PROGRAMME AND ABSTRACTS

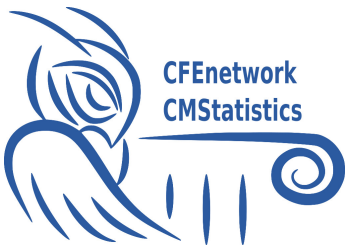
7th International Conference on

Econometrics and Statistics (EcoSta 2024)

http://cmstatistics.org/EcoSta2024

Beijing Normal University, Beijing, China

17-19 July 2024



ISBN: 978-9925-7812-3-2

Co-chairs:

Lixing Zhu, Massimiliano Caporin, Qing Mai and Catherine Liu.

EcoSta Editors:

Ana Colubi and Erricos J. Kontoghiorghes.

Scientific Programme Committee:

Elif Acar, Manabu Asai, Xuan Bi, Stefano Castruccio, Likai Chen, Yichen Cheng, Boris Choy, Xinwei Deng, Xiucai Ding, Wei Huang, Donggyu Kim, Pavel Krupskiy, Degui Li, Gen Li, Tianxi Li, Ting Li, Wai Keung Li, Xinyi Li, Chang-Yun Lin, Min-Qian Liu, Miles Lopes, Monia Lupparelli, Rogemar Mamon, Cristina Mollica, Sang-Yun Oh, Yu Philip, Matias Quiroz, Marco Reale, Zhao Ren, Stefano Rizzelli, Katsumi Shimotsu, CY Sin, Ekaterina (Katja) Smetanina, Eftychia Solea, Gilles Stupfler, Baoluo Sun, Masayuki Uchida, Boxiang Wang, Cheng Wang, Ning Wang, Runmin Wang, Tao Wang, Wanjie Wang, Weichi Wu, Wenbo Wu, Zhenke Wu, Dong Xia, Lingzhou Xue, Weixin Yao, Yiming Ying, Shan Yu, Jing Zeng, Emma Jingfei Zhang, Ting Zhang, Wei Zheng, Le Zhou and Hongxiao Zhu.

Local Organizing Committee:

School of Statistics of Beijing Normal University, EcoSta, CMStatistics and CFEnetwork.

Dear Colleagues,

It is a great pleasure to welcome you to the 7th International Conference on Econometrics and Statistics (EcoSta 2024). Following previous editions, the conference is held in a hybrid format that accommodates both in-person and virtual attendance, ensuring flexibility for participants based on their circumstances and local restrictions. The conference program has been thoughtfully tailored to facilitate the optimal presentation of research findings and networking opportunities.

EcoSta 2024 comprises about 230 sessions, three keynote talks, four invited sessions, and 885 presentations. These figures serve as a testament to the support of our research communities, highlighting the significance of this initiative. We trust that the EcoSta conference will continue to serve as an excellent platform for disseminating high-quality research in Econometrics and Statistics while fostering valuable networking opportunities.

The conference is jointly organized by the Computational and Methodological Statistics Working Group (CMStatistics), the Computational and Financial Econometrics Network (CFEnetwork), the journal Econometrics and Statistics (EcoSta), and Beijing Normal University. Building upon the achievements of previous editions, our aim is for this conference to become a leading event in the field of econometrics, statistics, and their applications.

The co-chairs express their gratitude to the scientific program committee, session organizers, and local organizing committee for their collective efforts in delivering a comprehensive program that covers various areas of econometrics and statistics. The School of Statistics at Beijing Normal University, as local hosts, along with the IFMSE ((International Foundation of Methodological Statistics & Econometrics) dedicated assistants and staff, has played an instrumental role in ensuring the smooth organization of the conference. We extend our heartfelt thanks to all of them for their invaluable support.

The Elsevier journals of Econometrics and Statistics (EcoSta) and Computational Statistics & Data Analysis (CSDA) are associated with CFEnetwork, CMStatistics, and the EcoSta 2024 conference. The participants are encouraged to join the networks and submit their papers to special or regular peer-reviewed EcoSta and the CSDA Annals of Statistical Data Science (SDS) issues.

The inaugural impact factor for the journal Econometrics and Statistics (EcoSta) in 2022, announced in June 2023, stands at 1.9. Meanwhile, Computational Statistics & Data Analysis (CSDA) continues to maintain its commendable and consistent performance, with an impact factor of 1.8 for 2022.

Finally, it is exciting to announce that the 8th International Conference on Econometrics and Statistics (EcoSta 2025) will take place at Waseda University, Tokyo, Japan, from 21-23 August 2025. The reputable Waseda University benefits from impeccable infrastructure, and the local host, Prof. Yan Liu, has previously hosted a very successful conference. A heartfelt invitation and enthusiastic encouragement for active participation in EcoSta 2025 are extended.

Ana Colubi and Erricos J. Kontoghiorghes.

