal effect estimation with weak overlap**
Presenter: **Alexander DAmour**, Google Research, United States
Co-authors: Alexander Franks

A key condition for obtaining reliable estimates of the causal effect of a treatment is overlap (a.k.a. positivity): the distributions of the features used to perform causal adjustment cannot be too different in the treated and control groups. In cases where the overlap is poor, causal effect estimators can become brittle, especially when incorporating weighting. To address this problem, a number of proposals (including confounder selection or dimension reduction methods) incorporate feature representations to induce better overlap between the treated and control groups. A key concern in these proposals is that the representation may introduce confounding bias into the effect estimator. We introduce deconfounding scores, which are feature representations that induce better overlap without biasing the estimation target. We show that deconfounding scores satisfy a zero-covariance condition that is identifiable in observed data. As a proof of concept, we characterize a family of deconfounding scores in a simplified setting with Gaussian covariates. We show that these scores can be used to construct estimators with good finite-sample properties in some simple simulations. In particular, we show that this technique could be an attractive alternative to standard regularizations often applied to IPW and balancing weights.

E0738: **Causal inference in high dimensions without sparsity**
Presenter: **Steve Yadlowsky**, Google Research, United States

The focus is on the problem of estimating the average treatment effect in the presence of fully observed, high dimensional confounding variables, where the number of confounders d is of the same order as the sample size n. To make the problem tractable, we posit a generalized linear model for the effect of the confounders on the treatment assignment and outcomes but do not assume any sparsity. Instead, we only require the magnitude of confounding to remain non-degenerate. Despite making parametric assumptions, this setting is a useful surrogate for some machine learning methods used to adjust for confounding in two-stage methods. In particular, the estimation of the first stage adds variance that does not vanish, forcing us to confront terms in the asymptotic expansion that normally are brushed aside as finite sample defects. We compare the parametric g-formula, IPW, and two common doubly robust estimators—augmented IPW (AIPW) and targeted maximum likelihood estimation (TMLE). When the outcome model estimates are unbiased, the g-formula outperforms the other estimators in both bias and variance. Among the doubly robust estimators, the TMLE estimator has the lowest variance. Existing theoretical results do not explain this advantage because the TMLE and AIPW estimators have the same asymptotic influence function. However, our model emphasizes differences in performance between these estimators beyond first-order asymptotics.

EO041 Room R03 NEW FLEXIBLE AND ROBUST STATISTICAL TOOLS FOR BIOMEDICAL RESEARCH Chair: Yanlin Tang

E0158: **Locally homogeneous censored quantile regression model with time-dependent covariates**
Presenter: **Tony Sit**, The Chinese University of Hong Kong, Hong Kong

Traditionally, censored quantile regression stipulates a specific, pointwise conditional quantile of the survival time given covariates. Although such a formulation provides a great deal of model flexibility and interpretability, the pointwise quantile estimates can sometimes be rather unstable across neighbouring quantile levels with substantially large variances. In view of this phenomenon, we propose a new class of censored quantile regression models with time-dependent covariates subject to right censoring. The resulting model can also be regarded as a generalisation of the accelerated failure time model for survival data in the sense that it relaxes the assumption of global homogeneity for the residual. In particular, such homogeneity for the residual, hence the conditional quantile, affecting the lifetime distribution is assumed only for a specific range of quantile levels. By introducing a class of weighted rank-based estimation procedure, our framework allows a localized quantile-based inference on the covariate effect with a less restrictive set of assumptions. Numerical studies demonstrate that the proposed estimator outperforms current alternatives under various settings in terms of smaller empirical bias and standard deviation. Finally, consistency and weak convergence of the proposed estimator are established via empirical process theory.

E0160: **Globally adaptive longitudinal quantile regression with high dimensional compositional covariates**
Presenter: **Huijuan Ma**, , China

A longitudinal quantile regression framework is proposed that enables a robust characterization of heterogeneous covariate-response associations in the presence of high-dimensional compositional covariates and repeated measurements of both response and covariates. We develop a globally adaptive penalization procedure, which can consistently identify covariate sparsity patterns across a continuum set of quantile levels. The proposed estimation procedure properly aggregates longitudinal observations over time. It ensures the satisfaction of the sum-zero coefficient constraint needed for the proper interpretation of the effects of compositional covariates. We establish the oracle rate of uniform convergence and weak convergence of the resulting estimators and further justify the proposed uniform selector of the tuning parameter to achieve global model selection consistency. We derive an efficient coordinate descent algorithm, where a maximization-minimization scheme is incorporated to facilitate stable and fast computation. We confirm our theoretical findings via extensive simulation studies. We apply the proposed method to a longitudinal study of cystic fibrosis children where the association between the gut microbiome and other diet-related biomarkers is of interest.

E0174: **Principal component analysis of hybrid functional and vector data**
Presenter: **Jeong Hoon Jang**, Indiana University, United States

A practical principal component analysis (PCA) framework is proposed that provides a nonparametric means of simultaneously reducing the dimensions of and modeling functional and vector (multivariate) data. We first introduce a Hilbert space that combines functional and vector objects as a single hybrid object. The framework, termed as PCA of hybrid functional and vector data (HFV-PCA), is then based on the eigen-decomposition of a covariance operator that captures simultaneous variations of functional and vector data in the new space. This approach leads to interpretable principal components with the same structure as each observation and a single set of scores that serves as a low-dimensional proxy for hybrid functional and vector data. To support the practical application of HFV-PCA, the explicit relationship between the hybrid PC decomposition and functional and vector PC decompositions is established, leading to a simple and robust estimation scheme where components of HFV-PCA are calculated using the components estimated from the existing functional and classical PCA methods. This estimation strategy allows the flexible incorporation of sparse and irregular functional data and multivariate functional data. We derive the consistency results and asymptotic convergence rates for the proposed estimators. We demonstrate the efficacy of the method through simulations and analysis of renal imaging data.

EO213 Room R04 ADVANCES IN FINITE MIXTURE MODELS: METHODOLOGY AND APPLICATIONS II Chair: Abbas Khalili

E0634: Uniform convergence rates for maximum likelihood estimation under two-component gaussian mixture models

Presenter: **Tudor Manole**, Carnegie Mellon University, United States
Co-authors: Nhat Ho

We derive uniform convergence rates for the maximum likelihood estimator and minimax lower bounds for parameter estimation in two-component location-scale Gaussian mixture models with unequal variances. We assume the mixing proportions of the mixture are known and fixed but make no separation assumption on the underlying mixture components. A phase transition is shown to exist in the optimal parameter estimation rate, depending on whether or not the mixture is balanced. Key to our analysis is a careful study of the dependence between the parameters of location-scale Gaussian mixture models, as captured through systems of polynomial equalities and inequalities whose solution set drives the rates we obtain.

E0719: Sparseness, consistency and model selection for Markov regime-switching autoregressives

Presenter: **Abbas Khalili**, McGill University, Canada
Co-authors: David Stephens

Markov regime-switching Gaussian autoregressive models, which aim to capture temporal heterogeneity exhibited by time series data, are studied. In constructing a Markov regime-switching model, several specifications must be made relating to both the state and observation models; in particular, the complexity of these models must be specified when fitting to a dataset. We propose new regularization methods based on the conditional likelihood for simultaneous autoregressive-order and parameter estimation with the number of regimes fixed and use a regularized Bayesian information criterion to select the number of regimes. Unlike the existing information-theoretic approaches, the new methods avoid an exhaustive search of the model space for model selection and thereby are computationally more efficient. We establish large sample properties of the proposed methods for estimation, model selection, and forecasting. We also evaluate the finite sample performance of the methods via simulations and illustrate their applications by analyzing a real dataset.

E0666: Bayesian hierarchical finite mixture of regression for histopathological imaging-based cancer data analysis

Presenter: **Yunju Im**, Yale University, United States
Co-authors: Yuan Huang, Jian Huang, Shuangge Ma

Cancer is heterogeneous. A finite mixture of regression (FMR) and other modeling techniques have been developed to accommodate such heterogeneity. “Classic” FMR analysis has usually been based on clinical, demographic, and molecular variables. More recently, histopathological imaging data – which is a byproduct of biopsy and hence enjoys broad data availability and high cost-effectiveness – has been increasingly used in cancer modeling. However, it is noted that its application to cancer FMR analysis remains limited. We further advance cancer FMR analysis based on histopathological imaging data. Significantly advancing from the existing analyses, our goal is to simultaneously use two types of histopathological imaging features, which are extracted based on domain-specific biomedical knowledge and using automated signal processing software, respectively. A significant modeling/methodological advancement is that we impose a hierarchy in the mixture structures to reflect the “increasing resolution” of the two types of imaging features. An effective and flexible Bayesian approach is proposed. Simulation shows its competitiveness over several highly relevant alternatives. The TCGA lung cancer data is analyzed, and interesting heterogeneous structures different from using the alternatives are found. Overall, this study provides an innovative new venue for FMR analysis for cancer and other complex diseases.

EO057 Room R05 COMPLEX DATA ANALYSIS USING MIXTURE MODELS AND EMPIRICAL LIKELIHOOD Chair: Suyeon Kang

E0590: Testing homogeneity in contaminated mixture models

Presenter: **Yuejiao Fu**, York University, Canada

Contaminated mixture models (CMMs) have wide applications in the real world. Testing homogeneity in the CMMs is an interesting and important research problem. We develop an EM-test for homogeneity in the general framework of the CMMs. The null limiting distribution of the test is shown to be a shifted mixture of chi-square distributions. Simulation studies demonstrate that the EM-test has excellent finite-sample performance. Two real-data examples illustrate the applications of the proposed method.

E0594: Statistical inference for normal mixtures with unknown number of components

Presenter: **Mian Huang**, Shanghai University of Finance and Economics, China

Statistical inference for normal mixture models with an unknown number of components has long been challenging due to non-identifiability, degenerated Fisher matrix, and boundary parameters. A penalized likelihood estimation procedure is proposed for mixtures of normals with an unknown number of components to achieve both the order selection consistency and the local root-n convergence rate for the component parameters estimators. We show that the proposed new estimator could avoid being trapped in certain degenerated regions of the non-identifiable subset of the parameter space for over-fitted normal mixture models so that a regular asymptotic quadratic Taylor expansion of the mixture log-likelihood could be derived. With a suitable penalty function on mixing proportions, the new estimator is consistent on the order selection and has an asymptotic normal distribution. Our derived sparsity conditions also reveal some surprising but interesting differences among some commonly used penalty functions and explain why the performance of some popularly used penalty functions, such as Lasso and SCAD, provide unsatisfactory results in the order selection of the mixture model.

E0727: Dimensionality reduction in mixtures of multivariate linear regression

Presenter: **Suyeon Kang**, University of California, Riverside, United States
Co-authors: Kun Chen, Weixin Yao

Motivated by the idea of reduced-rank estimation originally developed for non-mixture cases, we here study reduced rank mixtures of multivariate response regression models to provide more parsimonious and interpretable models. The proposed estimator simultaneously take into account the joint structure of the multivariate response and population heterogeneity. We show the complete derivation of iterative algorithms that perform parameter estimation in mixtures of multivariate response regression models with and without the reduced rank framework. Via the proposed paradigm, we have some desired features, such as the monotonicity of the penalized likelihood sequence. The consistency of the proposed estimators is established. The performances of the proposed reduced-rank methods are evaluated through simulation studies and real data analysis.

EO237 Room R06 NEW CHALLENGES IN COMPLEX DATA ANALYSIS Chair: Yichuan Zhao

E0518: **Non-crossing quantile regression for deep reinforcement learning**
Presenter: **Jianing Wang**, Shanghai University of Finance and Ecnomics, China
Co-authors: Fan Zhou, Xingdong Feng

Distributional reinforcement learning (DRL) estimates the distribution over future returns instead of the mean to more efficiently capture the intrinsic uncertainty of MDPs. However, batch-based DRL algorithms cannot guarantee the non-decreasing property of learned quantile curves, especially at the early training stage, leading to abnormal distribution estimates and reduced model interpretability. To address these issues, we introduce a general DRL framework by using non-crossing quantile regression to ensure the monotonicity constraint within each sampled batch, which can be incorporated with some well-known DRL algorithm. We demonstrate the validity of our method from both the theory and model implementation perspectives. Experiments on Atari 2600 Games show that some state-of-art DRL algorithms with the non-crossing modification can significantly outperform their baselines in terms of faster convergence speeds and better testing performance. In particular, our method can effectively recover the distribution information and thus dramatically increase the exploration efficiency when the reward space is extremely sparse.

E0543: **A variable selection method for the joint model of longitudinal and survival data with its application**
Presenter: **Tao Wang**, Yunnan Normal University, China

There has been extensive research for joint modelling methods of longitudinal and survival data in the last two decades motivated by the requirements of increasingly application, and the importance of such joint models has been increasingly recognized. However, the research on variable selection methods for joint models of longitudinal and survival outcomes with lower computational load is still getting on slowly. We propose a novel Bayesian variable selection method based on spike-and-slab lasso for semi-parametric joint models, consisting of a semi-parametric mixed-effects model for longitudinal data and a semi-parametric Cox proportional hazards model for survival data linked through shared random effects. We develop the computational program for such a variable selection method. Simulation studies and real data analysis demonstrate that our method performs well.

E0544: **Bayesian jackknife empirical likelihood**
Presenter: **Yichuan Zhao**, Georgia State University, United States

The empirical likelihood is a very powerful non-parametric tool that does not require any distributional assumptions. If the usual likelihood component in the Bayesian posterior likelihood is replaced with the empirical likelihood, then the posterior inference is still valid when the functional of interest is a smooth function of the posterior mean. However, it is not clear whether similar conclusions can be obtained for parameters defined in terms of U-statistics. We propose the so-called Bayesian jackknife empirical likelihood, which replaces the likelihood component with the jackknife empirical likelihood. We show, both theoretically and empirically, the validity of the proposed method as a general tool for Bayesian inference. Empirical analysis shows the small sample performance of the proposed method is better than its frequentist counterpart. Analysis of a case-control study for pancreatic cancer is used to illustrate the new approach.

EO093 Room R07 NEW FRONTIERS IN FUNCTIONAL DATA ANALYSIS Chair: Alexander Petersen

E0242: **Functional dimension reduction via average Fr'echet derivatives**
Presenter: **Kuang-Yao Lee**, Temple University, United States
Co-authors: Lexin Li

Sufficient dimension reduction (SDR) embodies a family of methods that aim to reduce dimensionality without loss of information in a regression setting. We propose a new method for nonparametric SDR, where both the response and the predictor are a function. We first develop the notions of functional central mean subspace and functional central subspace, which form the population targets of our functional SDR. We then introduce an average Fréchet derivative estimator, which extends the gradient of the regression function to the operator level, and use it to develop estimators for our functional dimension reduction spaces. We show that the resulting functional SDR estimators are unbiased and exhaustive, and more importantly, without imposing any distributional assumptions such as the linearity or the constant variance conditions commonly imposed by all existing functional SDR methods. We establish the uniform convergence of the estimators for the functional dimension reduction spaces while allowing both the number of Karhunen-Loève expansions and the intrinsic dimension to diverge with the sample size. We demonstrate the efficacy of the proposed methods through both simulations and a real data example.

E0335: **Forecasting of density functions with applications to cross-sectional and intraday returns**
Presenter: **Alexander Petersen**, Brigham Young University, United States
Co-authors: Han Lin Shang, Piotr Kokoszka, Hong Miao

The aim is the forecasting of probability density functions based on an observed density-valued functional time series. Density functions are nonnegative and have a constrained integral and thus do not constitute a vector space. The implementation of established functional time series forecasting methods for such nonlinear data is therefore problematic. Based on the log quantile density transformation and compositional data, two new methods are developed and compared to two existing methods. The comparison is based on the densities derived from cross-sectional and intraday returns. For such data, the log quantile density transformation is shown to dominate the existing methods. In contrast, the compositional method is comparable to an existing approach based on dynamic functional principal component analysis.

E0580: **Cross-component registration for multivariate functional data: Application to growth curves**
Presenter: **Alois Kneip**, University of Bonn, Germany
Co-authors: Cody Carroll, Hans-Georg Mueller

Multivariate functional data are becoming ubiquitous with the advance of modern technology and are substantially more complex than univariate functional data. We propose and study a novel model for multivariate functional data where the component processes are subject to mutual time warping. That is, the component processes exhibit a similar shape but are subject to mutual time-warping across their domains. To address this previously unconsidered mode of warping, we propose a new registration methodology based on a shift-warping model. The method differs from all existing registration methods for functional data in a fundamental way. Namely, instead of focusing on the traditional approach to warping, where one aims to recover individual-specific registration, we focus on shift registration across the components of a multivariate functional data vector on a population-wide level. The proposed estimates for these shifts are identifiable, enjoy parametric rates of convergence and often have intuitive physical interpretations, all in contrast to traditional curve-specific registration approaches. We demonstrate the implementation and interpretation of the proposed method by applying our methodology to the Zuerich Longitudinal Growth data and study its finite sample properties in simulations.

EO137 Room R08 RECENT ADVANCES IN COMPUTATIONAL AND METHODOLOGICAL STATISTICS Chair: Wanjie Wang

E0735: Graph matching beyond perfectly overlapping graphs

Presenter: **Wanjie Wang**, National University of Singapore, Singapore
Co-authors: Yi Yu, Yaofang Hu

Graph matching is a fruitful area in terms of both algorithms and theories. Suppose for the same set of nodes, two graphs are observed by different groups. Then how to match the nodes in the two graphs would be difficult. We exploit the degree information, which was previously used only in noiseless graphs and perfectly overlapping Erdos–Renyi random graphs matching. We are concerned with graph matching of partially overlapping graphs and stochastic block models, which are more useful in tackling real-life problems. We propose the edge exploited degree profile graph matching method and two refined variations. We conduct a thorough analysis of our proposed methods' performances in a range of challenging scenarios, including a zebrafish neuron activity data set and a coauthorship data set. Our methods are proved to be numerically superior to state-of-the-art methods.

E0420: Wasserstein distributionally robust optimization with local perturbations

Presenter: **Yongchan Kwon**, Stanford University, United States
Co-authors: Wonyoung Kim, Joong-Ho Won, Myunghee Cho Paik

Wasserstein distributionally robust optimization (WDRO) attempts to learn a model that minimizes the local worst-case risk in the vicinity of the empirical data distribution defined by Wasserstein ball. While WDRO has received attention as a promising tool for inference since its introduction, its theoretical understanding has not been fully matured. We propose a minimizer based on a novel approximation theorem and provide the corresponding risk consistency results. Furthermore, we develop WDRO inference for locally perturbed data that include the Mixup as a special case. We show that our approximation and risk consistency results naturally extend to the cases when data are locally perturbed. Numerical experiments demonstrate the robustness of the proposed method using image classification datasets. Our results show that the proposed method achieves significantly higher accuracy than baseline models on contaminated datasets.

E0423: Distributionally robust formulation of the graphical lasso

Presenter: **Sang-Yun Oh**, University of California, Santa Barbara, United States
Co-authors: Alexander Petersen, Pedro Cisneros-Velarde

Building on a recent framework for distributionally robust optimization, estimation of the inverse covariance matrix is considered for multivariate data. We provide a novel notion of a Wasserstein ambiguity set specifically tailored to this estimation problem, leading to a tractable class of regularized estimators. Special cases include penalized likelihood estimators for Gaussian data, specifically the graphical lasso estimator. As a consequence of this formulation, the radius of the Wasserstein ambiguity set is directly related to the regularization parameter in the estimation problem. Using this relationship, the level of robustness of the estimation procedure corresponds to the level of confidence with which the ambiguity set contains a distribution with the population covariance. Furthermore, the radius can be expressed in closed-form as a function of the ordinary sample covariance matrix. Taking advantage of this finding, we develop a simple algorithm to determine a regularization parameter for the graphical lasso, using only the bootstrapped sample covariance matrices, avoiding repeated evaluation of the graphical lasso algorithm during regularization parameter tuning, for example, with cross-validation. Finally, we numerically study the obtained regularization criterion and analyze the robustness of other automated tuning procedures used in practice.

Friday 25.06.2021 13:30 - 14:45 Parallel Session J – EcoSta2021

EO301 Room R01 SOCIAL ISSUES IN MACHINE LEARNING Chair: Yongdai Kim

E0751: **Within group fairness: A new concept for better between group fairness**
Presenter: **Yongdai Kim**, Seoul National University, Korea, South

As they have a vital effect on social decision-making, AI algorithms should be accurate and should not impose unfairness against certain sensitive groups. Various specially designed AI algorithms to ensure trained AI models to be fair between sensitive groups have been developed. We raise a new issue that between-group fair AI models could treat individuals in the same group unfairly. We introduce a new concept of fairness, so-called within-group fairness, which requires that AI models be fair for those in the same sensitive group and those in different sensitive groups.

E0752: **Meta ANOVA: A model-agnostic tool for interpretable machine learning**
Presenter: **Dongha Kim**, Sungshin Women's University, Korea, South
Co-authors: Yongdai Kim, Yongchan Choi

There are two things to be considered when we evaluate predictive models. One is prediction accuracy, and the other is interpretability. Over the recent decades, many predictive models of high prediction ability, such as ensemble-based models and deep neural networks, have been developed. But, these models are usually too complex, so it is hard to interpret their resulted predictions intuitively, and this limitation prevents them from being applied in many real-world fields that require accountability (e.g., medicine, finance, college admission, etc.). We develop a novel method called Meta ANOVA to provide an interpretable model for any given predictive model. Meta ANOVA first identifies significant input variables and high-order interactions without constructing any model. Then, it constructs a model only with identified input variables and high-order interactions that closely approximates a given predictive model. Meta ANOVA is a model agnostic algorithm and hence can be applied to any predictive models, including ensemble models and deep neural networks. By carrying out various experiments with synthetic and real data sets, we empirically demonstrate the superiority of Meta ANOVA.

E0753: **Challenges in making privacy protected data for public use**
Presenter: **Yonghee Lee**, University of Seoul, Korea, South

Preserving privacy when data are disseminated for public use has been an important issue for a long time. Nowadays, privacy is more important than ever since big data and machine learning demand more and more data. We will review some experience and practice in statistical methods for disclosure limitations that research groups in Korea have considered. Also, we will introduce recent issue and challenge in privacy protection for data dissemination and synthetic data for public use in Korea.

EO065 Room R02 SOME RECENT DEVELOPMENT ON TREATMENT EFFECT STUDY Chair: Wei Luo

E0713: **Doubly robust estimation for conditional treatment effect**
Presenter: **Chuyun Ye**, Beijing Normal University, China
Co-authors: Keli Guo, Lixing Zhu

A doubly robust approach is applied to estimate, when some covariates are given, the conditional average treatment effect under the parametric, semiparametric and nonparametric structure of the nuisance propensity score and outcome regression models. We then conduct a systematic study on the asymptotic distributions of nine estimators with different combinations of estimated propensity score and outcome regressions. The asymptotic properties with all models correctly specified are considered, with either propensity score or outcome regressions locally / globally misspecified; and with all models locally / globally misspecified. The asymptotic variances are compared, and the asymptotic bias correction under model-misspecification is discussed. The phenomenon that the asymptotic variance, with model-misspecification, could sometimes be even smaller than that with all models correctly specified is explored. We also conduct a numerical study to examine the theoretical results.

E0745: **On efficient dimension reduction with respect to the interaction between two response variables**
Presenter: **Wei Luo**, Zhejiang University, China

Theory and methodologies for dimension reduction with respect to the interaction between two response variables are proposed. This is crucial for effective dimension reduction in applications such as missing data analysis, causal inference, and graphical models. We introduce the concepts of the locally and the globally efficient dimension reduction subspaces, which induce reduced predictors that preserve the key feature for subsequent data analysis. These spaces can be low dimensional when neither of the two individual response variables is equipped with low-dimensional data structures, for which they cannot be recovered by the existing dimension reduction applications in general. Based on the existing inverse regression methods, we propose a family of dimension reduction methods called the dual inverse regression methods, which consistently estimate the locally efficient dimension reduction subspaces under mild assumptions and consistently estimate the globally efficient dimension reduction subspace when it exists. These methods are also easily implementable. In addition, we propose a sufficient and necessary condition for the existence of the globally efficient dimension reduction subspace that is handy to check. We illustrate the usefulness of the proposed dual inverse regression methods by simulations studies and a real data example at the end.

E0635: **A Neyman-type orthogonality-based significance test for structure functions**
Presenter: **Niwen Zhou**, Beijing Normal University, China

Structure functions are defined as a conditional expectation given covariates in which the response may contain unknown nuisance parameters. This includes classic regression functions and the conditional average treatment effects (CATE) as examples. We investigate the hypothesis testing problem that checks whether part of covariates significantly affects structure functions. As plug-in estimation of the unknown nuisance parameter could cause the limiting null distribution complicated, which leads to a more complicated or difficult procedure for critical value determination. Thus, we propose using the Neyman-type approach to select a representation of $Y(\eta)$ such that the conditional expectation can be orthogonal to the nuisance function. This can make the quasi-oracle property of the constructed distance-based test in the sense that the nuisance function asymptotically has no impact on the limiting distributions of the test statistic under both the null and alternatives. Further, the new test can still detect the local alternatives distinct from the null at the fastest possible rate in hypothesis testing. However, the conditional expectation is estimated by a nonparametric method. Numerical studies are conducted to examine the performance of the test, and the analysis for a real data example is implemented for illustration.