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 (Beijing time, UTC+8)

2024-07-17

Opening, 08:45 - 09:00

A - Keynote
EcoSta2024
09:00 - 09:50

Coffee break
09:50 - 10:20

B
EcoSta2024
10:20 - 12:00

Lunch break
12:00 - 13:30

C
EcoSta2024
13:30 - 15:10

Coffee break
15:10 - 15:40

D - Keynote
EcoSta2024
15:40 - 16:30

E
EcoSta2024
16:40 - 18:20

Welcome reception
19:00 - 20:30

2024-07-18

F
EcoSta2024
08:15 - 09:55

Coffee break
09:55 - 10:25

G
EcoSta2024
10:25 - 12:30

Lunch break
12:30 - 14:00

H
EcoSta2024
14:00 - 15:40

Coffee break
15:40 - 16:10

I
EcoSta2024
16:10 - 18:15

Conference dinner
19:30 - 22:00

2024-07-19

K
EcoSta2024
08:15 - 09:55

Coffee break
09:55 - 10:25

L
EcoSta2024
10:25 - 11:40

Lunch break
11:40 - 13:10

N
EcoSta2024
13:10 - 14:25

Coffee break
14:25 - 14:55

O
EcoSta2024
14:55 - 16:35

P - Keynote
EcoSta2024
16:45 - 17:35

Closing, 17:35 - 17:50

ACCESS TO THE CONFERENCE

- Participants can attend virtually or in person according to what they selected while registering.
- The in-person venue is Beijing Normal University, 19 Xinwai Ave, Beitaipingzhuang, Hai Dian Qu, Bei Jing Shi, China, 100875. Participants must enter the campus through the South Gate or the East Gate (see maps on page VIII). They will need to show their **passport** or the conference badge for that.
- Please note that the university opens at 8:00, so you will not be able to enter before that time.
- The keynote talks will take place in the Yingdong Education Building (see maps on page VIII) while the parallel sessions will take place in the Teaching Building No. 4 (see maps on page VIII and floor maps on page IX).
- The registration and coffee breaks will be located in the Yingdong Education Building (see maps on page VIII).
- For environmental sustainability reasons, the conference endeavors to minimize paper usage and overall consumption. While there will be a limited number of printed Books of Abstracts, bags, pens, and pads available for those who require them, we strongly encourage all participants to opt for digital materials by downloading them onto their personal devices. For those who do utilize printed materials, we kindly request that they be returned after use to be reused or recycled. QR codes will be displayed in the registration area. These codes will enable participants to quickly access essential information, further reducing the need for printed materials and promoting a paperless conference experience.
- The conference is live-streaming through Zoom, and it will not be recorded. The conference programme time is set at UTC+8.
- In order to access the virtual conference, you must first log in to the registration tool, get the daily password there and leave the session open. Then you should open another tab and go to the interactive programme (schedule). Click on the slot you wish to attend and then on the session. You will be redirected to Zoom, where you will need to use the daily password.
- Please note that for security reasons, the Zoom links will not be sent to the speakers, and they can only be found on the online interactive programme (schedule).
- More detailed instructions for virtual and in-person attendance, hybrid sessions, speakers, chairs, posters, networking, test sessions, as well as FAQ, can be found on the webpage.

TUTORIAL

The tutorial "Prediction-based statistical inference for multiple time series" will take place on Tuesday, the 16th of July 2024, 15:00-19:30 (UTC+8) in the Teaching Building No. 4, Room 108 (see maps on page VIII and floor maps on page IX). It will be delivered by Prof. Yan Liu, Waseda University, Japan. Only participants who have subscribed for the tutorial can attend. Registered participants will be able to access the tutorial either in person or virtually. The instructions to join online can be found on the website.

SOCIAL EVENTS

- The **lunches** for those who subscribed will take place in the Lanhui restaurant (see maps on page VIII). Information about the purchased lunches is embedded in the QR code on the conference badge. Participants must bring their conference badges in order to attend the lunches.
- **Welcome reception, 17th of July from 19:00 to 20:30:** Participants must bring their conference badges to attend the welcome reception. It will take place on the 4th floor of the hotel Holiday Inn Beijing Deshengmen, 71 Deshengmenwai Street, Xicheng District, Beijing (see maps on page VIII).
- **Conference dinner, Thursday, 18th of July, at 19:30:** The conference dinner is optional and registration is required. It will take place at the restaurant Qinghelou, Xinkang Road 2, Xicheng District, Beijing (see maps on page VIII). Participants must bring their conference badges to attend the conference dinner. Information about the purchased conference dinner ticket is embedded in the QR code on the conference badge.

Presentation instructions

Presentations must be shared through Zoom by all virtual and in-person speakers. The in-person speakers will use the room PC for that. The room PC will be connected to the corresponding virtual room on Zoom. Virtual speakers should have a stable internet connection, and make sure their video and audio work on Zoom. They will share their slides when the chair requires it, present their talk, and be ready to answer the questions after the presentation. Detailed instructions for speakers can be found on the website. Each speaker has 20 minutes for the talk and 3-4 minutes for discussion as a general rule. Strict timing must be observed.

Posters

The poster sessions will take place on Zoom. Posters do not need to be uploaded or sent in advance. After entering the poster room, presenters must select the breakout room with their poster code (e.g. E0123), share the poster and remain with the chat, camera, microphone, and audio ready throughout the entire session.