EO129 Room R04 FINTECH AND BIG DATA ANALYTICS Chair: Mike So

E0529: **Estimation of high dimensional vector autoregression via sparse precision matrix**

Presenter: **Manabu Asai**, Soka University, Japan

Co-authors: Benjamin Poignard

The problem of estimating sparse structural vector autoregression (SVAR) processes is considered. To do so, using the SCAD, MCP and LASSO regularisers, we rely on a sparse precision matrix within a general Bregman divergence setting, whose components characterize the SVAR coefficients. Under suitable regularity conditions, we derive error bounds for the regularised precision matrix for each Bregman divergence case. Moreover, we establish the support recovery property, including the case when the regulariser is non-convex. Empirical studies support these theoretical results.

E0624: **Hybrid resampling confidence intervals for change-point or stationary high-dimensional stochastic regression models**

Presenter: **Wei Dai**, The Chinese University of Hong Kong, Shenzhen, China

Hybrid resampling is used to address (a) the long-standing problem of inference on change times and changed parameters in change-point ARX-GARCH models, and (b) the challenging problem of valid confidence intervals, after variable selection under sparsity assumptions, for the parameters in linear regression models with high-dimensional stochastic regressors and asymptotically stationary noise. For the latter problem, we introduce consistent estimators of the selected parameters and a resampling approach to overcome the inherent difficulties of post-selection confidence intervals. For the former problem, we use a sequential Monte Carlo for the latent states (representing the change times and changed parameters) of a hidden Markov model. Asymptotic efficiency theory and simulation, and empirical studies demonstrate the advantages of the proposed methods.

E0641: **Efficient Bayesian updating in network models and its applications to financial risk management**

Presenter: **Thomas Chan**, Hong Kong University of Science and Technology, Hong Kong

Co-authors: Mike So

In the statistical inference of financial networks, large datasets with high-frequency updating are often considered. We study a dynamic network model to assess systemic risk in financial markets. Bayesian updating is adopted to incorporate new information in network modeling efficiently. We demonstrate how to efficiently conduct prediction and update estimates of parameters and latent variables with posterior inference. We adopt the proposed Bayesian updating to network modeling of financial returns of listed companies in Hong Kong to demonstrate our ideas.

EO159 Room R05 INSTABILITIES AND IRREGULARITIES IN VARIOUS DATA STRUCTURES Chair: Matus Maciak

E0232: **Changepoint and measurement errors**

Presenter: **Michal Pesta**, Charles University, Faculty of Mathematics and Physics, Czech Republic

Linear relations, containing measurement errors in input and output data, are considered. Parameters of these so-called errors-in-variables models can change at some unknown moment. The aim is to test whether such an unknown change has occurred or not. For instance, detecting a change in trend for a randomly spaced time series is a special case of the investigated framework. The designed changepoint tests are shown to be consistent and involve neither nuisance parameters nor tuning constants, which makes the testing procedures effortlessly applicable. A changepoint estimator is also introduced, and its consistency is proved. A boundary issue is avoided, meaning that the changepoint can be detected when being close to the extremities of the observation regime. As a theoretical basis for the developed methods, a weak invariance principle for the smallest singular value of the data matrix is provided, assuming weakly dependent and non-stationary errors. The results are presented in a simulation study, which demonstrates the computational efficiency of the techniques. The completely data-driven tests are illustrated through a calibration problem. However, the methodology can be applied to other areas such as clinical measurements, dietary assessment, computational psychometrics, or environmental toxicology.

E0295: **Testing constancy of correlation matrix based on self-normalization method**

Presenter: **Ji Eun Choi**, Pukyong National University, Korea, South

Co-authors: Dong Wan Shin

A new test for correlation matrix break is constructed based on the self-normalization method. The self-normalization test has a practical advantage over the existing test: easy and stable implementation not having the singularity issue and the bandwidth selection issue of the existing test, remedying size distortion problem of the existing test under (near) singularity, serial dependence, conditional heteroscedasticity or unconditional heteroscedasticity, having reasonable power. These advantages are demonstrated experimentally by a Monte-Carlo simulation and theoretically by showing no need to estimate the complicated covariance matrix of the sample correlations. We establish asymptotic null distribution and consistency of the self-normalization test. We apply the correlation matrix break tests to the stock log-returns of the companies of the 10 largest weight of the NASDAQ 100 index and five volatility indexes for options on individual equities.

E0297: **Implied volatility surface estimation via quantile regularization**

Presenter: **Matus Maciak**, Charles University, Czech Republic

Co-authors: Michal Pesta, Sebastiano Vitali

The implied volatility function and the implied volatility surface are both fundamental tools for analyzing financial and derivative markets. Still, various theoretical, practical, and computational limits occur in most of them. An innovative estimation method based on the idea of sparse estimation and atomic pursuit approach is introduced to overcome some of these limits: the quantile LASSO estimation implies robustness with respect to common market anomalies; the panel data structure allows for time-dependent modeling; linear constraints ensure the arbitrage-free validity; last but not least, the interpolated implied volatility values overcome the problem of changing maturity when observing implied volatility over time. Standard theoretical properties are derived, and the idea of interpolated volatilities is introduced. The finite sample properties are investigated using a simulation study. The whole methodology is applied to estimate the implied volatility surface of the Erste Group Bank AG call options quoted in the EUREX Deutschland market.

EO025 Room R06 STATISTICAL METHODS FOR FUNCTIONAL DATA Chair: Masaaki Imaizumi

E0428: **Fast convergence on perfect classification for functional data**
Presenter: **Masaaki Imaizumi**, The University of Tokyo, Japan
Co-authors: Tomoya Wakayama

The capacity of approaching perfect classification for functional data with finite samples is investigated. A perfect classifier for functional data has been shown to be easier to define than for finite-dimensional data. This result is based on a sufficient condition for the existence of a perfect classifier, named a Delaigle-Hall (DH) condition, which is only available for functional data. However, there is a danger that large sample size is required to achieve the perfect classification even though the DH condition holds because the convergence of misclassification errors of functional data is significantly slow. Specifically, a minimax rate of the convergence of errors with functional data has a logarithm order in the sample size. This complication is solved by proving that the DH condition also achieves fast convergence of the misclassification error. Therefore, we study a classifier with empirical risk minimization using reproducing kernel Hilbert space (RKHS) and analyse its convergence rate under the DH condition. The result shows that the convergence speed of the misclassification error by the RKHS classifier has an exponential order in sample size. Experimentally, we validate that the DH condition and the associated margin condition have a certain impact on the convergence rate of the RKHS classifier.

E0706: **Nonparametric density-on-density regression**
Presenter: **Han Lin Shang**, Macquarie University, Australia
Co-authors: Frederic Ferraty

The focus is on forecasting probability density functions. Density functions are nonnegative and have a constrained integral and thus do not constitute a vector space. Implementation of unconstrained functional time-series forecasting methods is therefore problematic for such nonlinear and constrained data. Under the framework of compositional data analysis, a novel forecasting method is developed based on a nonparametric function-on-function regression, where the response and the predictor are both probability density functions. Through a series of Monte-Carlo simulation studies, we evaluate the finite-sample performance of our nonparametric regression. Using monthly cross-sectional returns, intraday high-frequency returns, and age-specific period life-tables, we assess and compare finite-sample forecast accuracy between the proposed and several existing methods.

E0663: **On optimal prediction of missing functional data with memory**
Presenter: **Germain Van Bever**, Universite de Namur, Belgium
Co-authors: Pauliina Ilmonen, Lauri Viitasaari, Nourhan Shafik, Tommi Sottinen

The problem of reconstructing missing parts of functions based on their observed segments is considered. For Gaussian processes and arbitrary bijective transformations thereof, we provide theoretical expressions for the L2 -optimal reconstruction of the missing parts. These functions are obtained as solutions of explicit integral equations. In the discrete case, approximations of the solutions provide consistent expressions of all missing values of the processes. In Gaussian processes with a parametric covariance structure, the estimation can be conducted separately for each function and yields nonlinear solutions in the presence of memory. Simulated examples show that the proposed reconstruction indeed fares better than the conventional interpolation methods in various situations.

EO193 Room R07 RECENT ADVANCES IN CAUSAL INFERENCE Chair: Linbo Wang

E0731: **The promises of parallel outcomes**
Presenter: **Dehan Kong**, University of Toronto, Canada

Unobserved confounding presents a major threat to the validity of causal inference from observational studies. We introduce a novel framework that leverages the information in multiple parallel outcomes for the identification and estimation of causal effects. Under a conditional independence structure among multiple parallel outcomes, we achieve nonparametric identification with at least three parallel outcomes. We further show that under a set of linear structural equation models, causal inference is possible with two parallel outcomes. We develop accompanying estimating procedures and evaluate their finite sample performance through simulation studies and a data application studying the causal effect of the tau protein level on various types of behavioral deficits.

E0701: **Causal inference on non-linear spaces: Distribution functions and beyond**
Presenter: **Zhenhua Lin**, National University of Singapore, Singapore
Co-authors: Dehan Kong, Linbo Wang

Understanding causal relationships is one of the most important goals of modern science. So far, the causal inference literature has focused almost exclusively on outcomes coming from a linear space, most commonly the Euclidean space. However, it is increasingly common that complex datasets collected through electronic sources, such as wearable devices and medical imaging, cannot be represented as data points from linear spaces. We will present a formal definition of causal effects for outcomes from non-linear spaces, focusing on the Wasserstein space of cumulative distribution functions. Doubly robust estimators and associated asymptotic theory for these causal effects will be developed. The proposed framework extends to outcomes from certain Riemannian manifolds. As an illustration, we will use the framework to quantify the causal effect of marriage on physical activity patterns using wearable device data collected through the National Health and Nutrition Examination Survey.

E0687: **Robust estimation of treatment effects in a latent variable framework**
Presenter: **Mikhail Zhelonkin**, Erasmus University Rotterdam, Netherlands

Policy evaluation is one of the central problems in modern economics. Unfortunately, it is often impossible to perform randomized experiments in order to evaluate the treatment effects. Hence, the data from observational studies has to be used. In this case, the sample is typically non-random, and one has either to correct for selectivity or impose (conditional) independence assumption. Since this assumption is often irrealistic, the structural latent variable model is used. Although straightforward to compute and interpret, the parametric estimators have been criticized for sensitivity to the departures from the distributional assumptions. The alternative semi- and nonparametric estimators have complex identification and are limited to estimation of a certain parameter(s) of interest but do not allow for the general evaluation and interpretation of the model. We employ the latent-variable framework. We study the robustness properties of the estimators of three principal parameters: average treatment effect, the average treatment effect on the treated and local average treatment effect, and propose robust alternatives.

Friday 25.06.2021 14:55 - 16:35 Parallel Session K – EcoSta2021

EO027 Room R01 RECENT ISSUES IN ECONOMETRICS Chair: Jaeyoung Kim

E0673: **Jackknife GMM with many weak moment conditions**
Presenter: **Hyojin Han**, Hanyang University, Korea, South

Using many moment conditions can improve efficiency but also increases estimation bias. In order to address the problem of increasing bias, continuously updated estimator (CUE) and other generalized empirical likelihood estimators (GEL) have been proposed. However, they both require the covariance matrix to be bounded while the moment conditions may become highly correlated, and the covariance matrix can become singular under many moments. We consider a Jackknife GMM estimator with a nonsingular weighting matrix that is a regularized version of the inverse of the covariance matrix with a penalization term. We provide the consistency and asymptotic normality results of this estimator with sufficient conditions for them. The results can be extended to the cases where a continuum of moment conditions arise. Examples in that our results can be useful include models with conditional moment restrictions and linear IV models with many weak instruments.

E0700: **Exploring systematic risk through high-frequency panel regressions**
Presenter: **Ji Hyung Lee**, University of Illinois at Urbana-Champaign, United States
Co-authors: Torben Andersen, Viktor Todorov

The conditional expectation of systematic risk is studied using high-frequency asset return data. We model conditional expectation of systematic risk through linear regressions with various economic conditioning variables. An interesting nonstandard limit theory arises from the measurement errors of the systematic risk, which is estimated from intraday return data. We provide consistency and asymptotic normality that are strikingly different from conventional dynamic panel regressions.

E0705: **Estimation of a level shift in panel data with fractionally integrated errors**
Presenter: **Seong Yeon Chang**, Soongsil University, Korea, South

The focus is on the estimation of a common breakpoint in panel data. We consider the general case of fractionally integrated errors with memory parameter d in $(0.5, 0.5)$ and establish the consistency, convergence rate, and limiting distribution of the estimated common breakpoint. The ordinary least squares method is used for estimating the breakpoint in mean. We find that the convergence rate is invariant to the order of fractional integration. Simulation experiments are provided to illustrate some of the theoretical results.

E0709: **Optimal dynamic treatment regimes and partial welfare ordering**
Presenter: **Sukjin Han**, University of Bristol, United Kingdom

Dynamic treatment regimes are treatment allocations tailored to heterogeneous individuals. The optimal dynamic treatment regime is a regime that maximizes counterfactual welfare. We introduce a framework in which we can partially learn the optimal dynamic regime from observational data, relaxing the sequential randomization assumption commonly employed but instead using (binary) instrumental variables. We propose the notion of sharp partial ordering of counterfactual welfares with respect to dynamic regimes and establish a mapping from data to partial ordering via a set of linear programs. We then characterize the identified set of the optimal regime as the set of maximal elements associated with the partial ordering. We relate the notion of partial ordering with a more conventional notion of partial identification using topological sorts. Practically, topological sorts can be served as a policy menu for a policymaker. The framework can be applied beyond the current context, e.g., in establishing rankings of multiple treatments or policies across different counterfactual scenarios.

EO295 Room R02 PERSPECTIVES OF STATISTICS FOR STOCHASTIC PROCESSES Chair: Yuta Koike

E0561: **Stepwise model comparison for ergodic SDEs**
Presenter: **Shoichi Eguchi**, Osaka Institute of Technology, Japan
Co-authors: Hiroki Masuda

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. We consider pure-jump Lévy noise-driven SDEs as the candidate models and propose the AIC-type information criterion the stepwise model selection procedure.

E0281: **Marked Hawkes process and sparse estimation**
Presenter: **Masatoshi Goda**, University of Tokyo, Japan

The P-O (penalized method to ordinary method) estimator has the oracle properties and the polynomial convergence rate of selection consistency under suitable conditions. We extend this method to the case where there could be nuisance parameters, and the true value could be realized at the boundary of the parameter space. Moreover, we apply the method to a class of multivariate marked Hawkes processes with generalized exponential kernels.

E0401: **Moment convergence of the generalized maximum composite likelihood estimators for determinantal point processes**
Presenter: **Kou Fujimori**, Shinshu University, Japan
Co-authors: Yasutaka Shimizu, Sota Sakamoto

The maximum composite likelihood estimator for parametric models of determinantal point processes (DPPs) is discussed. Since the joint intensities of these point processes are given by determinant of positive definite kernels, we have the explicit form of the joint intensities for every order. This fact enables us to consider the generalized maximum composite likelihood estimator for any order. We introduce the two-step generalized composite likelihood estimator and show the moment convergence of the estimator under stationarity. Moreover, our results can yield information criteria for statistical model selection within DPPs.

E0564: **Drift estimation for a multi-dimensional diffusion process using deep neural networks**
Presenter: **Yuta Koike**, University of Tokyo, Japan
Co-authors: Akihiro Oga

Recently, many studies have shed light on the high adaptivity of deep neural network-based estimators in the framework of nonparametric regression. In particular, their superior performance has been established for various multivariate function classes. Motivated by this development, we propose estimating the drift coefficient of a multi-dimensional diffusion process by deep neural networks from its discrete observation data. Then, we derive their generalization error bounds and show that they achieve the minimax estimation rate up to a logarithmic factor.

EO081 Room R03 NEW MODELING AND ROBUST MODELING FOR VERSATILE DATA Chair: Catherine Liu

E0681: **FaMGLM: Factor analysis for a matrix-variated generalized linear model**
Presenter: **Yuzhe Zhang**, University of Science and Technology of China, China
Co-authors: Xu Zhang, Hong Zhang, Catherine Liu

Biomedical images act as an important biomarker role associated with various clinical traits in diagnostics of clinical trial and modern medical studies. The ultrahigh dimensionality of medical imaging data poses an unprecedented challenge to many classical statistical models and methods and have attracted ascending research interest in the past decade. Image data are mathematically considered as matrix variates. A major drawback when dealing with matrix-variated covariates within a tensor regression setting is the over-dimensional reduction and the loss of intrinsic data information. Rather than the regularization spirit, we propose a new generalized linear model with matrix-variated covariates based on factor analysis, named FaMGLM, as the working model of the conventional GLM model, including the matrix-variated covariate. The FaMGLM decomposes matrix-variated covariates into latent factor matrices based on factorization techniques and naturally enjoys a good interpretation of loading matrices. In numerical analysis, we compared our FaMGLM with those tensor regression and regularization methods in classification and prediction under various specific GLM models, including logistic regression, Poisson log-linear model and linear regression. We demonstrate a better discriminant power via real data analysis on the COVID-19 CT image dataset.

E0654: **Semi-parametric Bayesian inference to linear transformation model with censored data**
Presenter: **Chong Zhong**, The Hong Kong Polytechnic University, Hong Kong
Co-authors: Catherine Liu, Junshan Shen

A robust and flexible semi-parametric Bayesian inference procedure is developed for the linear transformation model that allows both error distribution and transformation function are unspecified under quite mild assumptions. Existing literature used to identify the model first and then develop statistical inference. The proposed procedure tackles the estimation of nonparametric and parametric components unified and simultaneously. It is implemented by choosing proper Gamma processes through I-spline functions and Dirichlet process mixture with Weibull kernel centering at a direct product of Gamma distribution as the priors of the unknown transformation function and the unspecific error term distribution, respectively. The MCMC sampler is built based on Hamiltonian Monte Carlo (HMC) and No-U-Turn sampler (NUTS) by Stan. We compare the proposed procedure with the method implemented by spBayesSurv R package.

E0734: **Robust Bayesian analysis based on trimmed mean regression**
Presenter: **Lulu Zhang**, The Hong Kong Polytechnic University (PolyU), Hong Kong

The use of Bayesian statistics in the social sciences is becoming increasingly widespread. However, seemingly high entry costs still keep many applied researchers from embracing Bayesian methods. Next to a lack of familiarity with the underlying conceptual foundations, the need to implement statistical models using specific programming languages remains one of the biggest hurdles. We will investigate Bayesian and robust Bayesian estimation of a wide range of parameters of interest in the context of Bayesian nonparametric under a broad class of trimmed mean regression and quantile regression. Dealing with uncertainty regarding the prior, we consider the Dirichlet and provide an explicit form of the resulting robust Bayes estimator.

E0677: **Feature screening for generalized linear model with network structure**
Presenter: **Xiangeng Fang**, The Hong Kong Polytechnic University, Hong Kong
Co-authors: Xu Zhang, Sheng Xu, Xuening Zhu, Catherine Liu

Feature screening is an essential step to deal with ultrahigh-dimensional data. When there is an auxiliary network structure, and the response is not continuous, it poses great challenges to developing new methodology and algorithms for dependent ultrahigh-dimensional data. There are sporadic works that either tackle dependent responses with marginal utility or dependent covariates but independent responses. We propose a new generalized linear autoregressive model which incorporates the network structure. We employ the pseudo-likelihood to do joint feature screening and establish a feasible algorithm. Theoretical results are discussed. The finite sample performance of the proposed method is assessed by a simulation study and illustrated by an empirical analysis of a dataset from the Chinese stock market.

EO029 Room R04 RECENT ADVANCES IN BAYESIAN INFERENCE III Chair: Antonio Lijoi

E0415: **Towards data imputation of large observations via Bayesian inference for multivariate extremes**
Presenter: **Isadora Antoniano-Villalobos**, Ca' Foscari University of Venice, Italy
Co-authors: Simone Padoan

Missing data is a known issue in statistics. In many environmental applications, the greatest interest is placed on large observations, e.g. of pollution levels, wind speed, precipitation or temperature, to name a few. In such contexts, usual data imputation methods may fail to reproduce the heavy tail behaviour of the quantities involved. Recent literature has proposed using multivariate extreme value theory to predict an unobserved component of a random vector given large observed values of the rest. This is achieved by estimating the angular measure controlling the dependence structure in the tail of the distribution. The idea can be used for effective data imputation at adequately large levels, provided that the model used for the angular measure is flexible enough to capture complex dependence structures. A Bayesian nonparametric model based on constrained Bernstein polynomials ensures such flexibility while allowing for tractable inference. An additional advantage of this approach is the natural way uncertainty about the estimation is incorporated into the imputed values through the Bayesian paradigm.

E0566: **A Bayesian survival model for time-varying coefficients and unobserved heterogeneity**
Presenter: **Peter Knaus**, WU Vienna University of Economics and Business, Austria
Co-authors: Daniel Winkler

Two sources of heterogeneity are often overlooked. On the one hand, time-varying hazard contributions of explanatory variables cannot be captured in the widely used Cox proportional hazard model. To this end, a dynamic survival model is investigated within a Bayesian framework. The specification allows parameters to evolve over time, thus accounting for time-varying effects gradually. On the other hand, unobserved heterogeneity across groups is often ignored, leading to invalid estimators. Accounting for such effects is made feasible for even large numbers of groups through a shared factor model, which picks up unexplained covariance in the error term. Building on a Markov Chain Monte Carlo scheme based on data augmentation allows the usage of shrinkage priors to avoid overfitting in such a highly parameterized model. In particular, the shrinkage priors are implemented to automatically detect which parameters should be included in the model and which should be allowed to vary over time. Finally, an R package that makes the routine easily available is introduced.

E0470: **Distributed methods for Bayesian regression: Contraction rate & uncertainty quantification**
Presenter: **Mohamed Amine Hadji**, Leiden University, Netherlands
Co-authors: Botond Szabo

In Bayesian regression, Gaussian processes have been widely used as priors in practice. Arguably, some of the most popular covariance kernels are the Matern and the squared exponential. However, these priors do not generally lead to scalable Bayesian methods and, therefore, are highly impractical for large data sets. In order to solve this issue, distributed methods, where the data of size n is divided over m machines, have been developed. After having processed the data locally, the machines collectively help to compute a global posterior. The main interest is to study the behavior of this global posterior, considering that the data is generated via a "true" underlying functional parameter. The two main problems we investigate are the rate at which the posterior contracts around the truth in L, and the reliability of uncertainty quantification, more precisely, the coverage properties of the corresponding credible sets.

E0550: **Variational Bayes for regression with Gaussian process priors: A frequentist Bayesian analysis**
Presenter: **Dennis Nieman**, VU Amsterdam, Netherlands
Co-authors: Harry Zanten, Botond Szabo

A Bayesian nonparametric regression model with Gaussian process priors is considered. In practice, sampling from the exact posterior distribution is computationally expensive. We study an approximative procedure called the variational method, which reduces computation time. Of particular interest is a variational framework that has gained popularity in the machine learning literature in the last decade. We investigate frequentist properties of the Bayesian approach: the data are assumed to be generated from a distribution with a true functional parameter, and conditions are given under which the contraction rate does not deteriorate under the approximation. The developed theory is applied to several examples of Gaussian process priors.

EO271 Room R05 THEORY OF MACHINE LEARNING AND DEEP LEARNING Chair: Masaaki Imaizumi

E0358: **Nonparametric link regression and its theoretical properties**
Presenter: **Akifumi Okuno**, The Institute of Statistical Mathematics, Japan
Co-authors: Keisuke Yano, Hidetoshi Shimodaira

Given d-dimensional data vectors and their link weights, i.e., strengths of graph-structured associations represented by a weighted adjacency matrix, we consider predicting link weights through the corresponding pair of data vectors. We call the problem as link regression (LR), where it reduces to link prediction if the weights are binary. We first apply nonparametric regression methods to the LR setting. Through numerical experiments and theoretical analysis, we also report that their asymptotic behaviors are different depending on the assumed design on the data vectors, i.e., random design and fixed design, unlike the classical regression methods.

E0394: **Theoretical analysis of graph classification problem**
Presenter: **Takanori Maehara**, RIKEN, Japan

The graph classification problem is a supervised learning problem in which the inputs are graphs, and the outputs are labels. This problem is one of the most actively studied problems in recent years. Several approaches have been proposed for this problem. A classical approach is graph kernels, which defines a kernel function on graphs, and a recent approach is graph neural networks, which propagates and aggregates information over graphs. Both of these approaches have been succeeded in many benchmark datasets and real-world problems. However, their theoretical properties are still not well understood. We aim at establishing a statistical learning theory for the graph classification problem from the viewpoint of graph limit theory.