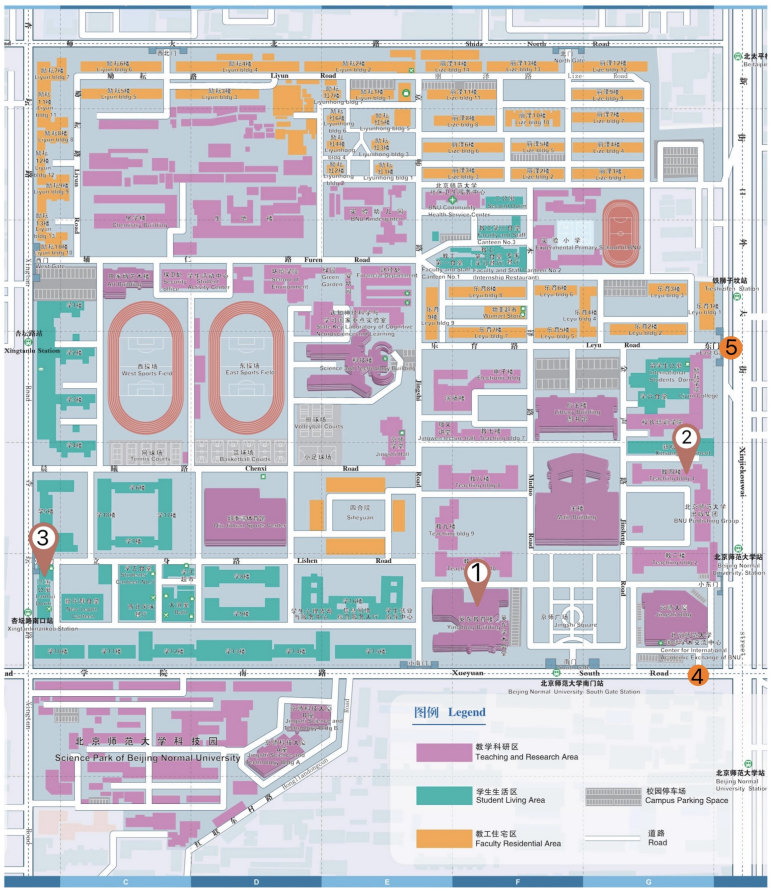
Session chairs

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 instructions for the (virtual or in-person) session chairs can be found on the website.

Test session

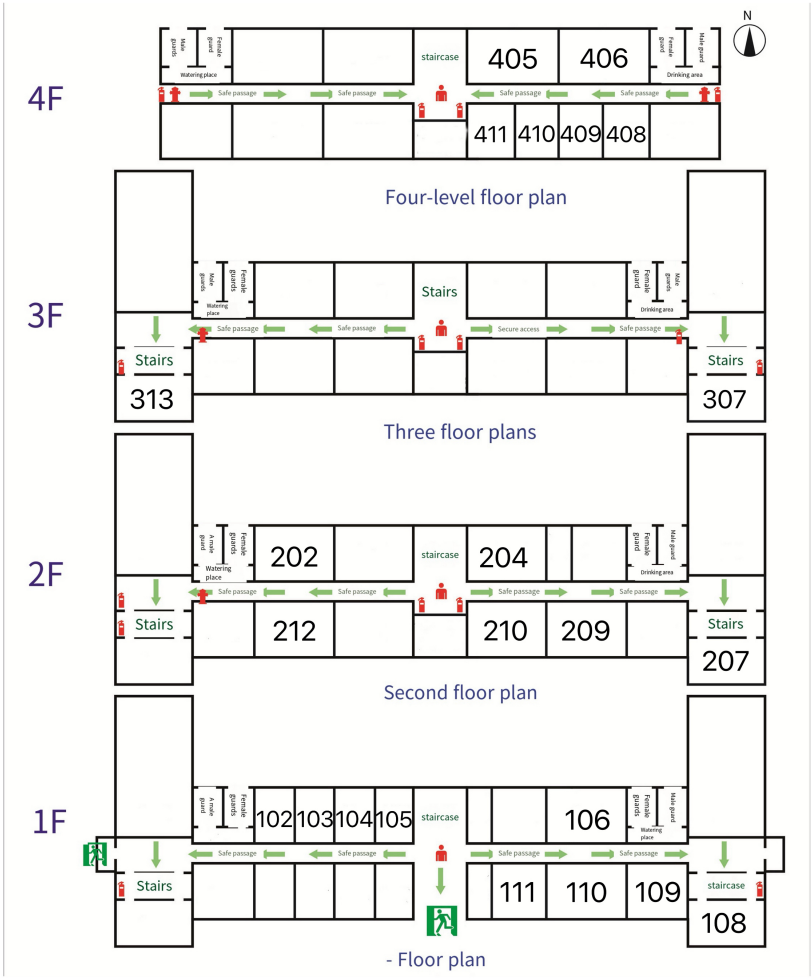
A virtual test session will be set up for Saturday the 12th of July 2024, from 15:00 to 15:30 UTC+8 (Beijing time). The participants will be able to join online through the Room 108 in the programme (e.g., through Parallel session B) to test their presentations, video, microphone and audio. Detailed indications for the test session can be found on the website.

Map of the venue and nearby area



1.Beijing Normal University-Yingdong Education Building (Coffee break/Keynote Speech/Registration)
2.Beijing Normal University-Teaching Building No.4 (Sessions)
3. Beijing Normal University-Lanhui Restaurant (Lunch)
4.South Gate
5.East Gate

Floor maps



4F :
For 16 people : room 408, 409, 410, 411
For 30 people : room 405, 406

3F :
For 30 people : room 313
For 35 people : room 307

2F :
For 20 people : room 212
For 30 people : room 202
For 35 people : room 204, 207, 209, 210

1F :
For 24 people : room 102, 103, 104, 105
For 40 people : room 111
For 60 people : room 106, 108, 109, 110

Teaching Building No. 4

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 & 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.

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

Wednesday 17.07.2024 09:00 - 09:50 Room: Auditorium Chair: Lixing Zhu Keynote talk I

Statistical generative learning leveraging pretrained large models

Speaker: **Jian Huang, The Hong Kong Polytechnic University, China**

The focus is on statistical generative models leveraging pre-trained large models. The basics of two generative learning approaches are first introduced: generative adversarial networks and denoising diffusion models. Two examples are then used to illustrate generative modeling with the help of a pre-trained large model. The first example considers protein data analysis based on data representations learned through a pre-trained model using a protein sequence database. The second example demonstrates a Bayesian fine-tuning approach for image generation leveraging a pretrained diffusion model.

Wednesday 17.07.2024 15:40 - 16:30 Room: Auditorium Chair: Catherine Liu Keynote talk II

Post-mortem interval: A functional data analysis for criminology

Speaker: **Frederic Ferraty, Mathematics Institute of Toulouse, France** Davide Pigoli, John Aston, Anjali Mazumder, Martin Hall, Cameron Richards

When a body is discovered at a crime scene, it is necessary to determine the time since death or, more formally, the post-mortem interval (PMI). If the body has been left outdoors for a long period, forensic entomology can estimate this post-mortem interval by examining evidence obtained from the growth of insect larvae on the body. The hatching period of the larvae plays an important role in this methodology as it corresponds to the date of abandonment of the corpse (the latter being assumed to be close to the victim's death). A method is proposed for estimating the hatching date of larvae (or maggots) based on their length, the temperature profile of the crime scene, and experimental data on larval development. This method requires the estimation of a time-dependent growth curve from experiments where the larvae were exposed to a relatively small number of constant temperature profiles. As temperature influences the rate of development, a crucial step is the temporal alignment of the curves at different temperatures. A dynamic model for the time-varying temperature profiles is then proposed based on the local growth rate estimated from the experimental data. This allows for the determination of the most likely time of hatching (and hence the PMI) for a sample of larvae from the crime scene of two criminal cases.

Friday 19.07.2024 16:45 - 17:35 Room: Auditorium Chair: Yan Liu Keynote talk III

Identifying impulse responses with instrumental variables

Speaker: **Myung Hwan Seo, Seoul National University, Korea, South** Bonsoo Koo, Seojeong Jay Lee

Macro shocks are often composites, yet their implications are overlooked in impulse response analysis. When an instrumental variable (IV) is used to identify a composite shock, it violates the common IV exclusion restriction. It is shown that the local projection-IV estimand is represented as a weighted average of component-wise impulse responses, but with possibly negative weights, which occur when the IV and shock components have opposite signs of correlation. An LP-IV estimand with negative weights does not have a causal interpretation. However, when combined appropriately with other LP-IV estimands, such a non-causal LP-IV estimand can provide identification of the component-wise impulse responses. Identification strategies are developed based on additional granular information or sign restrictions. In contrast, conventional approaches combining multiple instruments, such as two-stage least squares, do not provide identification of structural parameters. Applications confirm the composite nature of monetary policy shocks and reveal a non-defense spending multiplier exceeding one.

Wednesday 17.07.2024 10:20 - 12:00 Parallel Session B – EcoSta2024

EO010 Room 102 STATISTICS AND MACHINE LEARNING FOR FINANCIAL TIME SERIES AND INSURANCE DATA Chair: Yuning Zhang

E0888: **A new test for checking stationarity in variance for nonlinear time series with a trend**
Presenter: **Li Cai**, Zhejiang Gongshang University, China
Co-authors: Lei Jin
The assumption of constant variance is fundamental in numerous statistical procedures for time series analysis. A novel procedure, such as GARCH models and Markov-switching models, is introduced to assess the variance stationarity of nonlinear time series. Unlike others, the proposed test relies on systematic sampling via Walsh transformations. It is developed under the process with a nonconstant mean function. Asymptotic pairwise independence is established among various Walsh transformation coefficients, and a max-type statistic is defined. The asymptotic null distributions of the test statistics are obtained. Furthermore, the consistency of the proposed methods is established under a sequence of local alternatives, contingent upon additional conditions on the mean functions. Through a comprehensive simulation study, we evaluate the finite sample performance of the procedure and conduct comparisons with existing methodologies. The findings offer insights into analyzing financial time series data.

E0891: **DeepVol: A pre-trained universal asset volatility model**
Presenter: **Chao Wang**, The University of Sydney, Australia
Co-authors: Minh-Ngoc Tran, Richard Gerlach, Robert Kohn
DeepVol is a pre-trained deep-learning volatility model is introduced, which is more general than traditional econometric models. DeepVol leverages the power of transfer learning to effectively capture and model the volatility dynamics of all financial assets, including previously unseen ones, using a single universal model. This contrasts to the usual practice in the econometrics literature, which trains a separate model for each asset. The introduction of DeepVol opens up new avenues for volatility modeling in the finance industry, potentially transforming the way volatility is predicted.

E0928: **Portfolio optimization through regular vine copula model and computational intelligence method**
Presenter: **Nuttanan Wichitaksorn**, Auckland University of Technology, New Zealand
A new approach to portfolio investment strategy is introduced by combining the artificial immune system and genetic algorithms in computational intelligence with sentiment analysis. The key component of this strategy is a multivariate regular vine copula-based model with various bivariate copula functions where the marginal model is an intertemporal capital asset pricing with asymmetric exponential generalized autoregressive conditional heteroscedasticity models having a mixture of Gaussian distribution and two generalized Pareto distributions as the innovation. A parallel computing technique is applied to the proposed evolutionary algorithms to accelerate the convergence. In the empirical analysis, two scenarios, with and without the COVID-19 period, are investigated for the dependence structure in the financial markets. In the portfolios, a class of cryptocurrencies is included with the traditional stocks. The proposed model outperforms the benchmark models in terms of risk rewards and diversity indicators.