E0456: **Improving the trainability of deep neural networks: A perspective from the infinite width limit**
Presenter: **Ryo Karakida**, National Institute of Advanced Industrial Science and Technology (AIST)), Japan

Deep neural networks (DNNs) have performed excellently in various practical applications, but we still have many heuristics and arbitrariness been in their training. An interesting direction will be to reveal the hidden mechanics of training dynamics and uncover efficient training methods. The infinite width limit is a promising approach to analyze such training dynamics of DNNs. Under the infinite width limit with some rational assumptions, one can see a DNN as a linearized model. It enables us to analyze the convergence of training dynamics to a global minimum. Based on this approach, we reveal appropriate settings of gradient-based algorithms. First, we estimate the proper size of the learning rate for convergence. Eigenvalues of the Fisher information matrix play a fundamental role in this estimation. Next, we show that some normalization methods make the training less sensitive to the choice of the learning rate and improve the trainability. We will also discuss some other related topics on analyses in the infinite width limit.

E0642: **Asymptotic risk of overparameterized likelihood models: Double descent theory for deep neural networks**
Presenter: **Ryumei Nakada**, Rutgers University, United States
Co-authors: Masaaki Imaizumi

The asymptotic risk of a general class of overparameterized likelihood models, including deep models, is analyzed. The recent empirical success of large-scale models has motivated several theoretical studies to investigate a scenario wherein both the number of samples n and parameters p diverge to infinity and derive an asymptotic risk at the limit. However, these theorems are only valid for linear-in-feature models, such as generalized linear regression, kernel regression, and shallow neural networks. Hence, it is difficult to investigate a wider class of nonlinear models, including deep neural networks with three or more layers. We consider a likelihood maximization problem without the model constraints and analyze the upper bound of an asymptotic risk of an estimator with penalization. Technically, we combine a property of the Fisher information matrix with an extended Marchenko-Pastur law and associate the combination with empirical process techniques. The derived bound is general, as it describes both the double descent and the regularized risk curves, depending on the penalization. Our results are valid without the linear-in-feature constraints on models and allow us to derive the general spectral distributions of a Fisher information matrix from the likelihood. We demonstrate that several explicit models, such as parallel deep neural networks, ensemble learning, and residual networks, agree with our theory.

EO273 Room R07 MODELING SPATIAL DATA WITH COMPLEX DEPENDENCE STRUCTURES Chair: Pavel Krupskiy

E0233: **Global wind modeling with transformed Gaussian processes**
Presenter: **Jaehong Jeong**, Hanyang University, Korea, South

Uncertainty quantification of wind energy potential from climate models can be limited because it requires considerable computational resources and is time-consuming. We propose a stochastic generator that aims at reproducing the data-generating mechanism of climate ensembles for global annual, monthly, and daily wind data. Inferences based on a multi-step conditional likelihood approach are achieved by balancing memory storage and distributed computation for a large data set. In the end, we discuss a general framework for modeling non-Gaussian multivariate stochastic processes by transforming underlying multivariate Gaussian processes.

E0357: **Estimating high-resolution Red sea surface temperature hotspots, using a low-rank semiparametric spatial model**
Presenter: **Raphael Huser**, King Abdullah University of Science and Technology, Saudi Arabia
Co-authors: Arnab Hazra

Extreme sea surface temperature (SST) hotspots, i.e., high threshold exceedance regions, is estimated, for the Red Sea, a vital region of high biodiversity. We analyze high-resolution satellite-derived SST data comprising daily measurements at 16703 grid cells across the Red Sea over the period 1985-2015. We propose a semiparametric Bayesian spatial mixed-effects linear model with a flexible mean structure to capture spatially-varying trend and seasonality, while the residual spatial variability is modeled through a Dirichlet process mixture (DPM) of low-rank spatial Student-t processes (LTPs). Specifying cluster-specific parameters for each LTP mixture component, the bulk of the SST residuals influence tail inference and hotspot estimation only moderately. Our proposed model has a nonstationary mean, covariance and tail dependence, and posterior inference can be drawn efficiently through Gibbs sampling. In our application, we show that the proposed method outperforms some natural parametric and semiparametric alternatives. Moreover, we show how hotspots can be identified, and we estimate extreme SST hotspots for the whole Red Sea, projected for the year 2100. The estimated 95% credible region for joint high threshold exceedances include large areas covering major endangered coral reefs in the southern Red Sea.

E0497: **Spatio-temporal forecast by spatial VAR with co-integration and functional PCA on empirical dynamic quantile series**
Presenter: **Guoqi Qian**, The University of Melbourne, Australia
Co-authors: Xianbin Cao

A spatial auto-regression co-integration (SARC) model, combined with techniques of empirical dynamic quantile (EDQ) series and functional principal component analysis (FPCA), is proposed to analyze time-series data observed from multiple spatial locations and to provide spatio-temporal forecast. The data in consideration are often temporally non-stationary, spatially correlated, and high-dimensional in both time and space, posing statistical and computational challenges in analyzing them. Using the proposed method, we tackle these challenges by first applying the SARC model to a small number of EDQ series of the data so that the residual vector time series from the model become stationary and uncorrelated. We then apply FPCA to these residual vector series to forecast the EDQ values at a future time. Finally, we extrapolate the EDQ series forecasts to the whole space domain according to the estimated quantile levels of the corresponding spatial locations, resulting in further spatio-temporal forecasts in the intended space-time domain. The proposed SARC-EDQ-FPCA method is evaluated using simulated data before being applied to analyze a real spatio-temporal data set on a landslide.

E0520: **A model-free subsampling method based on minimum energy criterion**
Presenter: **Wenlin Dai**, Renmin University of China, China

A novel approach, termed adaptive subsampling, is proposed that is based on the minimum energy criterion (ASMEC). The proposed method requires no explicit model assumptions and smartly incorporates information on covariates and responses. ASMEC subsamples possess two desirable properties: space-filling and spatial adaptiveness to the full data. We investigate the theoretical properties of the ASMEC estimator under the smoothing spline regression model. We show that it converges at an identical rate to two recently proposed basis selection methods. The effectiveness and robustness of the ASMEC approach are also supported by a variety of simulated examples and two real-life examples.

EO207 Room R08 COMPLEX DATA ANALYSIS: STATISTICAL LEARNING AND DISCOVERY Chair: Subir Ghosh

E0628: **Practical data science projects of faculty of data science of Shiga University**
Presenter: **Akimichi Takemura**, Shiga University, Japan

In April of 2017, Shiga University established the first faculty of data science in Japan. Since then, the faculty has been collaborating with more than 100 companies and government organizations, e.g., Toyota motor company and related companies. These projects often involve large data from production lines, and we can apply complex machine learning techniques. Also, some companies are willing to provide practical projects for educational purposes. We discuss these activities of the data science faculty of Shiga University.

E0579: **Statistical learning via spectrally sparse smoothers**
Presenter: **Nathaniel Helwig**, University of Minnesota, United States

Statistical learning methods seek to find reliable prediction rules that can produce insightful discoveries about complex datasets. Many statistical learning methods use penalized likelihood estimation to estimate such prediction rules, which adds a roughness penalty to the (negative) log-likelihood function. The influence of the roughness penalty on the prediction rule is typically controlled via tuning parameters, which are selected using some form of cross-validation. Popular examples of penalized likelihood methods include regularized linear regression models (e.g., ridge, lasso, elastic net) and regularized nonparametric regression models (e.g., penalized splines, generalized additive models, smoothing spline ANOVA models). To combine the benefits of regression selection and smoothing methods, we propose a spectral parameterization of a penalized spline, which allows for an efficient application of elastic net regression to smooth and select eigenvectors of a kernel matrix. The classic solution for a penalized spline is a special case of the proposed kernel eigenvector smoothing and selection operator (kesso). Extensions for tensor product smoothers are developed for both the GAM and SSANOVA frameworks. Using simulated and real data examples, we demonstrate that the kesso offers practical and computational gains over typical approaches for fitting GAMs, SSANOVA models, and elastic net penalized GLMs.

E0618: **Split modeling for high-dimensional logistic classifier ensembles**
Presenter: **Stefan Van Aelst**, University of Leuven, Belgium
Co-authors: Anthony Christidis Christidis, Ruben Zamar

An ensemble of logistic classification models is learned simultaneously by optimizing a multi-convex objective function. To enforce diversity between the models, the objective function penalizes overlap between the models in the ensemble. Measures of diversity in classifier ensembles are used to show how our method learns the ensemble by exploiting the accuracy-diversity trade-off for ensemble models. In contrast to other ensembling approaches, the resulting ensemble model is fully interpretable as a logistic regression model and, at the same time, yields excellent prediction accuracy. The excellent performance of the method for high-dimensional binary classification is demonstrated via an extensive simulation study and gene expression data applications.

E0585: **Transfer learning on regression problem**
Presenter: **Hironori Fujisawa**, The Institute of Statistical Mathematics, Japan
Co-authors: Masaaki Takada

A novel transfer learning method is proposed for regression problems in a high-dimensional setting. Knowledge in a source domain can be effectively transferred to a target domain via a special L_1 regularization. When the target domain has the same environment as the source domain, the parameter estimate obtained in the source domain tends to be transferred, as is, to the target domain. Even if the target domain has a different environment from the source domain, the parameter estimate in the target domain can have consistency by weakening the L_1 regularization effect.

E0656: **Unsaturated log-linear model selection for categorical data analysis**
Presenter: **Subir Ghosh**, University of California, United States

A new metric, SAVE, is proposed for finding the best fitted unsaturated log-linear model to describe the categorical data in a contingency table with m categorical variables. Two kinds of extensions, standard and orthogonal, of an unsaturated log-linear model to the saturated model are the foundation of SAVE. The performance of SAVE in terms of the correct model parameter(s) detection is comparable with or even better than the commonly used metrics: Deviance, AIC, and BIC, as demonstrated in simulation studies.

EC338 Room R06 CONTRIBUTIONS IN FINANCIAL ECONOMETRICS II Chair: Wenying Yao

E0183: Explaining monetary spillovers: The matrix reloaded

Presenter: **Dora Xia**, Bank for International Settlements, Switzerland

Using monetary policy shocks for seven advanced economy central banks, measured at high-frequency, the strength and characteristics of interest rate spillovers to 47 advanced and emerging market economies are documented. The main goal is to assess different channels through which spillovers occur and why some countries interest rates respond more than others. We find that there is no evidence that spillovers relate to real linkages, such as trade shows. There is some indication that exchange rate regimes influence the extent of spillovers. By far the strongest determinant of interest rate spillovers is financial openness. Countries that have stronger bilateral (and aggregate) financial links with the US or euro area are susceptible to stronger interest rate spillovers. These effects are much more pronounced at the longer end of the yield curve, indicating that while countries retain policy rate independence, financial conditions are influenced by global yields.

E0682: Extracting time-varying betas of latent factors: Bridging econometrics and deep learning

Presenter: **Hao Ma**, USI Lugano and SFI, Switzerland

The aim is to develop a novel methodology in bridging econometrics and deep learning when studying the latent factor models with time-varying betas. Under very mild econometric assumptions, we prove that both the time-varying betas and the latent factors can be identified simultaneously. Notably, the identification condition turns out to be a pure prediction problem in a high-dimensional setting, making it necessary to resort to deep learning. With the Keras functional API, we construct a deep learning architecture that guarantees a one-to-one mapping between each of its modules and each condition of our econometric assumptions, providing a solution to the black box issue econometricians are concerned with. Furthermore, the model largely reduces the number of parameters by at least n/K times compared with the conditional autoencoder. The empirical results on the US stock market show that the estimated betas are mostly correlated with market beta and idiosyncratic volatilities, with tangibility, maximum returns, and return volatility coming right after. The out-of-sample performance reaches an R2 of 1.6% for an unbalanced panel of over 8000 stocks with a 5-year test period. For the US stock market, $n = 30,000$ is the number of firms, and $K = 100$ is the number of firm characteristics.

E0536: A multi-country model of the term structures of interest rates with a GVAR

Presenter: **Rubens Moura**, Universita catholique de Louvain, Belgium

Co-authors: Candelon Bertrand

Globalization induced Macro-Finance affine term structure models (ATSM) to account for cross-borders developments. Still, reference models face issues of tractability and identification of shocks. A Global Vector Autoregressive (GVAR) is used to model the dynamics of the risk factors within an affine term structure setting. Our framework is more parsimonious and offers a more purposeful strategy to identify structural shocks. As a result, model estimation is more tractable, and the economic results are more meaningful. Furthermore, the estimation of our GVAR-ATSM is about 3 to 5 times faster than alternative benchmark specifications. The GVAR-ATSM is illustrated by the yield curve of three Latin American economies (Brazil, Mexico, and Uruguay) and China. It appears that economic activity in China impacts the interest rate significantly in these Latin American countries.

E0560: Sparse and stable index tracking with lasso and time-varying liquidity control

Presenter: **Aurelien Renault**, ESILV, France

Co-authors: Matthieu Garcin, Natach Mangan, Anmar Al Wakil

Dynamic tracking of a stock index is proposed using a lasso approach. Dealing with a high-dimensional stock universe, this technique makes it possible to select a limited and relevant dynamic subsect of stocks to build a sparse portfolio. However, adopting a dynamic approach generates significant transaction costs between two re-balancing dates if we do not limit the evolution of the portfolio. To counter this problem, we propose a new penalty acting on the difference between two consecutive sets of weights to lighten turnovers in the spirit of a fused lasso and take into account the liquidity costs. We propose a method to select the model's free parameters, and we present an empirical study on a large real stock dataset and simulated indices. We work with time-varying liquidity measures computed with two different estimators, namely the Corwin-Schultz and the Abdi-Ranaldo measures.

Friday 25.06.2021 16:45 - 18:25 Parallel Session L – EcoSta2021

EO289 Room R01 Modelling of time series and spatiotemporal data Chair: Hiroshi Shiraishi

E0165: **On the sparsity of Mallows model averaging estimator**
Presenter: **Qingfeng Liu**, Otaru University of Commerce, Japan
Co-authors: Yang Feng, Ryo Okui

Mallows model averaging estimator is shown to be written as a least-squares estimation with a weighted L_1 penalty and additional constraints. By exploiting this representation, we demonstrate that the weight vector obtained by this model averaging procedure has a sparsity property in the sense that a subset of models receives exactly zero weights. Moreover, this representation allows us to adapt algorithms developed to efficiently solve minimization problems with many parameters and weighted L1 penalty. In particular, we develop a new coordinate-wise descent algorithm for model averaging. Simulation studies show that the new algorithm computes the model averaging estimator much faster and requires less memory than conventional methods when there are many models.

E0693: **On threshold panel time series regression of cross-sectional dependence with application to climate financial analysis**
Presenter: **Lulu Wang**, University of Southampton, United Kingdom
Co-authors: Zudi Lu

Threshold autoregression has been popular in nonlinear time series modelling. Although the threshold idea has been extended to panel data analysis, it basically assumes cross-sectional independence, which cannot facilitate analysis of the impacts of climate change on different stocks in a financial market. We propose considering a threshold panel time series regression model where both residuals and regressors are allowed to be cross-sectionally dependent. We then study the asymptotic distribution theory for our proposed least-squares based estimators. Under both the time series length T and cross-sectional size n tending to infinity, the estimated coefficients are shown to be asymptotically normal with the convergence rate of root-nT. We also provide a non-standard asymptotic distribution theory for the estimated threshold parameters. Some simulated examples with different cases of fixed effects and regressors are demonstrated with the finite sample performance on theoretical properties studied. An empirical application to study the effect of precipitation on the stocks in the FTSE100 shows that our proposed methods can facilitate climate financial analysis.

E0419: **Semiparametric estimation of optimal dividend barrier for Levy processes**
Presenter: **Hiroshi Shiraishi**, Keio University, Japan
Co-authors: Yasutaka Shimizu

How an insurance portfolio is used to provide dividend income for the insurance company's shareholders is an important problem in risk theory. The premium income as dividends is paid to the shareholders whenever the surplus attains a level barrier until the next claim occurs. We consider the semi-parametric estimation of the optimal dividend barrier in Levy processes. The optimal dividend barrier is defined as the level of the barrier that maximizes the expectation of the discounted dividends until ruin. We assume that a risk process is observed discretely. Based on the observed data, the estimated expected discounted dividends for each barrier are proposed, and the estimated optimal dividend barrier is defined as the maximizer of the objective function. A contribution in practice to decision-making on dividend barrier is made when a new product is launched or optimality of an existing dividend barrier is tested.

EO150 Room R02 New proposals for clustering complex data structures Chair: Cristina Mollica

E0479: **Multilevel latent class analysis: Stepwise estimation**
Presenter: **Zsuzsa Bakk**, Leiden university, Netherlands

A two-step estimator is proposed for multilevel latent class analysis with co-variates that separates the estimation of the measurement and structural model. Keeping the measurement model fixed in step two, when covariates are added to the model, it is possible to obtain an unbiased and efficient stepwise estimator. We investigate the bias and the efficiency of the proposed estimator via a simulation study. The results show that the proposed two-step estimator is less biased than the alternative three-step estimator and almost as efficient as the one-step estimator.

E0514: **LASSO-penalized clusterwise linear regression modeling with local least angle regressions**
Presenter: **Roberto Di Mari**, Department of Economics and Business, University of Catania, Italy
Co-authors: Roberto Rocci, Stefano Antonio Gattone

In clusterwise regression analysis, the goal is to predict a response variable based on a set of explanatory variables, where each predictor has different contributions to the response depending on the cluster. The number of candidates is typically large: whereas some of these variables might be useful, some others might contribute very little to the prediction. A well-known method to perform variable selection is the LASSO, where the penalty is calibrated by minimizing the Bayesian Information Criterion (BIC). However, available approaches to the computation of LASSO-penalized estimators are time-consuming and/or require approximate schemes making the tuning of the penalty cumbersome. In order to ease such computation, we introduce an expectation-maximization algorithm with closed-form updates. This is based on an iterative scheme where the component-specific lasso regression coefficients are computed according to a coordinate descent soft-thresholding update. The LARS algorithm is used to derive the full path of component-specific LASSO solutions for model selection. We show the advantage of this approach in terms of model selection and computation time reduction through a simulation study and illustrate it with an application to Major League Baseball salary data.

E0545: **Finite mixture models with repeated measures**
Presenter: **Yun Wei**, Duke University, United States

Finite mixture models are popular to model the heterogeneity among complex data. It is known that some finite mixture models suffer from slow rates for estimating the component parameters. Examples are mixtures of those weakly identifiable families like mixtures of gamma distributions. It is proposed to collect more samples from each mixture component To obtain faster parameter convergence rates. Hence each data is a vector of samples from the same mixture component. Such a model is known as a finite mixture model with repeated measures. It has been applied in psychological study. This model also belongs to the mixture of product distributions, with the special structure that the product distributions in each mixture component are also identical. Each data consists of conditionally independent and identically distributed samples in this setup and thus is an exchangeable sequence. It is shown that by taking repeated measures, i.e. collecting more samples from each mixture component, a finite mixture model that is not originally identifiable becomes identifiable. Moreover, the posterior contraction rates for the parameter estimation are also obtained, demonstrating that repeated measures are beneficial for estimating the component parameters. The results hold for general probability families, including all regular exponential families and can also be applied to hierarchical models.

E0571: **Advances in model-based clustering of high-dimensional data**
Presenter: **Claire Gormley**, University College Dublin, Ireland

The model-based clustering framework provides well-established methods that uncover sub-groups of observations in data. Such methods bestow several desirable benefits: reproducibility due to their statistical modelling basis, objectivity through the availability of principled model selection tools and interpretability through the provision of parameter estimates and their associated uncertainties. However, model-based clustering approaches begin to lose traction as data dimension increases, whether in terms of the number of observations, variables, timepoints etc. This loss of applicability is often due to stability issues associated with high dimensional covariance matrices, optimisation difficulties and/or the expensive nature of computing the likelihood function. We consider recent advances in model-based methods to clustering data where the number of variables p is large. In particular, we explore developments in factor analytic approaches, which are well-known models for big p data, and recent work utilising composite likelihood methods that facilitate the computation of intractable likelihood functions. The utility of such methods is illustrated through benchmark and real data sets.

EO083 Room R03 MODELING AND INFERENCE FOR COMPREHENSIVE DATA Chair: Catherine Liu

E0636: **Score tests with incomplete covariates and high-dimensional auxiliary variables**
Presenter: **Alex Kin Yau Wong**, Hong Kong Polytechnic University, Hong Kong

The presence of missing values often complicates the analysis of modern biomedical data. When variables of interest are missing for some subjects, it is desirable to use observed auxiliary variables, which are sometimes high-dimensional, to impute or predict the missing values to improve statistical efficiency. Although many methods have been developed for prediction using high-dimensional variables, it is challenging to perform valid inference based on the predicted values. We develop an association test for an outcome variable and a potentially missing covariate. The covariate can be predicted using selected variables from a set of high-dimensional auxiliary variables. We establish the validity of the test under data-driven model selection procedures. We demonstrate the validity of the proposed method and its advantages over existing methods through extensive simulation studies and an application to a major cancer genomics study.

E0646: **High-dimensional nonlinear matrix-variate normal factor model**
Presenter: **Xu Zhang**, Hong Kong University, China

The high-dimensional matrix-variate data are becoming ubiquitous in various fields with the advance of technology. For example, the single-cell RNA-seq data in bioinformatics, the network data in social science and images can be regarded as matrix-variate data. By vectorizing the matrix, standard vector models can be used. But it will result in loss of the intrinsic structure information between the rows and columns and lead to a much higher dimension problem. To handle this drawback, some methods are proposed to model the matrix-variate data directly, where factor structure is incorporated for dimension reduction. But these methods either need replications of the observation or have no statistical theoretical guarantee. As a result, we focus on the high-dimensional matrix-variate data with a single observation here and propose a nonlinear matrix-variate normal factor analysis (NMVNFA), whose statistical properties can be derived. Simulation studies and single-cell RNA-seq data analysis are conducted to illustrate the performance of the method.

E0534: **A nonparametric subgroup analysis for quantile regression**
Presenter: **Xiaoyu Zhang**, The Hong Kong University, China
Co-authors: Heng Lian, Di Wang, Guodong Li

Panel data of individuals drawn from a population consisting of unknown subgroups with different conditional quantiles are considered. Specifically, we may suppose that the observed individuals can be grouped into several classes whose members all share the same conditional quantile function. We propose a pairwise fusion penalised estimation procedure for nonparametric quantile regression to identify the unknown subgroup structure and model the nonlinear conditional quantile simultaneously. We derive the asymptotic properties of the procedure and investigate its finite sample performance through simulation studies and a real data example.

E0755: **Forecasting confirmed cases of the COVID-19 pandemic with a migration-based epidemiological model**
Presenter: **Catherine Liu**, The Hong Kong Polytechnic University, Hong Kong

The unprecedented coronavirus disease 2019 (COVID-19) pandemic is still a worldwide threat to human life since its invasion into the daily lives of the public in the first several months of 2020. Predicting the size of confirmed cases is important for countries and communities to properly prevent and control policies to curb the spread of COVID-19 effectively. Unlike the 2003 SARS epidemic and the worldwide 2009 H1N1 influenza pandemic, COVID-19 has unique epidemiological characteristics in its infectious and recovered compartments. This drives us to formulate a new infectious dynamic model for forecasting the COVID-19 pandemic within the human mobility network, named the SaucIR-model in the sense that the new compartmental model extends the benchmark SIR model by dividing the flow of people in the infected state into asymptomatic, pathologically infected but unconfirmed, and confirmed. Furthermore, we employ dynamic modeling of population flow in the model to incorporate spatial effects effectively. We forecast the spread of accumulated confirmed cases in some provinces of mainland China and other countries that experienced severe infection during the time period from late February to early May 2020. The numerical analysis validates the high degree of predictability of our proposed SaucIR model compared to the existing resemblance.

EO039 Room R04 RECENT ADVANCES IN BAYESIAN INFERENCE IV Chair: Sergios Agapiou

E0403: **Finite-dimensional discrete random structures in Bayesian nonparametrics**
Presenter: **Antonio Lijoi**, Bocconi University, Italy
Co-authors: Tommaso Rigon, Igor Pruenster

Discrete nonparametric priors are effective and well-developed tools in several applications. Their actual implementation most often boils down to the use of finite-dimensional approximations obtained by truncating their infinite series representation. We take a different perspective and define random discrete priors with finite support that converge to well-known models in infinite dimensions. In doing so, we will consider random measure-based constructions and focus on versions of the Pitman-Yor process and normalized random measures with independent increments. While gaining considerable flexibility, the proposals retain analytical tractability that makes them viable alternatives. Urn schemes and posterior characterizations are obtained in closed form, leading to exact sampling methods. Besides accurately approximating their infinite-dimensional counterparts, it will be shown one can improve over existing approaches that rely on truncations.