E0878: **Bayesian bi-directional self-exciting threshold autoregressive model and the application in loss reserving**
Presenter: **Yuning Zhang**, The University of Sydney Business School, Australia
Co-authors: Boris Choy, Wilson Ye Chen, Tak Kuen Siu
A Bayesian statistical approach is proposed for estimating a self-exciting threshold autoregressive model (SETAR) in bidirectional time series (bi-SETAR). While the frequentist SETAR, adapted into a bidirectional framework (SETAR-NN), has recently been utilized for claim reserving in run-off triangles, the proposed Bayesian approach introduces a more flexible and practical methodology for analyzing and estimating the SETAR-NN model. This approach focuses on providing probabilistic estimates for the structural parameters, emphasizing the threshold parameters. The Markov Chain Monte Carlo (MCMC) method is employed to simulate the posterior distributions of unknown parameters and predictive distributions. Applications in loss reserving and computing risk metrics are also demonstrated, and they are compared against the results from benchmark models.

EO157 Room 103 ADVANCES IN REGRESSION AND STOCHASTIC FRONTIER ANALYSIS WITH PANEL DATA Chair: Taining Wang

E0692: **A system approach to structural identification of production functions with multidimensional productivity spillovers**
Presenter: **Zhezhi Hou**, Southwestern University of Finance and Economics, China
Many recent researchers in production economics believe technological growth can enhance the productivity of all factors of production equally, or they can exhibit bias toward specific factors. A novel methodology is provided on the estimation of labor augmenting and Hicks Neutral productivity of firms in the presence of productivity spillovers originating from two sources - the spatiotemporal spillovers in Hicks Neutral productivity and in their labor augmenting counterpart. Previous work is extended by allowing these two forms of productivity to both follow the Markov process, which could be affected by the cross-sectional dependence in multidimensional productivity induced by spillovers and then develop an identification strategy based on a proxy variable method and first-order conditions.

E0912: **An estimator of a jump discontinuity in regression based on generated observations**
Presenter: **Feng Yao**, West Virginia University, United States
A new class of estimators is proposed for a jump discontinuity on nonparametric regression. While a vast amount of econometrics literature addresses this issue, the main approach in these studies is to use local polynomial (linear) estimators on both sides of the discontinuity to produce an estimator for the jump with desirable boundary properties. The approach extends the regression with generated observations from both sides of the discontinuity using a theorem of Hestenes. The extended regressions' generated observations are estimated and used to construct an estimator for the jump discontinuity that solves the boundary problems normally associated with classical kernel estimators. Asymptotic characterizations are provided for the jump estimators, including bias and variance orders and asymptotic distributions after suitable centering and normalization. Monte Carlo simulations show that the estimator for the jump can outperform that based on local polynomial (linear) regression.

E0923: **Understanding real estate matches through a semiparametric panel model of the matching function**
Presenter: **Taining Wang**, Capital University of Economics and Business, China
Co-authors: Feng Yao
A semiparametric panel model of the matching function is proposed, extending the conventional (log) Cobb-Douglas function to allow coefficients to vary with environmental variables. The model captures various unobserved searching frictions during the matching process by accounting for latent heterogeneity in both cross-sectional units and unobserved common shocks over time. Furthermore, the model allows for significant flexibility, enabling varying coefficients to appear in both constant and time-varying regressors (supply) and accommodating different environmental variables (mortgage rates and unemployment rates) in distinct coefficient functions. A two-step estimator is proposed without requiring a normalization of the fixed effects. The first step estimates the varying coefficients with series-based estimators, eliminating fixed effects through multiple

differencing. The second step performs a one-step kernel backfitting to improve the estimation efficiency. It is demonstrated through Monte Carlo simulations that the estimators are computationally efficient and perform well relative to a profile-based kernel estimator. Matches in the U.S. real estate market over 100 metro cities during 2012-2018 are studied. It is found that nonlinear effects of the mortgage and unemployment rates decrease the matching elasticity of housing sellers and buyers, and the magnitude of the effect varies with matching efficiency.

E1070: **Nonparametric panel data estimators of the nonparametric panel stochastic frontier analysis**
Presenter: **Kai Sun**, Shanghai University, China
Co-authors: Xin Geng
Procedures are proposed for estimating individual effects panel stochastic frontier models where the frontier function is unknown via nonparametric panel data estimators. The unobserved individual heterogeneity is decomposed into (i) individual effects that affect output through the frontier function and (ii) time-invariant technical inefficiency that affects output through the inefficiency function. The individual effects can be either fixed or random: they can affect output either as non-stochastic individual characteristics or as an individual-specific random shock. When the individual effects are random, the associated idiosyncratic term in the context of stochastic frontier analysis (SFA) is shown to be heteroskedastic, and therefore, a modified nonparametric random effects estimator of the conditional variance of the idiosyncratic term is also proposed. To guide practitioners to deciding between fixed versus random effects, a Hausman-type test statistic with a bootstrap procedure that works under the SFA setting is provided. Simulations show that the nonparametric panel data estimators and testing procedures perform well in finite samples. Finally, the panel data framework is applied to estimating a panel stochastic frontier model where the base inefficiency follows a half-normal or exponential distribution.