E0542: **Leveraging Bayesian finite mixture modeling for better clustering solutions**
Presenter: **Jan Greve**, WU Vienna University of Economics and Business, Austria
Co-authors: Bettina Gruen, Sylvia Fruehwirth-Schnatter, Gertraud Malsiner-Walli

In clustering applications, Bayesian methods based on mixture models have established a firm foothold. Particularly, infinite mixture models such as Dirichlet Process Mixtures (DPMs) developed in the area of Bayesian Nonparametrics have enjoyed considerable empirical success. Recently, its parametric and finite mixture counterpart, Mixture of Finite Mixtures (MFMs), started to gain attention due to its favorable theoretical properties. A comprehensive procedure covering the complete workflow of Bayesian cluster analysis based on MFMs is presented. Crucially, the proposed methodology distinguishes the number of latent mixture components in the model from those realized as clusters in data. An inference not possible via DPMs. The procedure starts with the elicitation of the induced prior on the number of realized clusters in the data and the computation of functionals over the prior partitions to reach a suitable prior and hyperparameter specification. The subsequent sampling is performed with a generic Markov Chain Monte Carlo scheme called telescoping sampling, which admits any component distribution, re-using the updating scheme available conditional on the component memberships. Finally, a suitable post-processing step to resolve permutation invariance of the posterior called label switching is outlined to complete the analysis. R packages that streamline the overall workflow are also briefly introduced.

E0555: **Distributed testing in nonparametric models**
Presenter: **Lasse Vuursteen**, Leiden University, Netherlands
Co-authors: Botond Szabo, Harry van Zanten

In distributed methods, data is considered to be spread out over multiple locations and not available at a single central location. Communication from these locations to a central decision-maker might be limited due to bandwidth, memory or privacy constraints. We study testing statistical hypotheses in a nonparametric distributed setting. We will discuss lower bounds for communication constraint testing protocols and exhibit methods that attain such bounds. In establishing the lower bounds for hypothesis testing, analysing the Bayes risk for a certain “least favourable” prior is common. When communication constraints are in place, choosing the right adversarial prior turns out to be key in establishing the impossibility results.

E0524: **Discrete nonparametric priors with fixed mean distributions**
Presenter: **Francesco Gaffi**, Bocconi University, Italy
Co-authors: Antonio Lijoi, Igor Pruenster

Functionals of random probability measures are objects of great interest from a probabilistic perspective. They also play an important role in Bayesian Nonparametrics. In the latter context, understanding the behavior of a finite-dimensional feature of a flexible and infinite-dimensional prior is crucial for prior elicitation. The classical line of research resorts to the Cifarelli-Regazzini identity and its extensions to determine the distribution of the mean of random probability measures, firstly in the Dirichlet case and then, in greater generality, for the Pitman-Yor process and normalized random measures. This presentation targets the inverse path: determining the (unique) parameter measure of a discrete random probability measure giving rise to a desired mean distribution. Taking this direction yields a better understanding of the set of mean distributions of notable nonparametric priors, giving moreover a way to enforce prior information directly. Such a task has been completed just in the Dirichlet case with a unit concentration parameter for solving a mostly unrelated problem in combinatorics. We provide results for the general Dirichlet case, the normalized stable and the Pitman-Yor processes, with an application to mixture models.

EO195 Room R05 THEORIES OF MODERN MACHINE LEARNING METHODOLOGIES Chair: Taiji Suzuki

E0377: **Optimality and superiority of deep learning for estimating functions in variants of Besov spaces**
Presenter: **Taiji Suzuki**, University of Tokyo / RIKEN-AIP, Japan
Co-authors: Atsushi Nitanda, Kazuma Tsuji

Deep learning has exhibited superior performance for various tasks. To understand this property, we investigate the approximation and estimation ability of deep learning on some variants of Besov spaces, such as anisotropic Besov space and variable exponent Besov space. The anisotropic Besov space is characterized by direction-dependent smoothness and includes several function classes investigated thus far. We demonstrate that the approximation error and estimation error of deep learning only depend on the average value of the smoothness parameters in all directions. Consequently, the curse of dimensionality can be avoided if the smoothness of the target function is highly anisotropic. Unlike existing studies, our analysis does not require a low-dimensional structure of the input data. We also investigate the minimax optimality of deep learning and compare its performance with that of the kernel method (more generally, linear estimators). The results show that deep learning has a better dependence on the input dimensionality if the target function possesses anisotropic smoothness and it achieves an adaptive rate for functions with spatially inhomogeneous smoothness. Finally, we also discuss the learning ability of deep learning in variable exponent Besov spaces. We will show that deep learning also adapts in that situation and achieves a better rate than linear estimators.

E0322: **Fast learning rates of averaged stochastic gradient descent for over-parameterized neural networks**
Presenter: **Atsushi Nitanda**, Kyushu Institute of Technology, Japan
Co-authors: Taiji Suzuki

The convergence of averaged stochastic gradient descent for over-parameterized two-layer neural networks on the regression problem is analyzed. We consider a condition where the target function is contained in the reproducing kernel Hilbert space spanned by the neural tangent kernel, and the network width is sufficiently large such that the learning dynamics fall into the neural tangent kernel regime. Under this setting, we show the global convergence of the averaged stochastic gradient descent and derive the fast convergence rate by exploiting the complexities of the target function and the neural tangent kernel depending on the data distribution.

E0334: **Approximation of Gaussian processes by bayesian neural networks**
Presenter: **Takuo Matsubara**, The Alan Turing Institute / Newcastle University, United Kingdom
Co-authors: Chris Oates, Francois-Xavier Briol

The Bayesian Neural Network (BNN) concept aims to endow a neural network with the formal structure of a generative probability model by placing a prior distribution on the parameters of the network. BNN has been employed in several important applications, e.g. uncertainty quantification, probabilistic classification and Bayesian optimisation. It has been observed that BNNs converge to Gaussian Processes (GPs) in certain limits that correspond to the number of parameters being increased. However, it is a hard problem to verify the practical advantages (if any) of such BNNs, particularly if the limiting GP is unsuitable for the problem at hand. The aim is to reverse this logic: first to identify a suitable GP and then approximate this GP using a BNN. This ought to deliver the scalability and trainability of a neural network whilst leveraging the fact that the suitability of the limiting GP is assured. The contribution of this work is to explore situations where BNNs are capable of approximating GPs and provide an explicit construction of a BNN based on quadrature techniques and the ridgelet transform in this context. Emphasis is placed on a detailed error analysis between the BNN and the target GP that is in several respects more rigorous than previous work on the ridgelet transform.

E0333: **Counterfactual mean embeddings**
Presenter: **Motonobu Kanagawa**, EURECOM, France
Co-authors: Krikamol Muandet, Sorawit Saengkyongam, Sanparith Marukatat

The counterfactual inference has become a ubiquitous tool in online advertisement, recommendation systems, medical diagnosis, and econometrics. Accurate modeling of outcome distributions associated with different interventions – known as counterfactual distributions – is crucial for the success of these applications. We propose to model counterfactual distributions using a novel Hilbert space representation called counterfactual mean embedding (CME). The CME embeds the associated counterfactual distribution into a reproducing kernel Hilbert space (RKHS) endowed with a positive definite kernel, which allows us to perform causal inference over the entire landscape of the counterfactual distribution. Based on this representation, we propose a distributional treatment effect (DTE) that can quantify the causal effect over entire outcome distributions. Our approach is nonparametric as the CME can be estimated under the unconfoundedness assumption from observational data without requiring any parametric assumption about the underlying distributions. We also establish a rate of convergence of the proposed estimator, which depends on the smoothness of the conditional mean and the Radon-Nikodym derivative of the underlying marginal distributions. Furthermore, our framework allows for more complex outcomes such as images, sequences, and graphs. Our experimental results on synthetic data and off-policy evaluation tasks demonstrate the advantages of the proposed estimator.

EO059 Room R06 MACHINE LEARNING AND ROBUSTNESS Chair: Yiming Ying

E0512: Robust persistence diagrams using reproducing kernels

Presenter: **Bharath Sriperumbudur**, Pennsylvania State University, United States

Persistent homology has become an important tool for extracting geometric and topological features from data whose multi-scale features are summarized in a persistence diagram. From a statistical perspective, however, persistence diagrams are very sensitive to perturbations in the input space. We develop a framework for constructing robust persistence diagrams from super-level filtrations of robust density estimators constructed using reproducing kernels. Using an analogue of the influence function on the space of persistence diagrams, we establish the proposed framework to be less sensitive to outliers. The robust persistence diagrams are shown to be consistent estimators in bottleneck distance, with the convergence rate controlled by the smoothness of the kernel. This, in turn, allows us to construct uniform confidence bands in the space of persistence diagrams. Finally, we demonstrate the superiority of the proposed approach on benchmark datasets.

E0491: Analysis of online learning algorithms

Presenter: **Zheng-Chu Guo**, Zhejiang University, China

Analyzing and processing large-scale data sets is becoming ubiquitous in the era of big data. Online learning algorithms have attracted increasing interest in recent years due to their low computational complexity and storage requirements. They have been applied to various learning tasks. Unlike batch learning, which processes the whole sample once, online learning processes the sample one by one and updates the output in time. We will give some mathematical analysis of online learning algorithms in a reproducing kernel Hilbert space (RKHS) for handling large-scale data.

E0197: Convergences of regularized algorithms with random projections

Presenter: **Junhong Lin**, Zhejiang University, China

The least-squares regression problem is studied over a Hilbert space, covering nonparametric regression over a reproducing kernel Hilbert space as a special case. We first investigate regularized algorithms adapted to a projection operator on a closed subspace of the Hilbert space. We prove convergence results with respect to variants of norms under a capacity assumption on the hypothesis space and a regularity condition on the target function. As a result, we obtain optimal rates for regularized algorithms with randomized sketches, provided that the sketch dimension is proportional to the effective dimension up to a logarithmic factor.

E0179: On qualitative robustness of divide-and-conquer methods

Presenter: **Andreas Christmann**, University of Bayreuth, Germany

The topic is at the intersection of machine learning for big data and robust statistics. Divide-and-conquer methods play an important role in machine learning and big data. In robust statistics, there are five main notions of robustness: qualitative robustness, sensitivity curve, influence function, maxbias, and breakdown point. The focus will be on the qualitative robustness of machine learning methods using a divide-and-conquer approach for the big data situation. Special cases are distributed learning and localized learning.

EO299 Room R08 ADVANCES IN TIME SERIES MODELLING Chair: Wai-keung Li

E0374: Automated estimation of heavy-tailed vector error correction models

Presenter: **Feifei Guo**, Hong Kong University of Science and Technology, Hong Kong

It has been challenging to determine the co-integrating rank in the vector error correction (VEC) model when its noise is a heavy-tailed random vector. We propose an automated approach via adaptive shrinkage techniques to determine the co-integrating rank and estimate parameters simultaneously in the VEC model with unknown order p when its noises are i.i.d. heavy-tailed random vectors with tail index $\alpha \in (0,2)$. It is shown that the estimated co-integrating rank and order p equal to the true rank and the true order p_0, respectively, with probability tending to 1 as the sample size $n \to \infty$, while other estimated parameters achieve the oracle property, that is, they have the same rate of convergence and the same limiting distribution as those of estimated parameters when the co-integrating rank and the true order p_0 are known. We also propose a data-driven procedure for selecting the tuning parameters. Simulation studies are carried to evaluate the performance of this procedure in finite samples. The techniques are applied to explore the long-run and short-run behavior of prices of wheat, corn and wheat in the USA. The results may provide new insight into the Lasso approach for both stationary and non-stationary heavy-tailed time series.

E0484: Some recent results on buffered (hysteretic) autoregressive model

Presenter: **Wai-keung Li**, The Education University of Hong Kong, Hong Kong

A buffered (hysteretic) autoregression extends the classical threshold autoregression by allowing a buffer for regime changes. We focus on asymptotic statistical inferences for the two-regime buffered autoregressive (BAR) model, with autoregressive unit-roots. We will also briefly survey some recent works on quantile estimation, smooth transition and error correction formulations of the BAR model.

E0508: Testing error distribution by kernelized Stein discrepancy in multivariate time series models

Presenter: **Donghang Luo**, The University of Hong Kong, China

Co-authors: Ke Zhu, Huan Gong, Dong Li

Knowing the error distribution is important in many multivariate time series applications. To alleviate the risk of error distribution misspecification, testing methodologies are needed to detect whether the chosen error distribution is correct. However, most of the existing procedures only deal with the multivariate normal distribution for some special multivariate time series models. Thus, they cannot be used to test for the often observed heavy-tailed and skewed error distributions in applications. We construct a new consistent test for general multivariate time series models based on the kernelized Stein discrepancy. To account for the estimation uncertainty and unobserved initial values, a resampling method is provided to calculate the critical values. The new test is easy to implement for a large scope of multivariate error distributions. Simulated and real data illustrate its importance.

E0586: Hybrid random weighting spectral test for multivariate white noise checking

Presenter: **Muyi Li**, Xiamen University, China

Vector autoregressive (VAR) models are one of the most popular tools in macroeconomic analysis. Hence the correct specification of the VAR models is crucial. To this end, we propose a frequency domain spectral test to check if the residuals from a fitted VAR model are multivariate white noises. The test statistics is a Cramer-von Mises (CM)-typed spectral test. The asymptotic null distribution can be obtained under mild conditions for more general unknown dependent structures on errors. In contrast to the time domain portmanteau tests, this spectral test is consistent and has nontrivial power against local alternatives by the order of the $\sqrt{n}$. Moreover, a blockwise hybrid random weighting method is employed to bootstrap critical values of the spectral CM test. The proposed bootstrapping procedure is easy to implement, and its first-order validity is justified. Monte Carlo simulation experiments and empirical data analysis are also reported.

EG296 Room R07 CONTRIBUTIONS IN FORECASTING I Chair: Manabu Asai

E0318: **A vine-copula HAR forecasting model: Application to Dow Jones stocks**
Presenter: **Martin Magris**, Aarhus University, Denmark

The heterogeneous autoregressive (HAR) model is extended by modeling the joint distribution of the four partial-volatility terms therein involved. Namely, today's, yesterday's, last week's and last month's volatility components. The joint distribution relies on a (C-) Vine copula construction, allowing to conveniently extract volatility forecasts based on the conditional expectation of today's volatility given its past terms. The proposed novel empirical application involves more than seven years of high-frequency transaction prices for ten stocks and evaluates the in-sample, out-of-sample and one-step-ahead forecast performance of our model for daily realized-kernel measures. The forecasting model proposed is shown to outperform the HAR benchmark under different models for marginal distributions, copula construction methods, and forecasting settings.

E0645: **Forecast comparison tests under fat-tails**
Presenter: **Jihyun Kim**, Toulouse School of Economics, France
Co-authors: Nour Meddahi, Mamiko Yamashita

Forecast comparison tests are widely implemented to compare the performances of two or more competing forecasts. The critical value is often obtained by the classical central limit theorem (CLT) or by the stationary bootstrap with regularity conditions, including the one where the second moment of the loss difference is bounded. However, the heavy-tailed nature of the financial variables can violate this moment condition. We show that if the moment condition is violated, the size of the test using the classical Normal asymptotics can be heavily distorted. The distortion is large, especially when the tail of the marginal distribution of the loss differences is heavy. As an alternative approach, we propose to use a subsampling method that is robust to fat tails. In the empirical study, we analyze several variance forecast tests. Examining several tail index estimators, we show that the second moment of the loss difference is likely to be unbounded, especially when the popular squared error (SE) function is used as a loss function. We also find that the outcome of the tests may change if the subsampling is used.

E0481: **A penalized two-pass regression to predict stock returns with time-varying risk premia**
Presenter: **Gaetan Bakalli**, Universtity of Geneva, Switzerland
Co-authors: Stephane Guerrier, Olivier Scaillet

A penalized two-pass regression with time-varying factor loadings is developed. The penalization in the first pass enforces sparsity for the time-variation drivers while maintaining compatibility with the no-arbitrage restrictions by regularizing appropriate coefficients. The second pass delivers risk premia estimates to predict equity excess returns. The Monte Carlo results and the empirical results on a large cross-sectional data set of US individual stocks show that penalization without grouping can yield nearly all estimated time-varying models violating the no-arbitrage restrictions. Moreover, the results demonstrate that the proposed method reduces the prediction errors compared to a penalized approach without appropriate grouping or a time-invariant factor model.

E0710: **A hybrid combined forecasting model of water discharge based on multiple linear regression and autoregressive models**
Presenter: **Khawla Khalid Mahmood**, Middle Trchnology University/Institute of Administration Al Russafa, Iraq

A precise forecast of river water discharge plays a vital role in the field of hydrology. It is essentially used to mitigate the risk of flood and manage the water resources. A method is presented to explain and forecast water discharge based on several hydrogeological and climatic Variables. The proposed hybrid combined model consists of a Combined Multiple Linear Regression (CMLR) and an ARIMA Process for the errors of this CMLR. The long, seasonal, and short-term components extracted using the low pass filter Kolmogorov-Zurbenko are used together to build the CMLR. The models for these components built based on some hydrogeological and climatic Variables. It has been proven that the decomposition of time series is fundamental before implementing any time series analysis. Daily data for the Mohawk River in the U.S. state of New York has been used to apply this methodology. It is shown that the proposed hybrid combined model provides better forecasts than using the CMLR without a structure designed to account for its errors.

Saturday 26.06.2021 08:00 - 10:05 Parallel Session M – EcoSta2021

EO197 Room R01 RECENT ADVANCES IN ALGORITHMS FOR STATISTICAL LEARNING Chair: Eric Chi

E0181: **Biconvex clustering with adaptive feature selection**
Presenter: **Jason Xu**, Duke University, United States

Convex clustering has recently gained popularity due to computational advances and useful heuristics that have rendered it practical. While it confers many advantages over traditional clustering methods, it is also limited in the face of high-dimensional data. Not only does the Euclidean measure of fit have less discriminating power, but pairwise affinity terms that rely on k-nearest neighbors (k-NN) become poorly specified. Attempts at sparse convex clustering also suffer from the latte. We introduce feature weights to the convex clustering objective to be optimized jointly. The resulting problem remains well-behaved as a biconvex problem, and admits fast algorithms with convergence guarantees and finite-sample bounds on prediction error. Importantly, it performs feature selection that is driven adaptively by learned clustering information. As the weights change the effective feature space throughout the algorithm, affinities based on k-NN can be recomputed across iterations, largely removing the strong dependence on carefully tuned heuristics to find appropriate affinities beforehand. We thoroughly validate the algorithm on real and simulated data.

E0245: **Significance testing for canonical correlation analysis in high dimensions**
Presenter: **Xin Zhang**, Florida State University, United States
Co-authors: Ian McKeague

The problem of testing for the presence of linear relationships between large sets of random variables will be considered based on a post-selection inference approach to canonical correlation analysis. The challenge is to adjust for selecting subsets of variables having linear combinations with maximal sample correlation. To this end, we construct a stabilized one-step estimator of the euclidean-norm of the canonical correlations maximized over subsets of variables of pre-specified cardinality. This estimator is shown to be consistent for its target parameter and asymptotically normal provided the dimensions of the variables do not grow too quickly with sample size. We develop a greedy search algorithm to accurately compute the estimator, leading to a computationally tractable omnibus test for the global null hypothesis that there are no linear relationships between any subsets of variables having the pre-specified cardinality. Further, we develop a confidence interval for the target parameter that takes the variable selection into account.

E0273: **Generalized liquid association analysis for multimodal neuroimaging**
Presenter: **Jing Zeng**, Florida State University, United States
Co-authors: Lexin Li, Xin Zhang

Alzheimer's disease (AD) is the leading form of dementia, and the number of affected people is drastically increasing along with the ageing of the worldwide population. A key question of AD research is to understand the spatial associative patterns between two pathological proteins, amyloid-beta and tau, as the subject's age varies. The problem can be formulated by studying the associations of two sets of random variables conditional on the third set of random variables, a topic that has received relatively little attention but is crucial for multimodal neuroimaging analysis in general. We propose a novel generalized liquid association analysis approach, which offers a new and unique angle to study associations among three sets of random variables. We establish a population dimension reduction model, transform the problem to sparse Tucker decomposition of a three-way tensor, and develop a higher-order singular value decomposition estimation algorithm. We derive the non-asymptotic error bound and asymptotic consistency of the proposed estimator. We analyze the motivating multimodal PET dataset and identify important brain regions that exhibit the most contrastive associations as age varies. We further complement the data analysis with additional simulations to demonstrate the efficacy of the proposed method.

E0533: **Proximity operator of the matrix perspective function and its applications**
Presenter: **Joong-Ho Won**, Seoul National University, Korea, South

The matrix perspective function, which is jointly convex in the Cartesian product of a standard Euclidean vector space and a conformal space of symmetric matrices, is shown to have a proximity operator in an almost closed form. The only implicit part is to solve a semi-smooth, univariate root-finding problem. We uncover the connection between our problem of study and the matrix nearness problem. Through this connection, we propose a quadratically convergent Newton algorithm for the root-finding problem. Experiments verify that the evaluation of the proximity operator requires at most 8 Newton steps, taking less than 5s for 2000 by 2000 matrices on a standard laptop. Using this routine as a building block, we demonstrate the usefulness of the studied proximity operator in constrained maximum likelihood estimation of Gaussian mean and covariance, pseudolikelihood-based graphical model selection, and a matrix variant of the scaled lasso problem.

E0551: **Dimension reduction and changepoint detection in network series**
Presenter: **Michael Weylandt**, University of Florida, United States

As social networks and Internet of Things systems become increasingly common, the analysis of network data, and data observed on those networks, holds great potential but poses several acute statistical challenges. Foremost among these is the small sample sizes typically associated with network data, often far less than the large scale and high-dimensionality of the systems of interest. To address this, we propose a framework for dimension reduction of networks observed over time and apply it to changepoint detection. The framework is flexible, allowing for both parametric and non-parametric models of the underlying network dynamics to be used. We prove the consistency of the proposed approach under several popular network models and provide efficient tensor decomposition algorithms suitable for use on large-scale networks. As a byproduct of our analysis, we present several new consistency results for high-dimensional tensor decompositions, which are likely to be of independent interest.

EO055 Room R02 RECENT APPLICATIONS OF LATENT VARIABLE MODELS Chair: Gongjun Xu

E0251: **High-dimensional principal component analysis with heterogeneous missingness**
Presenter: **Ziwei Zhu**, University of Michigan, Ann Arbor, China
Co-authors: Tengyao Wang, Richard Samworth

The effect of missing data in Principal Component Analysis (PCA) is being studied. In simple, homogeneous missingness settings with a noise level of constant order, we show that an existing inverse-probability weighted (IPW) estimator of the leading principal components can (nearly) attain the minimax optimal rate of convergence, and discover a new phase transition phenomenon along the way. For heterogeneous missingness settings, we introduce a new method for high-dimensional PCA, called "primePCA". Starting from the IPW estimator, "primePCA" iteratively projects the observed entries of the data matrix onto the column space of our current estimate to impute the missing entries, and then updates our estimate by computing the leading right singular space of the imputed data matrix. We prove that in the noiseless case, the error of "primePCA" converges to zero at a geometric rate when the signal strength is not too small and the true principal eigenspaces are incoherent. An important feature of our theoretical guarantees is that they depend on average, as opposed to worst-case, properties of the missingness mechanism. Our numerical studies on both simulated and real data reveal that "primePCA" exhibits very encouraging performance across a wide range of scenarios.

E0339: **Hypothesis testing methods for multivariate multi-occasion intra-individual change**
Presenter: **Chun Wang**, University of Washington, United States

In psychological and educational measurement, it is often of interest to assess change in an individual. The current study expanded on previous research by introducing methods that can evaluate individual change on multiple latent traits measured on multiple occasions. The four methods considered are the likelihood ratio test (LRT), the multivariate Wald test (MWT), the modified multivariate Wald test (MMWT), and the score test (ST). Simulation studies were conducted to examine the true positive rate (TPR) and the false positive rate (FPR) of the new methods under a conventional fixed-form test and a computerized adaptive test (CAT). Manipulated variables included the number of occasions, change magnitudes, patterns of change, and correlations between latent traits. Results revealed that, in terms of FPR, all methods except MWT had close adherence to the nominal significance level. Among the three methods, the LRT is recommended as it provided a balance between FPR and TPR. A larger change magnitude yielded higher TPR, regardless of the remaining factors. With the same test length, a CAT yielded higher TPR than a conventional test.

E0464: **Integrating sample similarities into latent class analysis: A tree-structured shrinkage approach**
Presenter: **Zhenke Wu**, University of Michigan, United States
Co-authors: Mengbing Li

The aim regards using multivariate binary observations to estimate the proportions of unobserved classes with scientific meanings. We focus on the setting where additional information about sample similarities is available and represented by a rooted weighted tree. Every leaf in the given tree contains multiple independent samples. Shorter distances over the tree between the leaves indicate higher similarity. We propose a novel data integrative extension to classical latent class models (LCMs) with tree-structured shrinkage. The proposed approach enables 1) borrowing of information across leaves, 2) estimating data-driven leaf groups with distinct vectors of class proportions, and 3) individual-level probabilistic class assignment given the observed multivariate binary measurements. We derive and implement a scalable posterior inference algorithm in a variational Bayes framework. Extensive simulations show a more accurate estimation of class proportions than alternatives suboptimally using the additional sample similarity information. A zoonotic infectious disease application is used to illustrate the proposed approach. Model limitations and extensions are briefly discussed.