EO019 Room 104 RECENT DEVELOPMENT IN HIGH-DIMENSIONAL INFERENCE AND MODELING Chair: Wenbo Wu

E0334: **CoxKnockoff: Controlled feature selection for the Cox model using knockoffs**
Presenter: **Daoji Li**, California State University Fullerton, United States
Although there is a huge literature on feature selection for the Cox model, none of the existing approaches can control the false discovery rate (FDR) unless the sample size tends to infinity. In addition, there is no formal power analysis of the knockoffs framework for survival data in the literature. To address those issues, a novel controlled feature selection approach is proposed using knockoffs for the Cox model. The proposed method is established to enjoy the FDR control in finite samples regardless of the number of covariates. Moreover, under mild regularity conditions, the power of the method is also shown to be asymptotically one as the sample size tends to infinity. To the best of knowledge, this is the first formal theoretical result on the power of the knockoffs procedure in the survival setting. Simulation studies confirm that the method has appealing finite-sample performance with desired FDR control and high power. The performance of the method is further demonstrated through a real data example.

E0830: **Reimaging semi-competing risks data analysis: Enhancing variable selection with preliminary dimension reduction**
Presenter: **Chenlu Ke**, Virginia Commonwealth University, United States
A new framework is introduced with an efficient algorithm for feature screening in the challenging context of ultrahigh dimensional semi-competing risk data. Specifically, the two-stage procedure initially employs a dual screening mechanism to select a coarse set of features that are potentially relevant to both terminal and nonterminal endpoints. This leads to the estimation of the augmented central subspace, pivotal for both endpoints and censoring, based on the selected features. In the second stage, refined sets of important features for the nonterminal and terminal events, respectively, are further identified using an inverse probability-of-censoring weighted filter, where the central subspace estimator is used to obtain the weights adjusting for censoring. The proposed framework is model-free, and it does not require independent censoring. Asymptotic properties are established under minor assumptions. The promising performance of the proposed method is demonstrated through simulations and gene expression data analysis.

E0910: **FDR control for high dimensional quantile variable selection**
Presenter: **Tianhai Zu**, University of Texas at San Antonio, United States
Co-authors: Zhigen Zhao, Yan Yu
Multiple testing is a significant challenge in genetic research, especially when investigating complex diseases. Quantile regression is increasingly critical for providing a nuanced understanding of heterogeneous relationships between genetic markers and complex conditions like diabetes. However, existing mechanisms for false discovery rate (FDR) control are not tailored to the quantile regression framework. To tackle these challenges, a novel FDR control method is proposed for linear quantile regression, utilizing data-splitting mirror statistics. The approach addresses current limitations in existing FDR control methods for quantile regression and is especially advantageous in preserving high power. Theoretical justifications are provided, highlighting that this is the first attempt for controlling FDR in linear quantile regression. Extensive simulations confirm the efficacy of the approach. Furthermore, its use case is demonstrated through a case study on diabetes data, with particular emphasis on high-risk quantiles. The method effectively identifies genetic factors across various diabetes risk quantiles that may benefit improved diagnostics and treatments.

E0591: **Mediation analysis with ultra-high dimensional confounders for the study on geriatric depression and Alzheimer's disease**
Presenter: **Yuexia Zhang**, The University of Texas at San Antonio, United States
Co-authors: Annie Qu, Yubai Yuan, Qi Xu, Fei Xue, Kecheng Wei
Depression and Alzheimer's disease (AD) are both prevalent diseases in older adults. Using the data sets from the Alzheimer's disease neuroimaging initiative (ADNI) study, whether geriatric depression is explored has a significant average treatment effect on AD and whether the effect is mediated by some important mediators. To estimate these causal effects consistently, ultra-high dimensional potential confounders are controlled for, including DNA methylation levels. A new ball correlation-based screening method is proposed for confounder selection in mediation analysis. A robust mediation analysis framework is utilized to achieve robustness against model misspecification. Simulation studies show that the proposed method has good finite-sample performance in terms of confounder and mediator selection, effect estimation, and inference. In the real data analysis, it is found that geriatric depression has a significantly positive causal effect on AD. New prevention and treatment strategies are also proposed for geriatric depression and AD by changing the selected confounders and mediators.

EO018 Room 105 CURRENT STATISTICAL INNOVATIONS IN INDUSTRIAL AND APPLIED STATISTICS Chair: Tsung-Jen Shen

E0917: **Robust group testing for prevalence estimation against uncertain test error mechanisms**
Presenter: **Shih-Hao Huang**, National Central University, Taiwan
In group testing applications, a primary focus is to estimate the prevalence of a trait in the presence of testing errors. However, the incorporation of testing error models may introduce substantial bias into prevalence estimation when misspecified and inflates variance when involving additional parameters. To mitigate these challenges, a robust estimation method is proposed within group testing frameworks to alleviate model misspecification bias. Additionally, efficient design algorithms are developed for data collection to complement the estimation technique, thereby enhancing prevalence estimation by reducing variance. The simulation experiments demonstrate that the approaches usually result in reasonably small mean squared errors compared to conventional ones.

E0964: **Cause-and-effect diagram-based supersaturated designs**
Presenter: **Chang-Yun Lin**, National Chung Hsing University, Taiwan
Supersaturated designs (SSDs) are often used for screening experiments, and cause-and-effect diagrams (CEDs) are useful quality tools that help engineers decide which factors to use in experiments. Based on engineers prior knowledge, certain factors (referred to as the primary factors) in the cause-and-effect diagram are considered more likely to be active than the others (referred to as the potential factors). Due to the unequal importance of the factors, the traditional E(s2) criterion is unsuitable for selecting supersaturated designs. A CED-based approach is proposed to select supersaturated designs that have smaller variances in the estimates of the primary factor effects and less severe aliasing between the effects of the primary and potential factors. Simulation studies show that the proposed method selects supersaturated designs that outperform other supersaturated designs in terms of power and type I error when engineers have accurate prior knowledge.

E0946: **Confirmatory analysis to identify relative importance of regressors for the linear regression model**
Presenter: **Tsung-Chi Cheng**, National Chengchi University, Taiwan
In various application analyses with regard to the linear regression models, the measurement of the individual relative importance of each explanatory variable in the model is of great practical significance and meaningfulness. Among the many approaches of relative importance, dominance analysis measures the relative importance of explanatory variables in an estimated model according to their contribution to the overall model fitting. Dominance analysis is quite intuitive, and its interpretation is very simple and straightforward. However, the so-called complete dominance is considered a purely qualitative comparison, while conditional dominance and general dominance are constructed in the context of quantitative concepts. Based on the framework of dominance analysis, the focus is the statistical hypothesis testing analysis for the confirmation of complete dominance, conditional dominance, and general dominance. The proposed methods are applied to construct the comprehensive measurement of subjective well-being (SWB) and identify those important factors affecting SWB.

E0918: **Predicting encounter distance for new species discovery**
Presenter: **Tsung-Jen Shen**, National Chung Hsing University, Taiwan
The purpose is to investigate the minimum distance needed to encounter the first individual of new species using restricted biodiversity data from a line transect. Understanding this parameter holds practical value in ecological research, particularly in assessing sampling intensity for species discovery. A straightforward estimator is developed to predict this minimal encounter distance and subsequently validate its efficacy through numerical simulations and empirical tests. The results demonstrate the estimator's robust predictive power, providing a reliable tool for researchers seeking to optimize sampling strategies for new species identification.

EO028 Room 106 ADVANCES OF STATISTICAL LEARNING METHODS AND THEIR APPLICATIONS Chair: Xuan Bi

E0841: **PASTA: Pessimistic assortment optimization**
Presenter: **Zhengling Qi**, The George Washington University, United States
A class of assortment optimization problems is considered in an offline data-driven setting. A firm does not know the underlying customer choice model but has access to an offline dataset consisting of the historically offered assortment set, customer choice, and revenue. The objective is to use the offline dataset to find an optimal assortment. Due to the combinatorial nature of assortment optimization, the problem of insufficient data coverage is likely to occur in the offline dataset. Therefore, designing a provably efficient offline learning algorithm becomes a significant challenge. To this end, an algorithm is proposed referred to as Pessimistic ASsortment opTimizAtion (PASTA for short), designed based on the principle of pessimism, that can correctly identify the optimal assortment by only requiring the offline data to cover the optimal assortment under general settings. In particular, a regret bound is established for the offline assortment optimization problem under the celebrated multinomial logit model. An efficient computational procedure is also proposed to solve the pessimistic assortment optimization problem. Numerical studies demonstrate the superiority of the proposed method over the existing baseline method.

E0233: **The non-overlapping approximation to overlapping group lasso**
Presenter: **Tianxi Li**, University of Minnesota, United States
The group lasso penalty is widely used to introduce structured sparsity in statistical learning, characterized by its ability to eliminate predefined groups of parameters automatically. However, when the groups are overlapping, solving the group lasso problem can be time-consuming in high-dimensional settings because of the non-separability induced by the groups. This difficulty has significantly limited the penalty's applicability in cutting-edge computational areas, such as gene pathway selection and graphical model estimation. A non-overlapping and separable penalty is introduced to efficiently approximate the overlapping group lasso penalty. The approximation substantially improves the computational efficiency in optimization, especially for large-scale and high-dimensional problems. It is shown that the proposed penalty is the tightest separable relaxation of the overlapping group lasso norm within a broad family of norms. Furthermore, the estimators based on the proposed norm are statistically equivalent to those derived from the overlapping group lasso in terms of estimation error, support recovery, and minimax rate under the squared loss. The method's effectiveness is demonstrated through extensive simulation examples and a predictive task of cancer tumors.

E0453: **Imaging mediation analysis for longitudinal outcomes**
Presenter: **Cai Li**, St. Jude Children's Research Hospital, United States
The focus is on improving cognitive outcomes for pediatric cancer survivors who undergo aggressive cancer treatments that may affect the central nervous system. Specifically, a new mediation framework is proposed for longitudinal neurocognitive outcomes pertaining to a clinical trial for medulloblastoma, the most common malignant brain tumour in children, using high-dimensional imaging mediators to identify causal pathways and corresponding white matter microstructures. The proposed approach takes into account both the spatial and temporal dependencies and smoothness of the mediators and outcomes, enhancing the detection power of informative voxels and accurately characterizing longitudinal patterns concurrently. The results offer insights into how to enhance long-term neurodevelopment and strategically spare brain regions that might be impacted by radiation therapy. This understanding will be crucial in planning future treatment protocols, ultimately benefiting brain cancer survivors. The validity and effectiveness of the method are affirmed through numerical studies.

E0861: **Powerful and resource-efficient meta-analysis of rare variant associations in large whole genome sequencing studies**
Presenter: **Zilin Li**, Northeast Normal University, China

Meta-analysis of whole-genome/exome sequencing (WGS/WES) studies provides an attractive solution to the problem of collecting large sample sizes for discovering rare variants associated with complex phenotypes. Existing rare variant meta-analysis approaches are not scalable to biobank-scale WGS data. The aim is to present MetaSTAAR, a powerful and resource-efficient rare variant meta-analysis framework for large-scale WGS/WES studies. MetaSTAAR accounts for relatedness and population structure, can analyze quantitative and dichotomous traits, and boosts the power of rare variant tests by incorporating multiple variant functional annotations. Through a meta-analysis of four lipid traits in 30,138 ancestrally diverse samples from 14 studies of the trans-omics for precision medicine (TOPMed) program, it is shown that MetaSTAAR performs rare variant meta-analysis at scale and produces results comparable to using pooled data. Additionally, several conditionally significant rare variant associations are identified with lipid traits. It is further demonstrated that MetaSTAAR is scalable to biobank-scale cohorts through meta-analysis of TOPMed WGS data and UK Biobank WES data of 200,000 samples.