E0465: **A latent variable model to measure fluency using response time and response accuracy**
Presenter: **Shiyu Wang**, University of Georgia, United States
Co-authors: Yinghan Chen, Houping Xiao

The recent Every Student Succeeds Act encourages schools to use an innovative assessment to provide feedback about students mastery level of grade-level content standards. Mastery of a skill requires the ability to complete the task with not only accuracy but also fluency. We offer a new sight on using both response times and response accuracy to measure fluency with the cognitive diagnosis model framework. Defining fluency as the highest level of a categorical latent attribute, we propose a polytomous response accuracy model and two forms of response time models to infer fluency jointly. A Bayesian estimation approach is developed to calibrate the newly proposed models. These models were applied to analyze data collected from a spatial rotation test. Results demonstrate that compared with the traditional CDM that using response accuracy only, the proposed joint models were able to reveal more information regarding test-takers spatial skills. A set of simulation studies were conducted to evaluate the accuracy of the model estimation algorithm and illustrate the various degrees of model complexities.

E0724: **Time-varying overlapping clustering method via latent factor model**
Presenter: **Kean Ming Tan**, University of Michigan, United States

Clustering is an important tool in interdisciplinary research such as genomics and neuroscience. One ubiquitous assumption for most clustering methods is that each variable belongs only to one cluster, and such an assumption may be unrealistic in many scientific settings. We will introduce a clustering procedure using a latent factor model that allows overlapping clusters, i.e., each variable can belong to multiple clusters. In particular, we focus on developing a method for clustering variables on time-varying data with clusters changing across time. The proposed method is also able to match the cluster labels across time. Theoretical guarantees are established consistent estimation of the clusters.

EO075 Room R04 ECOSTA JOURNAL PART B: STATISTICS Chair: Cristian Gatu

E0177: **A fast adaptive lasso for the Cox regression via safe screening rules**
Presenter: **Hong Wang**, Central South University, China

Recent studies have shown that safe feature elimination screening algorithms are useful alternatives in solving large scale and/or ultra-high dimensional Lasso-type problems. However, to the best of our knowledge, the plausibility of adapting the safe feature elimination screening algorithm to survival models is rarely explored. We first derive the safe feature elimination screening rule for the adaptive lasso Cox model. Then, using both simulated and real-world datasets, we demonstrate that the resulting algorithm can outperform Lasso Cox and adaptive Lasso Cox prediction methods in terms of its predictive performance. In addition to its good predictive performance, we illustrate that the proposed algorithm has a key computational advantage over the above-competing methods in computation efficiency.

E0433: **A method for identifying uncorrelated outlier signals from high dimensional data**
Presenter: **Hyo Young Choi**, University of Tennessee Health Science Center, United States
Co-authors: Steve Marron, Neil Hayes

A new method is proposed for detecting uncorrelated outlier signals from high dimensional data. To develop a new notion of outliers, we define outlier subspace consistency that describes the limiting properties of outlier signals in the limit as the dimension increases. In particular, we investigate the conditions under which outliers can be asymptotically well captured by a low dimensional subspace produced by PCA. Based on these theoretical results, we introduce a new method for identifying individual outlier signals and distinguishing the subspace where only main signals are associated by un-mixing principal components. The proposed method uses a projection pursuit approach to find the most extreme one-dimensional direction where a data point is outlying while minimizing its residual outlyingness. The resulting outlier signals identified by the method can reveal the interpretable signals responsible for outliers, which helps to understand the origins of aberrations in data. As an important application, the proposed method is used to identify structural alterations in mRNA transcripts in head and neck cancer patients. We identify a number of biologically important outliers, e.g., rare cancer variants, along with the successful characterization of the subspace associated with outliers, which holds promise for identifying otherwise obscured signals.

E0648: **Combining rules for F- and Beta-statistics from multiply-imputed data**
Presenter: **Ashok Chaurasia**, University of Waterloo, Canada

Missing values in data impedes the task of inference for population parameters of interest. Multiple Imputation (MI) is a popular method for handling missing data since it accounts for the uncertainty of missing values. Inference in MI involves combining point and variance estimates from each imputed data via Rubin's combining rules. A sufficient condition for these rules is that the estimator is approximately (multivariate) normally distributed. However, these traditional combining rules get computationally cumbersome for multicomponent parameters of interest and unreliable at a high rate of missingness (due to an unstable variance matrix). Combining rules for F- and Beta-statistics from multiply-imputed data is proposed for decisions about multicomponent parameters. These proposed combining rules have the advantage of being computationally convenient since they only involve univariate F- and Beta-statistics while providing the same inferential reliability as the traditional multivariate combining rules. The simulation study demonstrates that the proposed method has good statistical properties (maintaining low type I and type II error rates at relatively large proportions of missingness). The general applicability of the proposed method is demonstrated within a lead exposure study to assess the association between lead exposure and neurological motor function.

E0411: **Classification via local manifold approximation**
Presenter: **Didong Li**, Princeton University; University of California, Los Angeles, United States
Co-authors: David Dunson

Classifiers label data as belonging to one of a set of groups based on input features. It is challenging to obtain accurate classification performance when the feature distributions in the different classes are complex, with nonlinear, overlapping and intersecting supports. This is particularly true when training data are limited. To address this problem, we propose a new type of classifier based on obtaining a local approximation to the support of the data within each class in a neighborhood of the feature to be classified, and assigning the feature to the class having the closest support. This general algorithm is referred to as LOcal Manifold Approximation (LOMA) classification. As a simple and theoretically supported special case having excellent performance in a broad variety of examples, we use spheres for local approximation, obtaining a SPherical Approximation (SPA) classifier. We illustrate substantial gains for SPA over competitors on a variety of challenging simulated and real data examples.

EO233 **Room R05** **STATISTICAL INFERENCE WITH INCOMPLETE INFORMATION** **Chair: Yen-Chi Chen**

E0180: **Copula models for sensitivity analysis in observational causal inference**
Presenter: **Alexander Franks**, UC Santa Barbara, United States
Co-authors: Alexander DAmour, Jiajing Zheng

A fundamental difficulty with observational causal inference is that assumptions about unconfoundedness are not testable from observed data. As such, quantifying the sensitivity of causal conclusions to assumptions about confounders is of immense practical importance. One common approach for sensitivity analysis is to model the existence of latent confounders explicitly. Unfortunately, many existing latent confounder models perturb observable predictions and consequently degrade model fit. To address this issue, we propose a model based on the Gaussian copula, which quantifies the strength of dependence between outcome, treatment and confounder while leaving observed data marginals unperturbed. We generalize our method for use with multiple treatments, multiple outcomes as well as proxy variables. In particular, we provide results related to the recently discussed work on the benefits of modeling multiple causes. Our results show that, while identification is not achievable in general, we derive conditions under which small, bounded ignorance regions are achieved for specific treatment effects. We also provide corresponding and contrasting results for the case multivariate outcomes.

E0203: **Causal inference under interference and network uncertainty**
Presenter: **Daniel Malinsky**, Johns Hopkins University, United States
Co-authors: Rohit Bhattacharya, Ilya Shpitser

Classical causal and statistical inference methods typically assume the observed data consists of independent realizations. However, this assumption is inappropriate in many applications due to a network of dependencies between units in the data. Methods for estimating causal effects have been developed in the setting where the structure of dependence between units is known exactly. Still, in practice, there is often substantial uncertainty about the precise network structure. This is true, for example, in trial data drawn from vulnerable communities where social ties are difficult to query directly. We combine techniques from structure learning and interference literature in causal inference, proposing a general method for estimating causal effects under data dependence when this dependence structure is not known a priori. We demonstrate the utility of our method on synthetic datasets which exhibit network dependence.

E0345: **Pattern graph: A graphical approach to nonmonotone missing data**
Presenter: **Yen-Chi Chen**, University of Washington, United States

The aim is to introduce the concept of pattern graphs–a directed acyclic graph representing how response patterns are associated. Pattern graphs provide an elegant way to model non-monotone missing data. We introduce a selection model and a pattern mixture model formulation using the pattern graphs and show that they are equivalent. Pattern graphs lead to an inverse probability weighting estimator and an imputation-based estimator for estimating a parameter of interest. Asymptotic theories of the estimators are studied, and we provide a graph-based dynamic programming procedure for computing both estimators. We introduce three graph-based sensitivity analysis and study the equivalence class under a generalized version of pattern graphs.

E0404: **Estimation in Hawkes processes as a missing data problem**
Presenter: **Soeren Wengel Mogensen**, University of Copenhagen, Denmark

Hawkes processes are a popular model class for modeling interacting streams of events in continuous time, and they can represent, e.g., social media activity or transmission of infectious diseases. We formulate structure learning and causal effect estimation in Hawkes processes as missing data problems. This is done by using the inherent branching structure of Hawkes processes, and we show that one can efficiently solve these problems within this framework. This allows, for instance, a straightforward application of the EM algorithm in this context.

E0438: **Adversarial Monte Carlo meta-learning in partially identified problems**
Presenter: **Alex Luedtke**, University of Washington, United States

Traditionally, estimation in missing data and causal inference problems has been performed using the following three-step approach: (1) posit the existence of some unidentifiable full data distribution, (2) identify the quantity of interest as a feature of the observed data distribution, and (3) develop an estimator of this feature. We propose a new numerical approach for developing estimators in these problems that entirely circumvents the identification step (2). In our approach, missing data problems are framed as two-player games in which Nature adversarially selects a full data distribution that makes it difficult for the Statistician to answer the scientific question using a coarsening of data drawn from this distribution. The players' strategies are parameterized via neural networks, and optimal play is learned by modifying the network weights over many repetitions of the game. This approach performs favorably compared to standard practice in numerical experiments.

EO109 Room R07 EAC-ISBA SESSION: FRONTIERS OF SPATIAL AND TEMPORAL DATA MODELING Chair: Won Chang

E0502: A combined physical-statistical approach for estimating storm surge risk

Presenter: **Whitney Huang**, Clemson University, United States

Storm surge is an abnormal rise of seawater caused by a storm. According to the National Hurricane Center, storm surge is often the most damaging part of a hurricane. It poses the most severe threat to property and life in a coastal region. Thus, it is crucially important to assess the storm surge risk, typically summarized by r-year surge return level with return period r ranging from 10, 50, 100, or even much longer along a coastline. However, it is challenging to reliably estimate this quantity due to the limited storm surge observations in space and time. This talk presents an approach to integrate physical and statistical models to estimate extreme storm surge. Specifically, A physically-based hydrodynamics model is used to provide the needed interpolation in space and extrapolation in both time and atmospheric conditions. Statistical modeling is needed to 1) estimate the input distribution for running the computer model, 2) develop a statistical emulator in place of the computer simulator, and 3) estimate uncertainty due to input distribution, statistical emulator, missing/unresolved physics.

E0637: Dynamic spatio-temporal modeling

Presenter: **Jonathan Stroud**, Georgetown University, United States

We present new methods for inference in dynamic spatio-temporal models. We illustrate the methods with simulations and examples.

E0676: Bayesian convolutional networks-based generalized linear model

Presenter: **Yeseul Jeon**, Yonsei University of Applied Statistics, Korea, South

Co-authors: Seokjun Choi, Won Chang, Jaewoo Park

Neural networks provide complex function approximations between inputs and a response variable for a wide variety of applications. Examples include a classification for images and regression for spatially or temporally correlated data. Although neural networks can improve prediction performance compared to traditional statistical models, interpreting the impact of explanatory variables is difficult. Furthermore, uncertainty quantification for predictions and inference for model parameters are not trivial. To address these challenges, we propose a new Bayes approach by embedding convolutional neural networks (CNN) within the generalized linear models (GLM) framework. Using extracted features from CNN as informative covariates in GLM, our method can improve prediction accuracy and provide interpretations of regression coefficients. We show that the posterior distributions of model parameters asymptotically follow mixture normals. We apply our methods to simulated and real data examples, including non-Gaussian spatial data, brain tumor image data, and fMRI data. The algorithm can be broadly applicable to correlated data and quickly provide accurate Bayesian inference.

E0726: Multivariate spectral downscaling for PM2.5 species

Presenter: **Yawen Guan**, University of Nebraska - Lincoln, United States

Fine particulate matter (PM2.5) is a mixture of air pollutants that adversely affect human health. Understanding the health effects of the PM2.5 mixture and its individual species has been a research priority over the past two decades. However, the limited availability of speciated PM2.5 measurements continues to be a major challenge in exposure assessment for conducting large-scale population-based epidemiology studies. The PM2.5 species have complex spatial-temporal and cross dependence structures that should be accounted for in estimating the spatiotemporal distribution of each component. Two major sources of air quality data are commonly used for deriving exposure estimates: point-level monitoring data and gridded numerical computer model simulation, such as the Community Multiscale Air Quality (CMAQ) model. We propose a statistical method to combine these two data sources for estimating speciated PM2.5 concentration. Our method models the complex relationships between monitoring measurements and the numerical model output at different spatial resolutions, and we model the spatial dependence and cross dependence among PM2.5 species. We apply the method to combine CMAQ model output with major PM2.5 species measurements in the contiguous United States in 2011.

E0740: A scalable partitioned approach to model massive nonstationary non-Gaussian spatial datasets

Presenter: **Seiyon Lee**, George Mason University, United States

Co-authors: Jaewoo Park

Nonstationary non-Gaussian spatial data are common in many disciplines, including climate science, ecology, epidemiology, and social sciences. Examples include count data on disease incidence and binary satellite data on cloud mask (cloud/no-cloud). Modeling such datasets as stationary spatial processes can be unrealistic since they are collected over large heterogeneous domains (i.e., spatial behavior differs across subregions). Although several approaches have been developed for nonstationary spatial models, these have focused primarily on Gaussian responses. In addition, fitting nonstationary models for large non-Gaussian datasets is computationally prohibitive. We propose a scalable algorithm for modeling such data to address these challenges by leveraging parallel computing in modern high-performance computing systems. We partition the spatial domain into disjoint subregions and fit locally nonstationary models using a carefully curated set of spatial basis functions. Then, we combine the local processes using a novel neighbor-based weighting scheme. Our approach scales well to massive datasets (e.g., 1 million samples) and can be implemented in nimble, a popular software environment for Bayesian hierarchical modeling. We demonstrate our method to simulated examples and two large real-world datasets about infectious diseases and remote sensing.

EC337 Room R03 CONTRIBUTIONS TO STATISTICAL MODELLING AND APPLICATIONS I Chair: Ramses Mena

E0468: On the comparison of GLM-based charts for industrial processes

Presenter: **Tahir Mahmood**, The Open University of Hong Kong, Hong Kong

Co-authors: Anam Iqbal, Summera Kinat

In industry 4.0, many manufacturing processes are equipped with digital devices that are storing multiple data streams asynchronously. Usually, the information about the number of conforming or non-conforming products is used as a key indicator to assess the reliability and quality of a process. This information is often modeled by the Poisson count distribution. Based on it, several control chart studies have been developed. However, most of the time, few covariates are also measured along with the number of conforming or non-confirming products. For the fitting of such data, a generalized linear model (GLM) based on Poisson distribution is often used, and for monitoring purposes, GLM-based control charts are usually used. To the best of our knowledge, GLM-based Shewhart, EWMA, and CUSUM type setups are only available to monitor the Poisson distributed process. Although above-stated structures are enabled to detect small to moderate shifts in the process, several other charting schemes exist, which are more capable than the above-stated charts. We have designed more enhanced control charting schemes under GLM monitoring setup to provide a comparative analysis. It is revealed that the proposed methods have better performance as compared to all existing methods.

E0364: Sparse Gaussian mixture regression with application to flow cytometry data analysis

Presenter: **Sangwon Hyun**, University of Southern California, United States

Co-authors: Jacob Bien, Francois Ribalet, Mattias Cape

Flow cytometry data collected in the ocean can give valuable insight into the composition and dynamics of phytoplankton populations. We present a novel method for modeling time-varying flow cytometry data conditional on a large number of environmental covariates. We develop a novel mixture of multivariate sparse regressions model that can simultaneously estimate and identify the important covariates for each phytoplankton population. The method ties covariates to both the flow cytometry population centers as well as the relative abundances of these populations. The approach involves a lasso-penalized expectation-maximization procedure with additional convex constraints to facilitate interpretation of the estimated model. We apply the method to continuous-time flow cytometry data measured from the ocean, on a ship near Honolulu traveling from warmer, nutrient-sparse subtropical waters to cooler, more productive waters. The method provides a powerful framework for developing a fine-grained understanding of the environmental drivers of phytoplankton populations in the ocean.

E0457: **Variable selection under multi-collinearity using modified log penalty**
Presenter: **Chi Tim Ng**, Hang Seng University of Hong Kong, Hong Kong
Co-authors: Van Cuong Nguyen

A class of strictly concave penalty function is described to handle the multicollinearity issues in the regression analysis. As an example, a new penalty function called modified log penalty is introduced. The penalized estimator based on strictly concave penalties enjoys the oracle property under certain regularity conditions discussed in the literature. In the multicollinearity cases where such conditions are not applicable, the behaviors of the strictly concave penalties are discussed through examples involving strongly correlated covariates. Real data examples and simulation studies are provided to show the finite-sample performance of the modified log penalty in terms of prediction error under scenarios exhibiting multicollinearity.

E0702: **Efficient selection between hierarchical cognitive models: Cross-validation with variational Bayes**
Presenter: **Viet Hung Dao**, UNSW Business School, Australia
Co-authors: David Gunawan, Minh-Ngoc Tran, Robert Kohn, Guy Hawkins, Scott Brown

Model comparison is the cornerstone of theoretical progress in scientific research. The common practice relies on tools that evaluate competing models by balancing in-sample descriptive adequacy against model flexibility, with modern approaches advocating the use of marginal likelihood for hierarchical models. Cross-validation is another popular approach, but its implementation has remained out of reach for models evaluated in a Bayesian hierarchical framework, with the major hurdle being prohibitive computational cost. It is well known that Variational Bayes (VB) produces good predictive density estimates. We thus develop a novel VB algorithm with Bayesian prediction as a tool to perform the model comparison by cross-validation, which we refer to as CVVB. In particular, CVVB can be used as a model screening device that quickly identifies bad models. We demonstrate the utility of CVVB by revisiting a classic question in decision-making research: what latent components of processing drive the ubiquitous speed-accuracy trade-off? Our approach brings cross-validation within reach of theoretically important psychological models and makes it feasible to compare much larger families of hierarchically specified cognitive models than has previously been possible

E0756: **US stock market reaction to surprises in macroeconomic data announcements**
Presenter: **Lukas Petrasek**, Charles University Prague, Czech Republic

The purpose is to analyze the impact of surprising information in macroeconomic data releases on the US stock market using several machine learning techniques. We extend the standard Fama-French 5 factor model for surprise components of releases of the most important macroeconomic variables, such as GDP growth, inflation, or unemployment, in order to test various hypotheses. First, the potential state-dependency of the market reaction is investigated. We further improve the understanding of asymmetries in the responses to positive and negative news. The effects of additional data from the macroeconomic news releases, such as updates on the past figures or detailed numbers, are also covered. Lastly, we assess the eagerness of the market as proposed in the previous literature.

Saturday 26.06.2021 10:15 - 11:55 Parallel Session N – EcoSta2021

EO321 Room R01 RECENT ADVANCES IN CAUSAL MEDIATION ANALYSIS Chair: Yi Zhao

E0595: **Bayesian methods for multiple mediators: Relating principal stratification and causal mediation**
Presenter: **Chanmin Kim**, SungKyunKwan University, Korea, South
Co-authors: Michael Daniels, Joseph Hogan, Christine Choirat, Corwin Zigler

The goal is to develop new statistical methods to quantify relationships between emissions, ambient air pollution, and human health. We frame evaluation as a mediation analysis to assess the extent to which the effect of a particular control technology on ambient pollution is mediated through causal effects on power plant emissions. Since power plants emit various compounds that contribute to ambient pollution, we develop new methods for multiple intermediate variables that are measured contemporaneously, may interact with one another, and may exhibit joint mediating effects. Specifically, we propose new methods leveraging two related frameworks for causal inference in the presence of mediating variables: principal stratification and causal mediation analysis. We define principal effects based on multiple mediators and introduce a new decomposition of the total effect of an intervention on ambient pollution into the natural direct effect and natural indirect effects for all combinations of mediators. Both approaches are anchored to the same observed-data models, which we specify with Bayesian nonparametric techniques. We provide assumptions for estimating principal causal effects, then augment these with an additional assumption required for causal mediation analysis. The two analyses, interpreted in tandem, provide the first empirical investigation of the presumed causal pathways that motivate important air quality regulatory policies.

E0631: **Inference for individual mediation effects and interventional effects in sparse high-dimensional causal graphical models**
Presenter: **Abhishek Chakrabortty**, Texas A&M University, United States

The problem of identifying mediators that regulate the effect of a treatment on a response variable is considered. While there has been significant work on this classical topic, little work has been done when the set of potential mediators is high-dimensional (HD). A further complication arises when they are interrelated with unknown dependencies. We assume that the causal structure of the treatment, confounders, potential mediators and the response is an (unknown) directed acyclic graph (DAG). HD DAG models have previously been used to estimate causal effects from observational data, and methods called IDA and joint-IDA have been developed to estimate the effects of single and multiple interventions. Here we propose an IDA-type method called MIDA for estimating so-called individual mediation effects from HD observational data. Although IDA and joint-IDA estimators have been shown to be consistent in certain sparse HD settings, their asymptotic properties, such as convergence in distribution and inferential tools in such settings, have stayed unknown. We prove the HD consistency of MIDA for linear structural equation models with sub-Gaussian errors. More importantly, we derive distributional convergence results for MIDA in similar HD settings, which apply to IDA and joint-IDA estimators as well. To our knowledge, these are the first such distributional convergence results facilitating inference for IDA-type estimators. We also validate our results via numerical studies.

E0728: **Mediation in the analysis of metabolomic data**
Presenter: **Su Chu**, Brigham and Women's Hospital and Harvard Medical School, United States

Causal mediation analysis in the molecular study of disease is increasingly common but has yet to enjoy much application in the relatively newer discipline of metabolomics. Metabolomics is the study of all small molecules and their various interactions within a biological system. As the most proximal '-ome' to disease phenotype in the biological central dogma, the metabolome plays a mediating role in intrinsic biological and extrinsic environmental risk factors as they exert their effects on disease phenotype. It is thus a crucial mechanistic inflexion point for developing personalized, precision medicine intervention strategies. We survey techniques for integrative high dimensional metabolomic mediation, considering the specific biases that inherently characterize untargeted global profiling, and identify optimal strategies via simulation. We illustrate our approach with real data, and extensions to various study designs are also considered.

E0207: **Discussion**
Presenter: **Joseph Hogan**, Brown University School of Public Health, United States

The aim is to provide a discussion of the key issues raised in the three talks on causal mediation analysis and highlight new directions for research in this area.

EO231 Room R02 ADVANCES IN ANALYSIS METHODOLOGIES MEDICAL AND BIOMEDICAL APPLICATIONS Chair: Daniel Jeske

E0498: **Efficient estimation of general treatment effects using neural networks with a diverging number of confounders**
Presenter: **Ying Liu**, University of California, Riverside, United States
Co-authors: Xiaohong Chen, Shujie Ma, Zheng Zhang

The estimation of causal effects is a primary goal of behavioral, social, economic and biomedical sciences. Under the unconfounded treatment assignment condition, adjustment for confounders requires estimating the nuisance functions relating to outcome and/or treatment to confounders. The conventional approaches rely on either a parametric or a nonparametric modeling strategy to approximate the nuisance functions. Parametric methods can introduce serious bias into casual effect estimation due to possible misspecification, while nonparametric estimation suffers from the "curse of dimensionality". A new unified approach is proposed for efficient estimation of treatment effects using feedforward artificial neural networks when the number of covariates is allowed to increase with the sample size. We consider a general optimization framework that includes the average, quantile and asymmetric least squares treatment effects as special cases. Under this unified setup, we develop a generalized optimization estimator for the treatment effect with the nuisance function estimated by neural networks. We further establish the consistency and asymptotic normality of the proposed estimator and show that it attains the semiparametric efficiency bound. The proposed methods are illustrated via simulation studies and a real data application.

E0510: **Learning fine-resolution chromosome conformation interaction maps**
Presenter: **Wenxiu Ma**, University of California Riverside, United States

High-throughput methods based on chromosome conformation capture technologies have enabled us to investigate the three-dimensional (3D) genome organization at an unprecedented resolution. However, high-resolution maps of chromatin interactions require costly, extremely deep sequencing and have been achieved for only a small number of cell lines. Without sufficient sequencing depth, the observed chromatin interaction maps are very sparse and noisy, which imposes great statistical and computational challenges. We will present recent work on enhancing the resolution of chromatin interaction maps via a generative adversarial network framework.

E0581: **Real real-world evidence**
Presenter: **David Madigan**, Northeastern University, United States

In practice, learning healthcare systems rely primarily on observational studies generating one effect estimate at a time using customized study designs with unknown operating characteristics and publishing - or not - one estimate at a time. Instead, we propose a standardized process for performing observational research that can be evaluated, calibrated, and applied at scale to generate a more reliable and complete evidence base than previously possible, fostering a truly learning healthcare system. We demonstrate this process in the contexts of hypertension and depression.

E0589: **A Bayesian longitudinal trend analysis of count data with Gaussian processes**
Presenter: **Samantha VanSchalkwyk**, University of California, Riverside, United States
Co-authors: Daniel Jeske

The context of comparing two different groups of subjects that are measured repeatedly over time is considered. Our specific focus is on highly variable count data, which have a non-negligible frequency of zeros and have time trends that are difficult to characterize. These challenges are often present when analyzing bacteria or gene expression data sets. Traditional longitudinal data analysis methods, including Generalized Estimating Equations, can be challenged by the features present in these types of data sets. We propose a Bayesian methodology that effectively confronts these challenges. A key feature of the methodology is the use of Gaussian Processes to model the time trends flexibly. Inference procedures based on both sharp and interval null hypotheses are discussed, including the important hypotheses that test for group differences at individual time points. The proposed methodology is illustrated with next-generation sequencing data sets corresponding to two different experimental conditions. In particular, the method is applied to a case study containing bacteria counts of mice with chronic and non-chronic wounds to identify potential wound-healing probiotics. The methodology can be applied to similar next-generation sequencing data sets comparing two groups of subjects.

EO243 Room R04 RECENT ADVANCES IN THE MODELING OF FUNCTIONAL AND HIGH-DIMENSIONAL DATA Chair: Xiaowei Wu

E0168: **Bayesian causal discovery for reverse-engineering single-cell gene regulatory networks**
Presenter: **Yang Ni**, Texas A&M University, United States

Novel Bayesian causal discovery approaches will be presented. They are motivated by single-cell RNA-seq data. The proposed approaches are causally identifiable for purely observational, cross-sectional data under some causal assumptions.

E0565: **A scalable Bayesian method to leverage multiple quantitative functional annotations in GWAS of complex traits**
Presenter: **Jingjing Yang**, Emory University, United States
Co-authors: Junyu Chen

Many statistical methods have been proposed for integrating functional annotations in GWAS to prioritize potential causal variants and quantify heritability contribution per annotation. However, most of these methods are limited to categorical and non-overlapped functional annotations, assuming one variant can only be annotated with one categorical label. In order to relax this limitation, we propose a scalable Bayesian method BFGWAS_QUANT that jointly models genome-wide variants assuming quantitative functional annotations as covariates of a logistic model for the causal probabilities of genetic variants. The simulation studies showed that BFGWAS_QUANT estimates of annotation coefficients (i.e., enrichment of GWAS loci) converged to the truth. Both phenotype heritability estimates and GWAS power were improved up to 42% compared to not accounting for functional annotations. We applied BFGWAS_QUANT to the whole genome sequencing dataset for studying Alzheimer Dementia (AD) related phenotypes. We found that the quantitative annotation of probabilities of being cis-molecular quantitative trait loci in brain tissues had the highest enrichment. We demonstrated that the Bayesian polygenic risk scores (BPRS) derived from the BFGWAS_QUANT summary statistics with $N = 1,123$ had higher AD risk prediction accuracy than the standard PRS derived from a single variant GWAS summary statistics from IGAP with $N =$ 54K.

E0668: **Association testing for binary trees: A Markov branching process approach**
Presenter: **Xiaowei Wu**, Virginia Tech, United States
Co-authors: Hongxiao Zhu

A new approach is proposed to test associations between binary trees and covariates. This approach models binary trees with varying branch lengths by a binary fission Markov branching process. Based on a generalized linear regression model, we developed theoretical results and procedures for association testing, including variable selection and estimation of the covariate effects. The performance of such a modeling and inference approach was evaluated by simulation studies with both model-based data and model-free, semi-synthetic data generated from real brain-tumor images. One typical application of the proposed approach is to model dendrograms generated from hierarchical clustering on pixel intensities in biomedical images. Our final analysis of the glioblastoma multiforme brain-tumor data from The Cancer Imaging Archive identified multiple clinical and genetic variables that are potentially associated with brain-tumor heterogeneity.

E0697: **Multilevel-multiclass graphical model**
Presenter: **Inyoung Kim**, Virginia Tech, United States

The Gaussian graphical model has been a popular tool for investigating the conditional dependency structure between random variables by estimating sparse precision matrices. However, the ability to investigate the conditional dependency structure when a two-level structure exists among the variables is still limited. Some variables are considered as higher-level variables while others are nested in these higher-level variables - the latter are called lower-level variables. Higher-level variables are not isolated; instead, they work together to accomplish certain tasks. Therefore, our main interest is to simultaneously explore conditional dependency structures among higher-level variables and among lower-level variables. Given two-level data from heterogeneous classes, we propose a method to jointly estimate the two-level Gaussian graphical models across multiple classes, so that common structures in terms of the two-level conditional dependency are shared during the estimation procedure, yet unique structures for each class are retained as well. We also demonstrate the advantages of our approach using breast cancer patient data.

EO045 Room R05 RECENT ADVANCES IN TIME SERIES ANALYSIS Chair: Runmin Wang

E0162: **Revealing cluster structures based on mixed sampling frequencies: Application to the state-level labor markets**
Presenter: **Yeonwoo Rho**, Michigan Technological University, United States
Co-authors: Yun Liu, Hie Joo Ahn

Mixed data sampling (MIDAS) models have drawn much attention among professional forecasters for their capability for a concise yet data-driven summary of information in frequently observed variables. While a parametric MIDAS model provides a parsimonious tool to summarize information in high-frequency data, one parametric form may not necessarily be appropriate for all cross-sectional subjects. A penalized regression approach is proposed that lets the data reveal their underlying cluster structure. To ease this clustering procedure, a simple yet flexible nonparametric MIDAS specification is proposed. The proposed clustering algorithm delivers reasonable clustering results, both in theory and in simulations, without requiring knowledge of the true group membership. An empirical application on the state-level labor markets in the United States is presented, clustering the states based on the response of the unemployment rate to the regional gross domestic product growth and weekly initial unemployment insurance claims. The mixed-frequency Okun's law relationship suggests that the state-level labor markets can be clustered into two groups composed of 11 states and the rest distinguished by the cyclical sensitivity of the unemployment rates and the changing predictability of weekly initial claims through the quarter.

E0231: **Testing for the martingale difference hypothesis in multivariate time series models**
Presenter: **Ke Zhu**, University of Hong Kong, Hong Kong

A general class of tests is proposed to examine whether the error term is a martingale difference sequence in a multivariate time series model with a parametric conditional mean. These new tests are formed based on the recently developed martingale difference divergence matrix (MDDM). They provide formal tools to test the multivariate martingale hypothesis in the literature for the first time. Under suitable conditions, the asymptotic null distributions of these MDDM-based tests are established. Moreover, these MDDM-based tests are consistent to detect a broad class of fixed alternatives and have nontrivial power against local alternatives of order $n^{-1/2}$, where n is the sample size. Since the asymptotic null distributions depend on the data generating process and the parameter estimation, a wild bootstrap procedure is further proposed to approximate the critical values of these MDDM-based tests, and its theoretical validity is justified. Finally, the usefulness of these MDDM-based tests is illustrated by simulation studies and one real data example.

E0259: **Volatility martingale difference divergence matrix for dimension reduction of multivariate volatility**
Presenter: **Chung Eun Lee**, University of Tennessee, Knoxville, United States
Co-authors: Xiaofeng Shao

The so-called volatility martingale difference divergence matrix (VMDDM)is proposed to quantify the conditional variance dependence of a random vector Y given X, building on the recent work on the martingale difference divergence matrix (MDDM) measures the conditional mean dependence. We further generalize VMDDM to the time series context and apply it to make dimension reduction for multivariate volatility, following the recent work. Unlike the latter work, our metric is easy to compute, can fully capture nonlinear serial dependence and involves fewer user-chosen numbers. Furthermore, we propose a variant of VMDDM and apply it to the estimation of the conditional uncorrelated components model. Simulation and data illustration show that our method can perform well compared to the existing ones with the less computational time and can outperform others in cases of strong nonlinear dependence.

E0267: **Hypothesis testing for high-dimensional time series via self-normalization**
Presenter: **Runmin Wang**, Southern Methodist University, United States
Co-authors: Xiaofeng Shao

Self-normalization has attracted considerable attention in the recent literature of time series analysis, but its scope of applicability has been limited to low-/fixed-dimensional parameters for low-dimensional time series. We propose a new formulation of self-normalization for inference about the mean of high-dimensional stationary processes. Our original test statistic is a U-statistic with a trimming parameter to remove the bias caused by weak dependence. Under the framework of nonlinear causal processes, we show the asymptotic normality of our U-statistic with the convergence rate dependent upon the order of the Frobenius norm of the long-run covariance matrix. The self-normalized test statistic is then constructed on the basis of recursive subsampled U-statistics and its limiting null distribution is shown to be a functional of time-changed Brownian motion, which differs from the pivotal limit used in the low-dimensional setting. We also present applications to testing for bandedness of the covariance matrix and testing for white noise for high-dimensional stationary time series and compare the finite sample performance with existing methods in simulation studies.

EO275 Room R06 NONLINEARITY IN APPLIED ECONOMETRICS Chair: Feng Yao

E0161: **Estimation of a partially linear seemingly unrelated regressions model: Application to a translog cost system**
Presenter: **Kai Sun**, Shanghai University, China
Co-authors: Xin Geng

A partially linear seemingly unrelated regressions model is studied to estimate a translog cost system that consists of a partially linear translog cost function and input share equations. A simple and feasible estimation procedure is proposed. We show that both our parametric and nonparametric component estimators are consistent, asymptotically normal, and more efficient relative to the single-equation counterparts. We highlight that the relative efficiency gain of the nonparametric estimator for a particular equation, based on the Cholesky decomposition, improves with its position in the system and is maximized when this equation is placed at the end. A model specification test for parametric functional forms is proposed, and how to correct the between- and within-equation heteroscedasticity is also discussed. An Italian banking data set is used to estimate the translog cost system, yielding policy implications for risk management in banking.

E0208: **Financial openness and investment allocations in Chinese real sectors**
Presenter: **Zhouheng Wu**, Guangdong University of Foreign Studies, China
Co-authors: Shenguo Yuan

The impacts of financial openness on the investment allocations between fixed investment and financial investment in real sectors of China are examined, applying the system GMM method and 2003 to 2016 annual data of non-financial listed firms. The empirical results show that, first, financial openness leads to the reallocation of investment from a fixed asset to a financial asset. Second, the negative impacts of financial openness are heterogeneous across sectors. They are also different when we distinguish the impacts of capital inflow and capital outflow. The impacts of capital inflows are larger than that of capital-output. At the sectoral level, the fixed asset investment ratio in the agricultural and industrial sectors is largely affected by capital inflow and capital outflow. However, the fixed asset investment ratio in the non-financial service sector is only affected by capital outflow. Final, we found both direct and indirect impacts of financial openness on real sectors. Financial openness would strengthen the negative impacts of relative return and risk factors on the fixed asset investment.

E0388: **A fixed effect additive stochastic frontier model with interactions and distribution free inefficiency**
Presenter: **Taining Wang**, Capital University of Economics and Business, China
Co-authors: Feng Yao, Subal Kumbhakar

A semiparametric additive stochastic frontier model with interactions for panel data is proposed, where inputs and environment variables can enter the frontier individually and interactively through unknown functions. We disentangle time-invariant unobserved heterogeneities of firms from their technical inefficiency, which can be helpful to avoid overestimating the inefficiency level. The inefficiency has its mean function known up to finite parameters, influenced by its determinants that may or may not appear in the frontier. More importantly, we do not impose distribution assumption on the composite error for inefficiency estimation. Our model can incorporate a large number of frontier or inefficiency determinants flexibly, with the curse of dimensionality rising only in the order of interaction functions. We illustrate the appealing finite-sample performance of the proposed estimator and two related tests through the Monte Carlo study and apply the world production frontier model with 116 countries during 2001-2013.

E0506: **Efficient nonparametric estimation of regression discontinuity**
Presenter: **Feng Yao**, West Virginia University, United States

Regression discontinuity is proposed to be estimated non parametrically with a modified local constant and local linear estimators. The approach is based on a modified kernel density estimator that estimates the extension of the density. The estimator requires no kernel modification near the boundary, and standard kernels can be used. Unlike the nonparametric regression discontinuity estimation based on local constant estimators, our estimator exhibits the same rates at the boundary and interior points of the domain. We illustrate its encouraging finite sample performance with a simulation study.

EO261 Room R07 RECENT ADVANCES IN FUNCTIONAL DATA ANALYSIS Chair: Yuhang Xu

E0620: **Statistical modeling and inference for next-generation functional data**
Presenter: **Guannan Wang**, College of William & Mary, United States
Co-authors: Lily Wang, Yueying Wang

With the rapid growth of modern technology, many large-scale imaging studies have been or are being conducted to collect massive datasets with large volumes of imaging data, thus boosting the investigation of next-generation functional data. These enormous collections of imaging data contain interesting information and valuable knowledge, which has raised the demand for further advancement in functional data analysis. We mainly focus on modeling and inference of the next-generation functional data. We propose using flexible multivariate splines over triangulation or tetrahedral partitions to handle the irregular domain of the images common in brain imaging studies and other biomedical imaging applications. The proposed spline estimators are shown to be consistent and asymptotically normal under some regularity conditions. We also provide a computationally efficient estimator of the covariance function and derive its uniform consistency. Finally, we discuss the inferential capabilities of the proposed method. To be more specific, we develop simultaneous confidence corridors for the mean of the next-generation functional data. The procedure is also extended to the two-sample case in which we focus on comparing the mean functions of random samples drawn from two populations. The proposed method is applied to analyze brain Positron Emission Tomography (PET) data of Alzheimer's Disease.

E0621: **Growth dynamics for plant high-throughput phenotyping studies using hierarchical functional data analysis**
Presenter: **Yuhang Xu**, Bowling Green State University, United States
Co-authors: Yehua Li, Yumou Qiu

In modern high-throughput plant phenotyping, images of plants of different genotypes are repeatedly taken throughout the growing season, and phenotypic traits of plants (e.g., plant height) are extracted through image processing. It is of interest to recover whole trait trajectories and their derivatives at both genotype and plant levels based on observations made at irregular discrete time points. We propose to model trait trajectories using hierarchical functional principal component analysis (HFPCA) and show that recovering derivatives of the trajectories is reduced to estimating derivatives of eigenfunctions, which is solved by differentiating eigenequations. Simulation studies show that the proposed procedure performs better than its competitors in recovering both trait trajectories and their derivatives. Interesting characteristics of plant growth dynamics are revealed in the application to a modern plant phenotyping study.

E0629: **Low-rank covariance function estimation via functional unfolding**
Presenter: **Raymond Ka Wai Wong**, Texas A&M University, United States
Co-authors: Jiayi Wang, Xiaoke Zhang

Multidimensional function data arise from many fields nowadays. The covariance function plays an important role in the analysis of such increasingly common data. We will present a novel nonparametric covariance function estimation approach under the reproducing kernel Hilbert spaces (RKHS) framework that can handle both sparse and dense functional data. We extend multilinear rank structures for (finite-dimensional) tensors to functions, allowing for flexible modeling of covariance operators and marginal structures. The proposed framework can guarantee that the resulting estimator is automatically semi-positive definite and can incorporate various spectral regularizations. The trace-norm regularization, in particular, can promote low ranks for both covariance operator and marginal structures. Despite the lack of a closed-form, under mild assumptions, the proposed estimator can achieve unified theoretical results that hold for any relative magnitudes between the sample size and the number of observations per sample field. The rate of convergence reveals the phase-transition phenomenon from sparse to dense functional data.

E0640: **Temporal-dependent principal component analysis of two-dimensional functional data**
Presenter: **Kejun He**, Renmin University of China, China
Co-authors: Shirun Shen, Lan Zhou

A novel model is proposed to analyze the temporal-dependent two-dimensional functional data on an irregular domain. Illustrated by a study of Texas temperature, the method assumes that the functional principal component scores of two-dimensional functional data are temporally correlated and models the scores as latent time series. To overcome the challenge that the two-dimensional functions of interest can be irregularly and sparsely observed, we use the bivariate spline basis on triangulations. These ideas are integrated into a unified model and an expectation-maximization (EM) algorithm along with Kalman filter, and smoother is developed to estimate the unknown parameters. A simulation study is conducted to demonstrate that the proposed model outperforms its alternative. We finally use the proposed model to analyze the dataset of Texas temperature and the results are consistent with the scientific conclusions in domain knowledge.

EO061 Room R08 MATHEMATICS OF DATA SCIENCE Chair: Ding-Xuan Zhou

E0204: **Chebyshev-type cubature formulas on weighted spheres**
Presenter: **Han Feng**, City University of Hong Kong, Hong Kong

The strict Chebyshev-type cubature formula (CF) (i.e., equal-weighted CF) is presented for doubling weights on the unit sphere equipped, with the usual surface Lebesgue measure and geodesic distance. The main interest is in the minimal number of nodes required in a strict Chebyshev-type CF. Precisely, given a normalized doubling weight on the unit sphere, we will establish the sharp asymptotic estimates of the minimal number of distinct nodes, which admits a strict Chebyshev-type CF. In addition, if the weight function is essentially bounded, the nodes involved can be configured well-separately in some sense. The proofs of these results rely on constructing new convex partitions of the unit sphere that are regular with respect to the weight. The weighted results on the unit sphere also allow us to establish similar results on strict Chebyshev-type CFs on the unit ball and the standard simplex.

E0389: **Riemannian distances between infinite-dimensional covariance operators and Gaussian processes**
Presenter: **Minh Ha Quang**, RIKEN, Japan

The aim is to present several recently formulated Riemannian distances between infinite-dimensional positive definite Hilbert-Schmidt operators on a Hilbert space, particularly covariance operators associated with functional random processes. The main focus is on the affine-invariant Riemannian distance, corresponding to the Fisher-Rao distance between centered Gaussian measures, the Log-Hilbert-Schmidt distance, and the family of Alpha Procrustes distances, including the Bures-Wasserstein distance, corresponding to the 2-Wasserstein distance between centered Gaussian measures. Consistent finite-dimensional approximations of the infinite-dimensional distances are also established. The theoretical formulation is illustrated with numerical results on covariance operators associated with Gaussian processes.

E0516: **Generalized linear regression with grouped and right-censored count data**
Presenter: **Xin Guo**, The Hong Kong Polytechnic University, Hong Kong
Co-authors: Qiang Fu, Tian-Yi Zhou, Yue Wang, Kenneth Land

Count responses with grouping and right censoring (GRC) have long been used in surveys to study a variety of behaviors, status, and attitudes in criminology, demography, epidemiology, marketing, sociology, psychology, and other related disciplines. Yet, the corresponding methodologies for generalized linear regression are far from developed, and grouping or right-censoring decisions of count responses still rely on arbitrary choices made by researchers. To implement generalized linear regression for GRC counts, we developed a hybrid line search algorithm for parameter inference, demonstrated the finite-sample performance of the estimators via artificial and real data, and derived theory on their asymptotic behaviors. For the design of the grouping scheme, we proposed a new algorithm, named Fisher Information Maximizer (FIM), which finds the global maximizer of a score function based on the Fisher information matrix, among infinitely many grouping schemes. By developing a general regression-based framework based on which data collection and statistical analysis can be unified, a novel tool to design and analyze GRC counts is offered.

E0559: **Learning theory of weighted distance weighted discrimination**
Presenter: **Jun Fan**, Hong Kong Baptist University, Hong Kong

High dimension low sample size (HDLSS) statistical analysis of binary classification has attracted great interest recently. We study a weighted generalized distance weighted discrimination (DWD) algorithm associated with varying Gaussian kernels to overcome the data-piling issue and imbalanced data issue of existing classification algorithms in the HDLSS context. We derive fast learning rates for the proposed algorithm under mild conditions within the framework of learning theory. The parameter in the generalized DWD loss function plays an important role in our error analysis.

EG298 Room R03 CONTRIBUTIONS IN FORECASTING II Chair: Donggyu Kim

E0572: **Reducing the risk in tail risk forecasting models**
Presenter: **Dan Li**, Queensland University of Technology, Australia
Co-authors: Adam Clements, Christopher Drovandi

The purpose is to demonstrate that existing quantile regression models used for forecasting Value-at-Risk (VaR) and expected shortfall (ES) are sensitive to initial conditions. A Bayesian quantile regression approach is proposed for estimating joint VaR and ES models. By treating the initial values as unknown parameters, sensitivity issues can be dealt with. Furthermore, a new additive-type model is developed for the ES component that is robust to initial conditions. A novel approach using the open-faced sandwich (OFS) method is proposed, which improves uncertainty quantification in risk forecasts. Simulation and empirical results highlight the improvements in risk forecasts ensuing from the proposed methods.

E0657: **Predictive quantile regression with mixed roots and increasing dimensions**
Presenter: **Rui Fan**, Rensselaer Polytechnic Institute, United States
Co-authors: Ji Hyung Lee, Youngki Shin

The benefit of using the adaptive LASSO for predictive quantile regression is studied. It is common that financial predictors in predictive quantile regression have various degrees of persistence and exhibit different signal strengths in explaining the dependent variable, such as stock returns. We show that the adaptive LASSO has the consistent variable selection and the oracle properties under the simultaneous presence of stationary, unit root and cointegrated predictors. Some encouraging simulation and out-of-sample prediction results are reported.

E0703: **Forecast encompassing in modal regression for equity premium prediction**
Presenter: **Yaojue Xu**, University of California, Riverside, United States
Co-authors: Tae-Hwy Lee

While the equity premium may not be predictable in mean using many macroeconomic and financial predictors, it may well be predictable in some quantiles especially in tails or in the mode. The mode has its own merits relative to mean and quantile. It is robust when the distribution is skewed. Is the mode of the financial returns more predictable than the mean? With this in mind, we develop a novel framework for Granger-causality (GC) test in the predictive regression for the conditional mode based on the forecast-encompassing principle. We show that the encompassing statistic is asymptotically standard normal with zero mean under the null hypothesis that there is no GC in the modal regression. The Monte-Carlo simulation shows the encompassing test has a proper size and excellent power in finite samples. We apply the encompassing statistic to test if financial and macroeconomic variables Granger-cause the mode of the equity premium distribution. The mode prediction results for the equity premium are generally more significant than the mean prediction results.

E0729: **Back to the present: Learning about the euro area through a now-casting model**
Presenter: **Thiago Ferreira**, Federal Reserve Board, United States
Co-authors: Danilo Cascaldi-Garcia, Domenico Giannone, Michele Modugno

A model is built for simultaneously now-casting economic conditions in the euro area and its three largest member countries—Germany, France, and Italy. The model formalizes how market participants and policymakers monitor the euro area by incorporating all market-moving indicators in real-time. We find that area-wide and country-specific data provide informative signals to now-cast the economic conditions in the euro area and member countries. The model provides accurate predictions of economic conditions in real-time over a period that covers the past three recessions.

Saturday 26.06.2021 13:55 - 15:35 Parallel Session P – EcoSta2021

EO329 Room R01 RECENT ADVANCES IN MODEL SELECTION AND RELATED TOPICS Chair: Shinpei Imori

E0264: **Robust bayesian regression with synthetic posterior**
Presenter: **Shintaro Hashimoto**, Hiroshima University, Japan

A robust Bayesian regression method using synthetic posterior and shrinkage priors for regression coefficients is proposed. Our synthetic posterior is based on a robust loss function, and an efficient posterior computation algorithm using Bayesian bootstrap within Gibbs Sampling is provided. The performance of the proposed method is evaluated through simulation.

E0299: **On the verifiable identification condition in NMAR missing data analysis**
Presenter: **Kosuke Morikawa**, Osaka University and The University of Tokyo, Japan
Co-authors: Kenji Beppu

Missing data often causes undesired properties such as bias and loss of efficiency. By modeling the distribution of complete data and its missing-data mechanism, incorporating them into the likelihood can solve the problem. However, especially when the missing-data mechanism is NMAR (Not missing at random), there are two problems: (i) we cannot verify sufficient conditions for the distribution of complete data; (ii) guaranteeing model identifiability is difficult even for relatively simple models. Some recent studies tackle the first problem (i) by modeling the distribution of observed data, not complete data, which is impossible to obtain. As for problem (ii), we have derived sufficient conditions for model identifiability under nonignorable nonresponse by specifying that the distribution of the outcome model is a normal or normal mixture, but the missing-data mechanism is any parametric model. The new conditions can check whether assumed models are identifiable by the observed data in NMAR missingness.

E0527: **A consistent variable selection method with GIC in multivariate linear regression even when dimensions are large**
Presenter: **Ryoya Oda**, Hiroshima University, Japan
Co-authors: Hirokazu Yanagihara

The focus is on variable selection criteria for selecting explanatory variables in a normality-assumed multivariate linear regression. Since our setting is that the dimensions of response and explanatory variables may be large, but the sum does not exceed the sample size, it is not practical to calculate a variable selection criterion over all combinations of explanatory variables. Thus, we deal with a variable selection method via the KOO method, which is useful in terms of run-times when the dimension of explanatory variables is large, with the generalized information criterion (GIC). We obtain conditions for consistency of the GIC used in the KOO method under the following high-dimensional asymptotic framework: the sample size tends to infinity, the sum of the dimensions of response and explanatory variables divided by the sample size converges to a positive constant within [0,1), and the dimension of true explanatory variables may tend to infinity. Then, using the obtained conditions, we propose a consistent criterion used in the KOO method under the high-dimensional asymptotic framework. Through simulation experiments, it is shown that the probability of selecting true explanatory variables by our proposed criterion is high, and the run-time is fast even when the dimensions are large.

E0627: **Variable selection in high-dimensional multivariate linear regression models with group structure**
Presenter: **Shinpei Imori**, Hiroshima University, Japan

A variable selection problem is studied in high-dimensional multivariate linear regression models. In multivariate linear regression models, it is often assumed to have common explanatory variables for each response variable. However, the condition that the same explanatory variables are used for each response variable cannot express group structure among response variables. We mitigate this assumption and derive a sufficient condition so that Cp-type criteria have asymptotic efficiency when the number of response variables and explanatory variables is large.

EO051 Room R02 RECENT ADVANCES IN STATISTICS Chair: Yunjin Choi

E0306: **Identifiability of structrual equation models**
Presenter: **Gunwoong Park**, University of Seoul, Department of Statistics, Korea, South

The identifiability assumption of structural equation models (SEMs) is considered in which each variable is determined by an arbitrary function of its parents plus an independent error. It has been shown that linear Gaussian structural equation models are fully identifiable if all error variances are the same or known. Hence, we prove the identifiability of SEMs with both homogeneous and heterogeneous unknown error variances. The new identifiability assumption exploits error variances and edge weights; hence, it is strictly milder than prior work on the identifiability result. We further provide a statistically consistent and computationally feasible learning algorithm. We verify through simulations that the proposed algorithm is statistically consistent and computationally feasible in the high-dimensional settings and performs well compared to state-of-the-art US, GDS, LISTEN, PC, and GES algorithms. We also demonstrate through real human cell signalling and mathematics exam data that our algorithm is well-suited to estimating DAG models for multivariate data compared to other methods used for continuous data.

E0348: **Bayesian fixed-domain asymptotics for covariance parameters in gaussian random field models**
Presenter: **Cheng Li**, National University of Singapore, Singapore

Gaussian random fields are commonly used for modeling spatial processes. We focus on the Gaussian process with Matern covariance functions. Under fixed-domain asymptotics, it is well known that when the dimension of data is less than or equal to three, the microergodic parameter can be consistently estimated with asymptotic normality while the length-scale (or range) parameter cannot. Motivated by this frequentist result, we establish a Bernstein-von Mises theorem for the covariance parameters. Under the fixed-domain asymptotics, we show that the posterior distribution of the microergodic parameter converges in total variation norm to a normal distribution with shrinking variance, while the length-scale parameter remains unidentifiable and its posterior does not converge even if the sample size increases to infinity. As a result, the posterior predictive mean squared error enjoys the asymptotic efficiency, as if the true microergodic parameter were known. We illustrate these asymptotic results in numerical experiments.

E0530: **Global rates of convergence in mixture density estimation**
Presenter: **Arlene Kyoung Hee Kim**, Korea University, Korea, South
Co-authors: Adityanand Guntuboyina

The estimation of a monotone decreasing density f_0 represented by a scale mixture of uniform densities is considered. The rates of convergence of the MLE has been previously conjectured to be $n^{-1/3}$ with a log factor whose power depends on d, but the proof has not been provided yet. We first derive a general bound on the Hellinger accuracy of the MLE over convex classes. Using this bound with an entropy calculation, we provide a different proof for the convergence of the MLE for $d = 1$. Then, we consider a possible multidimensional extension. We can prove, for $d \geq 2$, that the rate is as conjectured under the assumption that the density is bounded from above and below and supported on a compact region. We are exploring strategies for weakening the assumptions.

E0549: **Semiparametric causal mediation analysis under unmeasured mediator-outcome confounding**
Presenter: **BaoLuo Sun**, National University of Singapore, Singapore

Although the exposure can be randomly assigned in studies of mediation effects, any form of direct intervention on the mediator is often infeasible. As a result, unmeasured mediator-outcome confounding can seldom be ruled out. We propose semiparametric identification of natural direct and indirect effects in the presence of unmeasured mediator-outcome confounding by leveraging heteroskedasticity restrictions on the observed data law. For inference, we develop semiparametric estimators that remain consistent under partial misspecification of the observed data model. We illustrate the robustness of the proposed estimators through both simulations and an application to evaluate the effect of self-efficacy on fatigue among health care workers during the COVID-19 outbreak.

EO089 **Room R03** INNOVATIVE THEORY AND APPLICATIONS OF HIDDEN MARKOV MODELS **Chair: Xinyuan Song**

E0604: **Order selection for regression-based hidden Markov model**
Presenter: **Yiqi Lin**, The Chinese University of Hong Kong, Hong Kong

Hidden Markov models (HMMs) describe the relationship between two stochastic processes: an observed process and an unobservable finite-state transition process. Owing to their modeling dynamic heterogeneity, HMMs are widely used to analyze heterogeneous longitudinal data. Traditional HMMs frequently assume that the number of hidden states (i.e., the order of HMM) is a constant and should be specified prior to analysis. This assumption is unrealistic and restrictive in many applications. We consider the regression-based hidden Markov model (RHMM) while allowing the number of hidden states to be unknown and determined by the data. We propose a novel likelihood-based double penalized method, along with an efficient expectation-conditional maximization with iterative thresholding-based descent (ECM–ITD) algorithm, to perform order selection in the context of RHMM. An extended Group-Sort-Fuse procedure is proposed to rank the regression coefficients and impose penalties on the discrepancy of adjacent coefficients. The order selection consistency and convergence of the ECM–ITD algorithm are established under mild conditions. Simulation studies are conducted to evaluate the empirical performance of the proposed method. An application of the proposed methodology to a real-life study on Alzheimer's disease is presented.

E0605: **Bayesian quantile non-homogeneous Hidden Markov models**
Presenter: **Yanlin Tang**, East China Normal University, China

The hidden Markov model (HMM) is a useful tool for simultaneously analyzing a longitudinal observation process and its dynamic transition process. Existing HMMs have mainly focused on mean regression for the longitudinal response. However, the tails of the response distribution are as important as the centre in many substantive studies. We propose a quantile HMM to provide a systematic method to examine the whole conditional distribution of the response given the hidden state and potential covariates. Instead of considering homogeneous HMMs, which assume the probabilities of between-state transitions are independent of subject- and time-specific characteristics, we allow the transition probabilities to depend on exogenous covariates, thereby yielding non-homogeneous Markov chains and making the proposed model more flexible than its homogeneous counterpart. We develop a Bayesian approach coupled with efficient Markov chain Monte Carlo methods for statistical inference. Simulation studies are conducted to assess the empirical performance of the proposed method. An application of the proposed methodology to a cocaine use study provides new insights into the prevention of cocaine use.

E0601: **A new Bayesian joint model for longitudinal and survival data with latent variables**
Presenter: **Xiaoxiao Zhou**, The Chinese University of HongKong, Hong Kong

A new joint modeling approach is developed that incorporates latent variables to analyse longitudinal and survival data. A hidden Markov model, comprised of two components, is proposed to deal with the continuous longitudinal response variables. The primary goal to describe the relationship between the observation and unobservable finite-state transition processes. The first component of HMM is a transition model for elucidating how potential covariates influence the probabilities of transitioning from one state to another. The second component is a parametric conditional model for characterizing the state-specific effects of observed covariates and latent risk factors, measured by multiple observed variables through factor analysis. We also generalize the conventional proportional hazards model to accommodate the latent risk factors in the analysis of survival data. A shared random effect is introduced to the conditional model and survival model to emphasize the association between the observation process and time-to-event process. A Bayesian approach coupled with efficient Markov chain Monte Carlo methods is developed to conduct statistical inference. The performance is evaluated by simulation studies. Moreover, an application to a joint survey of cognitive function and mortality is presented.

E0616: **Bayesian analysis of hidden Markov varying coefficient models with zero-effect regions**
Presenter: **Hefei Liu**, Qujing Normal University, China

In psychological, social, behavioral, and medical studies, hidden Markov models have been extensively applied to the simultaneous modeling of heterogeneous observation and hidden transition in longitudinal data analysis. However, the existing hidden Markov models are developed in a constant-coefficient framework. A novel hidden Markov varying coefficient model is considered, in which the dynamic relations in the effects of potential covariates on the response variable can be investigated. A Bayesian inference method is developed to estimate the model, especially to detect the zero-effect regions. The empirical performance of the proposed method is evaluated through simulation studies.

EO047 **Room R04** RECENT ADVANCES IN BAYESIAN INFERENCE V **Chair: Marta Catalano**

E0499: **Adaptive Bayesian inference for current status data on a grid**
Presenter: **Minwoo Chae**, Pohang University of Science and Technology, Korea, South

A Bayesian approach is studied to infer an event time distribution in the current status model where observation times are supported on a grid of potentially unknown sparsity, and multiple subjects share the same observation time. The model leads to a very simple likelihood, but statistical inferences are non-trivial due to the unknown sparsity of the grid. In particular, for an inference based on the maximum likelihood estimator, one needs to estimate the density of the event time distribution which is challenging because the event time is not directly observed. We consider Bayes procedures with a Dirichlet prior on the event time distribution. With this prior, the Bayes estimator and credible sets can be easily computed via a Gibbs sampler algorithm. Our main contribution is to provide a thorough investigation of frequentist's properties of the posterior distribution. Specifically, it is shown that the posterior convergence rate is adaptive to the unknown sparsity of the grid. If the grid is sufficiently sparse, we further prove the Bernstein–von Mises theorem which guarantees frequentist's validity of Bayesian credible sets. A numerical study is also conducted for illustration.

E0570: **Frequentist coverage of empirical Bayesian uncertainty quantification using deep neural network regression**
Presenter: **Stefan Franssen**, Leiden University, Netherlands
Co-authors: Botond Szabo

In the past 5 years there has been a breakthrough in our understanding in the behaviour of (sparse) Deep Neural network regression. For β-Hölder spaces, Schmidt-Hieber gave near minimax convergence rates, and the work has been extend to Besov spaces by Suzuki. These works give guarantees for the square loss of (near) minimizers of the empirical square loss, which imply that Deep Neural Networks following their designs will have good uncertainty quantification. In spite of this progress, there has not been any rigorous way of quantifying uncertainty in the estimates of Deep Neural Networks. We provide both an Empirical Bayesian methodology to provide uncertainty quantification and a theoritical analysis with frequentist coverage guarantees. We also ran a simulation study which illustrates the coverage properties.

E0489: **Tree boosting for learning probability measures**
Presenter: **Naoki Awaya**, Duke University, United States
Co-authors: Li Ma

Learning probability measures based on an i.i.d. sample is a fundamental inference task but is challenging when the sample space is high-dimensional. Inspired by the success of tree boosting in high-dimensional classification and regression, we propose a tree boosting method for learning high-dimensional probability distributions. We formulate concepts of "addition" and "residuals" on probability distributions in terms of compositions of a new, more general notion of multivariate cumulative distribution functions (CDFs) than classical CDFs. This then gives rise to a simple boosting algorithm based on forward-stagewise (FS) fitting of an additive ensemble of measures. The output of the FS algorithm allows analytic computation of the probability density function for the fitted distribution. It also provides an exact simulator for drawing independent Monte Carlo samples from the fitted measure. Typical considerations in applying boosting—namely choosing the number of trees, setting the appropriate level of shrinkage/regularization in the weak learner, and evaluating variable importance—can be accomplished analogously to traditional boosting in supervised learning. Numerical experiments confirm that boosting can substantially improve the fit to multivariate distributions compared to the state-of-the-art single-tree learner and is computationally efficient.

E0495: **Trees of random probability measures and Bayesian nonparametric modelling**
Presenter: **Filippo Ascolani**, Bocconi University, Italy
Co-authors: Antonio Lijoi, Igor Pruenster

A way to generate trees of random probability measures is introduced, where a hierarchical procedure gives the link between two nodes: starting from a common root, each node of the tree is endowed with a random probability measure, whose baseline distribution is again random and given by the associated node in the previous layer. The data can be observed at any node of the tree, and different branches may have a different length: the split mechanism can also be considered random or based on covariates of interest. When the branches have the same length and the observations are linked only to the leaves, we recover the well-known family of discrete hierarchical processes prove that, if the distribution at each node is given by the normalization of a completely random measure (NRMI), the model is analytically tractable: conditional on a suitable latent structure, the posterior is still given by a deep NRMI. Furthermore, the asymptotic behaviour of the number of clusters is derived when either the sample size at a particular layer diverges, or the number of levels grows. Finally, the extension to kernel mixtures is discussed.

EO035 Room R05 RECENT DEVELOPMENT OF VAR AND SVAR MODEL Chair: Koichi Maekawa

E0328: **Forecasting public investment using daily stock returns**
Presenter: **Hiroshi Morita**, Hosei University, Japan

The predictability of public investment in Japan is investigated using the daily excess stock returns of the construction industry, to contribute to the recent discussion on fiscal foresight. To examine the relationship between monthly public investment and daily stock returns without any prior time aggregation, we employ the VAR model with MIDAS regression and estimate the optimal weights for connecting high-frequency and low-frequency data in addition to VAR coefficients and the variance-covariance structure. We find that the VAR model with MIDAS regression reduces the mean square prediction error in out-of-sample forecasting by approximately 15 and 2.5 percents compared to the no-change forecast and VAR model forecasting with prior time aggregation, respectively. Moreover, using the local projection method, we find evidence of the fiscal news shock estimated in our proposed model delaying positive effects on output, consumption, hours worked, and real wage when news shocks actually result in increasing public investment. This finding suggests the New Keynesian structure of the Japanese economy.

E0353: **A dynamic econometric analysis of the dollar-pound exchange rate in an era of structural breaks and policy regime shifts**
Presenter: **Takamitsu Kurita**, Fukuoka University, Japan
Co-authors: Jennifer L Castle

A newly-developed partial cointegration system is employed allowing for level shifts to examine whether economic fundamentals form the long-run determinants of the dollar-pound exchange rate over a recent period characterized by structural breaks and policy regime shifts. We uncover a class of local data generation mechanisms underlying long-run and short-run dynamic features of the exchange rate using a set of economic variables that explicitly reflect quantitative monetary policy and the influence of a forward exchange market. The impact of the Brexit referendum is evaluated by examining forecasts when the dollar-pound exchange rate fell substantially around the vote.

E0680: **Application of non-Gaussian SVAR model to the analysis of Japans quantitative easing monetary policy**
Presenter: **Tadashi Nakanishi**, Hiroshima University, Japan
Co-authors: Koichi Maekawa, Takashi Senda

In recent years the independent component analysis (ICA), originally developed in machine learning, has been introduced to time series econometrics. The number of papers using ICA is increasing. The advantage of the ICA approach is that it can give an alternative way to avert identification problem under non-Gaussian disturbances. Another advantage is that if the contemporaneous coefficient matrix in Structural VAR is lower triangular, then the causal order of economic variables can be easily detected. This paper attempts to check the effectiveness ICA-based SVAR model by Monte Carlo experiment and to compare the performance of several existing macro models. Furthermore, we apply this model to detect the effect of quantitative easing monetary policy by the Bank of Japan.

E0371: **Estimation of non-Gaussian structural VAR model under a flexible quasi-log-likelihood function**
Presenter: **Koichi Maekawa**, Hiroshima University of Economics, Japan
Co-authors: Tadashi Nakanishi

Maximum likelihood estimation for the non-Gaussian Structural VAR model is considered when the non-Gaussian density function of the error term is unknown. In such a case, we cannot construct a precise log-likelihood function. To overcome this problem, we propose a method to search for an appropriate quasi log-likelihood function among the Pearson family of probability density functions. We carried out Monte Carlo experiments assuming the structural errors generated by non-Gaussian density function, such as t, Laplace, and Hyperbolic secant distributions. The results show that the generated errors are fitted well to a density function of Pearson Type VII. Then we construct a quasi log-likelihood function using a selected Pearson Type VII and calculate the ML estimator of structural coefficients. The performance of ML estimation is satisfactory.

EO073 Room R07 ECOSTA JOURNAL PART A: ECONOMETRICS Chair: Erricos John Kontoghiorghes

E0409: **Size does matter: Big data machine learning portfolios**
Presenter: **Simon Hediger**, University of Zurich, Switzerland
Co-authors: Gianluca De Nard, Markus Leippold

The importance of the training phase is underlined when applying a machine learning algorithm in empirical asset pricing, for example, when forecasting stock returns. We argue that both stock-specific and (complete) over-arching approaches are suboptimal and that a better choice lies in between those two. We propose an easy to implement and fast stock grouping scheme based on the stocks' market capitalization. We highlight that the data generating process is tremendously different for large-caps, small-caps and micro-caps. We recommend that machine learning algorithms should be applied to specific groups rather than for all stocks at once. In an extensive empirical analysis, we demonstrate that training the machine learning algorithms group-specific results in a stellar and significant out-of-sample performance gain in terms of return predictability and portfolio performance (e.g., R^2_{OOS} above five percent and Sharpe ratio above four).

E0182: **A generalization of Friedman's permanent income hypothesis with a large, negative income shock**
Presenter: **Seyoung Park**, University of Nottingham, United Kingdom
Co-authors: Steven Kou

There is an unstoppable technological revolution. New forms of work driven by artificial intelligence and big data analysis are emerging. According to the recent related research, 47% of jobs in the U.S. runs the risk of being automated, thereby disrupting labor markets and affecting workers negatively in the long run. Potentially catastrophic loss of income is an omnipresent risk. It increases concern about future income uncertainty. So many people face the challenge of determining how to continue being able to afford what they can currently afford, i.e., how to attain a smooth profile of future consumption. The aim is to generalize Friedman's permanent income hypothesis (PIH) with a large, negative income shock (LNIS) and evaluate whether the generalized framework can explain how people would respond to the LNIS to increase their resilience through the lens of their precautionary savings. Having generalized the PIH with the LNIS, we examine its effect on interest rates using a general equilibrium analysis. We make the following two contributions. First, the precautionary savings could increase as wealth increases, consistent with the US data. Second, the LNIS could influence a decrease in equilibrium interest rate, which is particularly relevant to today's low-interest-rate environment.

E0386: **Bayesian estimation of realized EGARCH model to forecast tail risks**
Presenter: **Vica Tendenan**, The University of Sydney, Australia
Co-authors: Richard Gerlach, Chao Wang

A Bayesian framework is developed for the realized exponential generalized autoregressive conditional heteroskedasticity (Realized EGARCH) model, which is used to forecast tail risks such as value at risk and expected shortfall. The Realized EGARCH model is extended by considering the combination of the Gaussian, standardized student-t, and skewed-t distributions for the return equation with the Gaussian distributions in the measurement equation(s). We consider different types of realized measures, such as realized variance, realized kernel, and realized range. The Bayesian estimation is conducted by employing Markov chain Monte Carlo (MCMC) procedures by using the robust adaptive Metropolis algorithm (RAM) in the burn-in period and the standard random walk Metropolis in the sample period. The Bayesian estimator is compared with maximum likelihood estimators and shows more favourable results. The one-step-ahead forecast of the tail risks of six international equity index market is conducted for over a period of 1000 days. The forecast performance of the model is evaluated via VaR and ES backtests.

E0402: **Adaptive robust large volatility matrix estimation based on high-frequency financial data**
Presenter: **Donggyu Kim**, KAIST, Korea, South

Several novel statistical methods have been developed to estimate large integrated volatility matrices based on high-frequency financial data. They require sub-Gaussian or finite high-order moments assumptions for observed log-returns, which cannot account for the heavy tail phenomenon of stock returns. Recently, the robust estimator is developed to handle the heavy-tailed distributions with bounded fourth moments assumption. However, we often observe that the tail index of observed log-returns is less than 4, and the heavy-tailedness are heterogeneous over the asset and time period. To deal with the heterogeneous heavy-tailed distributions, we develop an adaptive robust integrated volatility estimator which employs pre-averaging and truncation schemes according to the daily tail indexes. We call this adaptive robust pre-averaging realized volatility (ARP) estimator. We show that the ARP estimator has the sub-Gaussian tail concentration with only finite 2α-moments for any $\alpha > 1$. In addition, we establish matching upper and lower bounds to show that the estimation procedure is optimal. To estimate large integrated volatility matrices using the approximate factor model, the ARP estimator is further regularized by the principal orthogonal complement thresholding method (POET). The numerical study is conducted to check the finite sample performance of the ARP estimator.

EC331 Room R06 CONTRIBUTIONS TO STATISTICAL MODELLING AND APPLICATIONS II Chair: Liming Xiang

E0472: **Unsupervised learning through generalized mixture model**
Presenter: **Samyajoy Pal**, LMU Munich, Germany
Co-authors: Christian Heumann

A generalized way of building mixture models using different distributions is explored. The EM algorithm is used with some modifications to accommodate different distributions within the same model. The model uses any point estimate available for the respective distributions to estimate the mixture components and model parameters. The focus is on the application of mixture models in unsupervised learning problems, especially cluster analysis. The convenience of building mixture models using the generalized approach is further emphasised by appropriate examples, exploiting the well-known maximum likelihood and Bayesian estimates of the parameters of the parent distributions.

E0477: **Statistical learning: Bernoulli time series modelling for discrete decision choice**
Presenter: **Miguel Angel Ruiz Reina**, Universidad de Malaga, Spain

Not infrequently, decisions have been temporally binary modelled under uncertainty. Binary data time series are used in natural science, pure science, computer science, or social science. We present the temporal Bernoulli modelling in uncertainty contexts that allows extracting knowledge for Statistical Learning, finding patterns not previously described and without information. The proposed Bernoulli regression uses an uncertainty factor with high explanatory power; firstly, we can find endogeneity problems with the error term. The Generalised Method of Moments method, solving endogeneity problems and guaranteeing the consistency of the parameters. The goal is to intelligently provide future decision tools with modelling the past in future uncertainty contexts. The model is compared with other prediction models in the literature (Entropy Model, SARIMA and ARDL + Seasonality). The Matrix U1 Theil is a decision tool that guarantees the model's high forecasting capacity presented over the others. The decision-making framework can be applied in a wide variety of domains. This modelling can be applied to supervised learning and optimisation contexts. The analysis of real data sets confirms the methodological framework; in particular, we are interested in modelling economic agents' decision temporarily in a mutually exclusive context.

E0523: **Development of data mining trajectories based on the statistical matching resources**

Presenter: **Elena Zarova**, Plekhanov Russian University of Economics, Russia

Currently, data mining methods are increasingly used to identify and assess hidden patterns, that is, structures and relationships that are not predetermined by the a priori hypotheses of the researcher. Analysis of published materials reveals two problems in this context: (1) a fragmentary use of data mining methods, leading to a decrease in their potential information efficiency; (2) a limitedness of the initial information space. To solve the first problem, the author proposes algorithms for the formation of optimal trajectories for the systemic application of data mining methods. A system of criteria is proposed. They are based on the statistical significance of the results of the method at each stage of data mining and the goals of the following stage in the general trajectory of data mining. To solve the second problem, it is proposed to use the methods of statistical matching. This increases the set of features to be explored and provides new opportunities for data mining trajectories. The proposed approaches and algorithms were tested on the data of Russian official statistics based on the results of labor force surveys and income and expenditure surveys using the R packages. This provided more reliable and accurate data on income distribution and its determinants. These results are practically meaningful for national statistical offices.

E0463: **Clustering data with nonignorable missingness usingsemi-parametric mixture models**

Presenter: **Matthieu Marbac**, CREST - ENSAI, France

Co-authors: Marie du Roy de Chaumaray

The focus is on clustering continuous data sets subject to nonignorable missingness. We perform clustering with a specific semi-parametric mixture, avoiding the component distributions and the missingness process to be specified under the assumption of conditional independence given the component. Estimation is performed by maximizing an extension of smoothed likelihood allowing missingness. This optimization is achieved by a Majorization-Minorization algorithm. We illustrate the relevance of our approach by numerical experiments. Under mild assumptions, we show the identifiability of our model, the monotony of the MM algorithm, and the estimator's consistency. We propose an extension of the new method to mixed-type data that we illustrate on a real data set.

Saturday 26.06.2021 15:45 - 17:25 Parallel Session Q – EcoSta2021

EO105 Room R01 THEORIES AND METHODOLOGIES FOR HIGH-DIMENSIONAL DATA ANALYSIS Chair: Aki Ishii

E0217: **On asymptotic normality of CDM-PCA in HDLSS**
Presenter: **Shao-Hsuan Wang**, National Central University, Taiwan
Co-authors: Su-Yun Huang, Ting-Li Chen

Principal component analysis in high dimension low sample size setting has been an active research area in recent years. A cross data matrix-based method showed the asymptotic normality for estimates of spiked eigenvalues, and also consistency for corresponding estimates of PC directions was previously proposed. However, the asymptotic normality for estimates of PC directions is still lacking. We have extended previous work to include the investigation of the asymptotic normality for the leading CDM-based PC directions and to compare it with the asymptotic normality for the classical PCA. Numerical examples are provided to illustrate the asymptotic normality.

E0355: **Clustering by kernel PCA with Gaussian kernel and tuning for high-dimensional data**
Presenter: **Yugo Nakayama**, Kyoto University, Japan
Co-authors: Kazuyoshi Yata, Makoto Aoshima

Theories and methodologies for high-dimensional data have become increasingly important in many fields. The clustering based on the linear principal component analysis (PCA) for high dimensional mixture models has been previously considered. We study asymptotic properties of the kernel PCA (KPCA) for high-dimensional data. We give the clustering based on the KPCA. In particular, we investigate the asymptotic properties of the KPCA with the Gaussian kernel. We give theoretical reasons why the Gaussian kernel is effective for clustering high-dimensional data. Since the clustering performance is influenced by a scale parameter involved in the Gaussian kernel, we discuss a choice of the scale parameter yielding a high performance of the KPCA with the Gaussian kernel in numerical simulations and actual data analyses.

E0316: **Tests for covariance structures in high-dimensional data**
Presenter: **Kazuyoshi Yata**, University of Tsukuba, Japan
Co-authors: Aki Ishii, Makoto Aoshima

Testing high-dimensional covariance structures is considered: (i) scaled identity matrix, (ii) diagonal matrix, and (iii) intraclass covariance matrix. We produce a new test statistic for each covariance structure by using the extended cross-data-matrix methodology. We show that the test statistic is an unbiased estimator of its test parameter for each covariance structure. We prove that the test statistic has a consistency property and establishes the asymptotic normality. We propose a new test procedure for (i) to (iii) and evaluate its asymptotic size and power theoretically when both the dimension and sample size grow. Finally, we demonstrate the proposed test procedure by using a microarray data set.

E0688: **Change-testing for high-dimensional econometric factor models**
Presenter: **Ansgar Steland**, University Aachen, Germany

Change-point tests for high-dimensional time series, especially factor models, are presented to test for changes in the dependence structure. The test is related to bilinear projections of the sample covariance matrix and covers CUSUM statistics, self-standardized CUSUMs, as well as subsample CUSUMs, maximized over all possible locations. A multiple testing approach is used to handle multiple projections, which allows analyzing (pseudo) eigenvalues and eigenspaces of the data. This leads to a method that is highly efficient from a computational viewpoint. Simulations show that our approach competes well with known methods and has decent statistical properties for high-dimensional settings, both in terms of level and power under various alternatives. The approach is illustrated by analyzing the well known Fama-French factors during the COVID-19 financial crash.

EO346 Room R04 RECENT ADVANCES IN BAYESIAN INFERENCE VI Chair: Minwoo Chae

E0427: **A Bayesian nonparametric approach for functional regression with application to sport data**
Presenter: **Raffaele Argiento**, Universita Cattolica del Sacro Cuore, Italy
Co-authors: Silvia Montagna, Alessandro Lanteri

In sports analytics, there is often interest in predicting elite athletes' performance at a future sporting event, given his/her competitive results tracked throughout the athlete's career and other (time-varying) covariates. Such predictions can be useful for scouting purposes and to build red flag indicators of unexpected increases in athlete performance for targeted anti-doping testing. We propose a predictive model for the longitudinal trajectory of athletes performance where we characterize the curve with a sparse basis expansion allowing individual time-dependent covariates to impact the shape of the estimated trajectories. Moreover, we introduce random intercepts, distributed according to a nonparametric hierarchical process, in order to induce clustering while borrowing statistical information across curves. In particular, we assume a hierarchical normalized generalized gamma process to grants great flexibility in clustering and accuracy in prediction. We apply our model to a longitudinal study on shot put athletes, where their competitive results are tracked throughout their career

E0482: **Ultimate Polya gamma samplers: Efficient MCMC for possibly imbalanced binary and categorical data**
Presenter: **Gregor Zens**, Vienna University of Economics and Business, Austria
Co-authors: Helga Wagner, Sylvia Fruehwirth-Schnatter

Modeling binary and categorical data is one of the most commonly encountered tasks of applied statisticians and econometricians. While Bayesian methods in this context have been available for decades now, they often require a high level of familiarity with Bayesian statistics or suffer from issues such as low sampling efficiency. To contribute to the accessibility of Bayesian models for binary and categorical data, we introduce novel latent variable representations based on Polya Gamma random variables for a range of commonly encountered discrete choice models. From these latent variable representations, new Gibbs sampling algorithms for binary, binomial and multinomial logistic regression models are derived. All models allow for a conditionally Gaussian likelihood representation, rendering extensions to more complex modeling frameworks such as state-space models straight-forward. However, sampling efficiency may still be an issue in these data augmentation based estimation frameworks. To counteract this, MCMC boosting strategies are developed and discussed in detail. The merits of our approach are illustrated through extensive simulations and a real data application.

E0546: **Learning multimorbidity and its temporal dynamics with the Wright-Fisher Indian buffet process**
Presenter: **Woojung Kim**, University of Warwick, United Kingdom
Co-authors: Paul Jenkins, Christopher Yau

A multimorbidity trajectory charts the time-dependent acquisition of disease conditions in an individual. This is important for understanding and managing patients with a complex array of multiple chronic conditions, particularly later in life. We have developed a model based on a Bayesian nonparametric feature allocation model with a Wright-Fisher Indian Buffet Process prior. The model, which we call the Multimorbidity Wright-Fisher Indian Buffet Process (mWFIBP), defines a generative process in which a set of individuals' diseases is drawn from latent multimorbidity clusters whose dependency structure across time is governed by the Wright-Fisher diffusion. We demonstrate the utility of our model in applications to simulated data and disease event data from patient electronic health records. In both settings, we show how the mWFIBP can obtain an intelligible representation of latent multimorbidity clusters and their time susceptibility and predict future disease acquisition.

E0547: **Modeling univariate and multivariate stochastic volatility in R with stochvol and factorstochvol**
Presenter: **Darjus Hosszejni**, WU Vienna University of Economics and Business, Austria
Co-authors: Gregor Kastner

Stochastic volatility (SV) models are nonlinear state-space models that enjoy increasing popularity for fitting and predicting heteroskedastic time series. However, due to the large number of latent quantities, their efficient estimation is non-trivial and software that allows fitting SV models to data easily is rare. We aim to alleviate this issue by presenting novel implementations of four SV models delivered in two R packages. Several unique features are included and documented. Unlike previous versions, stochvol is now capable of handling linear mean models, heavy-tailed SV, and SV with leverage. Moreover, we newly introduce factorstochvol, which caters for multivariate SV. Both packages offer a user-friendly interface through the conventional R generics and a range of tailor-made methods. Computational efficiency is achieved via interfacing R to C++ and doing the heavy work in the latter. In the paper at hand, we provide a detailed discussion on Bayesian SV estimation and showcase the use of the new software through various examples.

EO315 Room R05 ADVANCED COMPUTATIONAL STATISTICS FOR MACHINE LEARNING Chair: Jong-june Jeon

E0265: **An efficient parallel block coordinate descent algorithm for large-scale precision matrix estimation using GPUs**
Presenter: **Donghyeon Yu**, Inha University, Korea, South
Co-authors: Young-Geun Choi, Seunghwan Lee

Large-scale sparse precision matrix estimation has attracted wide interest from the statistics community. The convex partial correlation selection method (CONCORD) has recently been credited with some theoretical properties for estimating sparse precision matrices. The CONCORD obtains its solution by a coordinate descent algorithm (CONCORD-CD) based on the convexity of the objective function. However, since a coordinate-wise update in CONCORD-CD is inherently serial, a scale-up is nontrivial. We propose a novel parallelization of CONCORD-CD, namely, CONCORD-PCD. CONCORD-PCD partitions the off-diagonal elements into several groups and updates each group simultaneously without harming the computational convergence of CONCORD-CD. We guarantee this by employing the notion of edge coloring in graph theory. It turns out that CONCORD-PCD simultaneously updates off-diagonal elements in which the associated edges are colorable with the same color. As a result, the number of steps required for updating off-diagonal elements reduces from $p(p-1)/2$ to $p-1$ (for even p) or p (for odd p), where p denotes the number of variables. We prove that the number of such steps is irreducible. A numerical study shows that the SIMD-parallelized PCD algorithm implemented in graphics processing units (GPUs) boosts the CONCORD-CD algorithm multiple times.

E0288: **Accelerated gradient method for convex-concave saddle-point problems**
Presenter: **Donghwan Kim**, KAIST, Korea, South

Some recent machine learning problems, such as a generative adversarial network (GAN), require solving large-dimensional saddle-point problems. Gradient descent ascent type methods are widely used to solve such problems. We present some recent progress on accelerating the gradient method for convex-concave saddle-point problems.

E0405: **On proper local scoring rules for spherical data**
Presenter: **Keisuke Yano**, The University of Tokyo, Japan
Co-authors: Fumiyasu Komaki, Yuya Takasu

Statistical analysis of data on spheres often suffers from intractable likelihoods. This intractability comes from the fact that many statistical models for spherical data do not have normalizing constants with closed forms. One of the most important examples is the Fisher-Bingham model. Several methods for obtaining MLE have been proposed: the saddle-point approximation; the holonomic gradient method. We present yet another method for parameter estimation of spherical data instead of MLE. Our method has several merits: (i) avoiding the calculation of normalizing constants; (ii) possessing root n-consistency. The key idea is to construct a divergence between probability measures that do not depend on normalizing constants using the Poincare lemma. The method is an extension of score matching and proper local scoring rules for continuous distributions on the real line by previous researches. We will discuss how to extend the results to mixture models.

E0665: **Regularized Babington Smith ranking models**
Presenter: **Jong-june Jeon**, University of Seoul, Korea, South
Co-authors: Sang Jun Moon

The regularization method of the Babington Smith model, from which academically important ranking models are derived, is presented. By regularizing the model's parameters, we construct a continuum class of ranking models that bridges the Babington Smith model and the Bradley-Terry-Mallows model. Through the regularization, we can account for an unusual characteristic such as intransitivity of preferences in the model. We also propose the computational algorithm based on contrastive divergence to estimate the parameters in our model and investigate its convergence property of the algorithm.

EO111 Room R07 RECENT ADVANCES IN COMPLEX TIME SERIES ANALYSIS Chair: Haeran Cho

E0434: **Collective anomaly detection in high-dimensional VAR models**
Presenter: **HyeYoung Maeng**, Lancaster University, United Kingdom
Co-authors: Paul Fearnhead, Idris Eckley

There is an increasing interest in detecting collective anomalies: potentially short periods of time where the features of data change before reverting to normal behaviour. We propose a new method for detecting a collective anomaly in VAR models. Our focus is on situations where the change in the VAR coefficient matrix at an anomaly is sparse (i.e. a small number of entries of the VAR coefficient matrix change). To tackle this problem, we propose a test statistic for a local segment built on the lasso estimator of the change in model parameters. This enables us to detect a sparse change more efficiently, and our lasso-based approach becomes especially advantageous when the anomalous interval is short. We show that the new procedure controls Type 1 error and has asymptotic power tending to one. The practicality of our approach is demonstrated through simulations and two data examples involving New York taxi trip data and EEG data.

E0455: **Threshold block fused lasso method for break point detection in high-dimensional time series models**
Presenter: **Abolfazl Safikhani**, University of Florida, United States
Co-authors: Abolfazl Safikhani, Yue Bai, George Michailidis

Fused lasso has been used as a powerful algorithm for anomaly detection in time series literature. However, it is well known that fused lasso cannot consistently perform parameter estimation or breakpoints detection. That is why it needs to be coupled with additional steps to yield satisfactory performance in estimation/detection. We propose a threshold block fused lasso (TBFL) procedure in which the parameters can only change at block endpoints instead of every time points as in fused lasso. We show that with proper choice of block sizes, TBFL achieves consistent modeling parameter estimates. Moreover, with hard thresholding of the estimated jumps, TBFL consistently estimates the number of breakpoints and their relative locations. The performance of TBFL for finite samples is investigated through extensive simulation studies and real data example.

E0496: **Changepoint detection for complex time series data**
Presenter: **Euan McGonigle**, University of Bristol, United Kingdom
Co-authors: Haeran Cho

A method is proposed to detect multiple changepoints in the mean of a univariate time series in the presence of an autocorrelated, time-varying, and possibly non-Gaussian error structure. A common approach for detecting mean changes is to utilise localised testing procedures, such as those based on cumulative sums (CUSUMs) or moving sums (MOSUM). In such cases, the test statistics must be scaled appropriately, using an estimate of the long-run variance or related quantity, in order to distinguish between genuine changepoints and fluctuations due to autocorrelation. We propose a robust wavelet-based approach to estimate the necessary scaling factors of test statistics used in localised testing procedures. We describe how our methodology can be utilised within well-studied changepoint detection methods, for example, the MOSUM procedure. This enables the consistent estimation of the location and number of mean changepoints under general time series error settings. The proposed method is shown to perform well in a variety of changepoint and error scenarios via a simulation study. The method is also applied to a data application in the environmental sciences, highlighting the potential uses.

E0698: **Temporal aggregation on locally stationary time series**
Presenter: **Chao Zheng**, University of Southampton, United Kingdom

Locally stationary processes, as an important type of the non-stationary time series in which data are modelled as locally approximately stationary, has attracted a lot of attention. We consider the temporal aggregation and systematic sampling of a locally stationary process, which transfer an original time series at high frequency to be available only at every certain period. We show that the local stationarity is preserved during these aggregations. Also, we investigate the forecasting performance of original versus aggregated processes with respect to the best linear predictor and give ratio consistency results. This provides practitioners motivations to only look at the aggregated series if the scale of the original ones is too expensive to predict. We also discuss how to select the aggregation window and the parameters for linear prediction and assess the aggregation effect via extensive numerical experiments.

EC332 Room R06 CONTRIBUTIONS IN ECONOMETRIC MODELLING Chair: Alexandra Soberon

E0718: **A semiparametric panel data model with common factors and spatial dependence: The knowledge production function**
Presenter: **Alexandra Soberon**, Universidad de Cantabria, Spain
Co-authors: Juan Manuel Rodriguez-Poo, Antonio Musolesi

A new semiparametric heterogeneous panel data model is proposed which simultaneously handles complex and relevant empirical problems: (i)functional misspecification by modelling stochastic observed common factors with a nonparametric function instead of assuming the usual parametric form,(ii) cross-sectional dependence arising simultaneously from common factors and spatial dependence, for the latter neither imposing a specific parametric spatial diffusion process nor requiring the specification of a given interaction matrix, but being directly derived from the data, (iii)heterogeneous relations. We propose a new estimator that extends the common correlated effect (CCE) approach to such a semiparametric spatially augmented framework. Asymptotic normal distributions are derived when both the time and cross-sectional dimension are large. Small sample properties of the estimators are investigated by Monte Carlo experiments. An empirical application on the knowledge capital production function is conducted.

E0569: **Negatively valued functions in economic modeling**
Presenter: **Christos Kountzakis**, University of the Aegean, Greece
Co-authors: Konstantinos Gkillas

Production functions are mathematical representations of the relation of some inputs (e.g. capital and labor) and output. They offer an account of the forces making for changes in these determinants of the level of the output and/or details about the quantitative relationships between inputs and output. Production externalities could be negative, positive, or both. Negative externalities occur when the production of a good imposes a harmful/external cost on third parties, e.g., in society. In particular, we propose altering the function by including a factor representing an externality. The proposed formula allows the outcome to be negative without affecting the balance of inputs in the products production function, and marginal increases in inputs do not allow having a magnified impact on the output to become positive (e.g., idled inputs to be put back into use). We establish a model for the effect of the external factor on the output by relaxing the assumption that the output level is strictly positive. From an econometric point of view, we propose a simple yet novel solution to this issue to treat the econometric issues of estimation. To this end, we suggest representing complex arguments by simpler, more intuitive arguments in polar coordinates, which, in turn, allow us to estimate the parameters of a model via conventional econometric methods.

E0583: **Cryptoassets markets as dynamic networks: Investigating connectedness, investment horizons, and systemic risk**
Presenter: **Jan Sila**, Univerzita Karlova, Czech Republic
Co-authors: Ladislav Kristoufek

Cryptoassets are often believed to be majorly driven by Bitcoin as the clear market leader. As the market has been rapidly developing and restructuring in time, a comprehensive examination covering this time variability and investors' temporal structure is still missing. We study a portfolio of 11 major cryptoassets between 2016 and 2021 utilizing a novel methodology of the dynamic networks and describe the time-frequency dynamics of the system connectedness. We show that cryptoassets only exhibit some causal structure at times of significant upward movements, such as at the end of 2017 and the end of 2020. Otherwise, the system is driven purely by contemporaneous (not causal) correlations in returns and volatility. The shocks to the system are absorbed very quickly, and they do not create medium or long-horizon persistence, which sets the cryptomarkets apart from the standard asset classes.

E0307: **Credit rating downgrade risk on equity returns**
Presenter: **Periklis Brakatsoulas**, Charles University, Faculty of Social Sciences, Czech Republic

An asset pricing model is developed to capture credit rating downgrade risk and suggest a new methodology to generate firm-level downgrade probabilities. Using credit transition matrices and rating histories from US issuers, we provide empirical evidence for a statistically significant positive downgrade risk premium. Stocks at a higher risk of failure tend to deliver higher returns. The performance of the model remains robust across several panel data estimation methods. Panel Granger causality test results further indicate a Granger-causal relationship from credit rating transition probabilities to excess returns. We thus provide the basis for further development and empirical validation of Fama-French-type models under financial distress.

EG240 Room R02 CONTRIBUTIONS IN NONPARAMETRIC STATISTICS AND APPLICATIONS Chair: Germain Van Bever

E0462: **Pair copula constructions of point-optimal sign-based tests for predictive linear and non-linear regressions**
Presenter: **Kaveh Salehzadeh Nobari**, Lancaster University, United Kingdom

Pair copula constructed point-optimal sign tests are proposed in the context of linear and nonlinear predictive regressions with endogenous, persistent regressors and disturbances exhibiting serial (nonlinear) dependence. The proposed approach entails considering the entire dependence structure of the signs to capture the serial dependence and building feasible test statistics based on pair copula constructions of the sign process. The tests are exact and valid in the presence of heavy-tailed and nonstandard errors, as well as heterogeneous and persistent volatility. Furthermore, they may be inverted to build confidence regions for the parameters of the regression function. Finally, we adopt an adaptive approach based on the split-sample technique to maximize the power of the test by finding an appropriate alternative hypothesis. In a Monte Carlo study, we compare the performance of the proposed "quasi"-point-optimal sign tests based on pair copula constructions by comparing its size and power to those of certain existing tests that are intended to be robust against heteroskedasticity. The simulation results maintain the superiority of our procedures to existing popular tests.

E0692: **Quantile versions of inequality curves and inequality measures**
Presenter: **Alicja Jokiel-Rokita**, Wroclaw University of Science and Technology, Poland

Classical inequality curves and inequality measures are defined for distributions with the finite mean value. We consider various quantile versions of known inequality curves such as the Lorenz curve, the Bonferroni curve and the Zenga curve, and quantile-based versions of the inequality measures such as the Gini index, the Bonferroni index and the Zenga index, respectively. We propose various nonparametric estimators of the curves and indexes considered, and we compare their accuracy in a simulation study. We also illustrate the quantile versions of inequality curves and the quantile-based inequality measures based on real data.

E0730: **Wavelet regression for symplectic data from a Bayesian and nonparametric Bayes perspective**
Presenter: **Andrej Srakar**, University of Ljubljana, Slovenia

Regression for symplectic, i.e. compositional data, has so far been largely considered only from a parametric point of view. It has also been extended to nonparametric situations, introducing local constant and local linear smoothing for regression with compositional data and simplicial splines. We extend their analysis to wavelet regressions, constructing wavelet transforms, deriving father and mother wavelets using Legendre polynomial based sequential approach to orthogonalization. We extend their perspective for wavelet construction in any topological and symplectic space, enabling its usage for modelling compositional data of any dimension. To derive the wavelet regression estimator, we use a Bayesian approach, namely multivariate wavelet priors, evaluating the fit with a recent Stein-based procedure. We extend this further to the nonparametric Bayes perspective using random Bernstein polynomials. The new regression estimators are derived for all three cases: simplicial-real, simplicial-simplicial, and real-simplicial regression. The performance of the estimators is studied in delta-type asymptotic analysis and simulation study. We apply the findings to two economic datasets on income inequality and international trade.

E0562: **From the non-parametric estimation of tail dependence coefficients to portfolio diversification**
Presenter: **Matthieu Garcin**, Leonard de Vinci Pole Universitaire, France
Co-authors: Maxime Nicolas

A theoretical expression is derived for the mean squared error of a non-parametric estimator of tail dependence coefficients (TDC), depending on a threshold that defines which rank corresponds to the tails of a distribution. We propose a new method to select the threshold optimally. It combines the theoretical mean squared error of the estimator and a parametric estimation of the copula linking observations in the tails. Using simulations, we compare this method with the plateau-finding algorithm. We then use the estimated TDCs in a problem of portfolio diversification, in which the TDCs are incorporated in the distance used for clustering a stock dataset.

EG377 Room R03 CONTRIBUTIONS IN SURVIVAL ANALYSIS Chair: Binyan Jiang

E0205: **Analysis of regression discontinuity designs using censored data**
Presenter: **Youngjoo Cho**, Konkuk University, Korea, South
Co-authors: Chen Hu, Debashis Ghosh

In medical studies, the treatment assignment may be determined by a clinically important covariate that predicts patients' survival risk. There is a class of methods from the social science literature known as regression discontinuity (RD) designs that can be used to estimate the treatment effect in this situation. Under certain conditions, such an effect enjoys a causal interpretation. However, few authors have discussed the use of RD for censored data. We show how to estimate causal effects under the regression discontinuity design for censored data. The proposed estimation procedure employs a class of censoring unbiased transformations that includes inverse probability weighting and a doubly robust transformation. Simulation studies demonstrate the utility of the proposed methodology.

E0658: **Hierarchical multi-parameter regression survival models**
Presenter: **Fatima-Zahra Jaouimaa**, University of Limerick, Ireland
Co-authors: Il Do Ha, Kevin Burke

Standard survival models introduce covariates through a single (scale) parameter, and we refer to this standard practice as Single-Parameter Regression (SPR). In contrast, Multi-Parameter Regression (MPR) allows covariates to enter the model through multiple distributional parameters, i.e., scale and shape. Its flexibility has been highlighted in the context of survival data. We extend this to handle multivariate survival data by introducing random effects in both the scale and the shape regression components. We consider various possible dependence structures for these random effects (independent, shared, and correlated) and estimation proceeds using an h-likelihood approach. As the shape parameter may be viewed as a dispersion parameter for log-time, our proposal bears similarities to Double Hierarchical Generalized Linear Modelling (DHGLM). We investigate the performance of our estimation procedure using simulated data and also consider a real data example.

E0664: **Flexible two-piece distributions for right censored survival data**
Presenter: **Worku Biyadgie Ewnetu**, Hasselt University, Belgium
Co-authors: Irene Gijbels, Anneleen Verhasselt

Flexible distributions are essentially used in lifetime data analysis when the data exhibit complex structures such as heavy, light tails, only partial information is observed (censored data), and the hazard function shows nonmonotonic shapes. In recent years, a large family of asymmetric distributions and maximum likelihood estimation for the parameters in that family has been studied in the complete data case. We extend this family of two-piece asymmetric distributions for modelling randomly right-censored data. The flexible two-piece asymmetric distributions can be generated by employing various symmetric about the origin distributions and known monotonic link functions. One of the interesting features of the proposed family is that the location parameter coincides with a specific quantile of the distribution. Given flexibility, this three-parameter family, along with the known link functions, is suitable to encapsulate different shapes of the hazard function (increasing, decreasing, bathtub and upside-down bathtub or unimodal shapes). Maximum likelihood estimators under nonstandard conditions are studied, and their asymptotic properties are established. The finite-sample performance of the estimators is investigated via simulations, and the methodology is illustrated on two real data examples.

E0685: **Penalized multi-parameter regression modelling**

Presenter: **Laura McQuaid**, University of Limerick, Ireland

Co-authors: Kevin Burke

Multi-parameter regression (MPR) modelling refers to the approach whereby covariates enter a parametric model through multiple distributional parameters simultaneously (e.g., scale and shape parameters), allowing more complex covariate effects to be captured. On the other hand, penalized estimation procedures such as the least absolute shrinkage and selection operator (LASSO) and adaptive LASSO are commonly used to perform continuous variable selection - but they have primarily been developed for classical regression problems where covariates enter only through a single distributional parameter. Therefore, we develop a penalized MPR modelling framework and investigate its performance through simulation studies and real data analysis. We consider the application area of survival analysis, but the methodology can equally be applied to other areas.

Authors Index