EO140 Room 108 FRONTIERS AT THE INTERSECTION OF STATISTICS AND MACHINE LEARNING Chair: Zhenke Wu

E0632: **A robust angle-based transfer learning**
Presenter: **Tian Gu**, Columbia University, United States

The increasing numbers of large-scale biobanks and institutional data networks have brought unique opportunities to link patient genomics, electronic health records, and survey data for studying complex human diseases, primarily to address the diminished model performance in minority and disadvantaged groups due to their low representation in biomedical studies. A novel angle-based transfer learning (angleTL) method is proposed to improve risk prediction in underrepresented populations by integrating data from multiple biobanks, different ancestries, and related outcomes. It protects data privacy by learning from pre-trained models in external data sources without sharing patient-level data and accounts for potential data heterogeneity. Theoretical guarantees are provided for the model performance and insights regarding when the external model can be helpful to the target model. It is shown that angleTL unifies several benchmark methods by construction, with examples using data from the UK biobank and the electronic medical records and genomics (eMERGE) network.

E0634: **A robust test for the stationarity assumption in sequential decision making: Towards better policy under batch learning**
Presenter: **Zhenke Wu**, University of Michigan, United States

Reinforcement learning (RL) is a powerful technique that allows an autonomous agent to learn an optimal policy to maximize the expected return. The optimality of various RL algorithms relies on the stationarity assumption, which requires time-invariant state transition and reward functions. However, deviations from stationarity over extended periods often occur in real-world applications like robotics control, health care and digital marketing, resulting in sub-optimal policies learned under stationary assumptions. A doubly robust procedure is proposed for testing the stationarity assumption and detecting change points in offline RL settings, e.g., using data obtained from a completed sequentially randomized trial. The proposed testing procedure is robust to model misspecifications and can effectively control type-I error while achieving high statistical power, especially in high-dimensional settings. Simulations and a real-world interventional mobile health example illustrate the advantages of the method in detecting change points and optimizing long-term rewards in high-dimensional, non-stationary environments.

E0717: **Is this model reliable for everyone? Testing for strong calibration**
Presenter: **Jean Feng**, UCSF, United States

In well-calibrated risk prediction models, the average predicted probability is close to the true event rate for any given subgroup. Such models are reliable across heterogeneous populations and satisfy strong notions of algorithmic fairness. However, auditing machine learning models for strong calibration is difficult due to the number of potential subgroups. As such, common practice is to only assess calibration with respect to a few subgroups. Recent developments in goodness-of-fit testing offer potential solutions but are not designed for settings with weak signals or small, poorly calibrated subgroups. A new testing procedure is introduced based on the following insight: if observations can be reordered by their expected residuals, there should be a change in the association between the predicted and observed residuals along this sequence if a poorly calibrated subgroup exists. This reframes the problem of calibration testing into one of changepoint detection, for which powerful methods already exist. A sample-splitting procedure is first introduced where a portion of the data is used to train candidate models for predicting the residual, and the remaining data are used to perform an adaptive score-based cumulative sum (CUSUM) test. This test is then extended to incorporate cross-validation while maintaining Type I error control. Compared to existing methods, the proposed procedure consistently achieved higher power in simulation studies and real-world data analyses.

E0913: **Score identity distillation: Exponentially fast distillation of pretrained diffusion models for one-step generation**
Presenter: **Mingyuan Zhou**, University of Texas at Austin, United States

Score identity distillation (SiD), an innovative data-free method, is introduced to distil the generative capabilities of pre-trained diffusion models into a single-step generator. SiD facilitates an exponentially fast reduction in Frechet inception distance (FID) during distillation and approaches or exceeds the FID performance of the original teacher diffusion models. By reformulating forward diffusion processes as semi-implicit distributions, three score-related identities are leveraged to create an innovative loss mechanism. This mechanism achieves rapid FID reduction by training the generator using its own synthesized images, eliminating the need for real data or reverse-diffusion-based generation, all accomplished within a significantly shortened generation time. Upon evaluation across four benchmark datasets, the SiD algorithm demonstrates high iteration efficiency during distillation and surpasses competing distillation approaches, whether they are one-step or few-step, data-free, or dependent on training data in terms of generation quality. This achievement not only redefines the benchmarks for efficiency and effectiveness in diffusion distillation but also in the broader field of diffusion-based generation.

EO204 Room 109 THEORY-DRIVEN MACHINE LEARNING METHODS Chair: Ben Dai

E0456: **Model privacy: A unified framework to understand model stealing attack and defense**
Presenter: **Ganghua Wang**, University of Minnesota, United States
Co-authors: Jie Ding, Yuhong Yang

The use of machine learning (ML) has become increasingly prevalent in various domains, thereby highlighting the importance of understanding and ensuring its safety. One pressing concern is the vulnerability of ML applications to model-stealing attacks, where adversaries attempt to recover a learned model from limited query-response interactions, such as through cloud-based services or on-chip artificial intelligence (AI) interfaces. Existing literature has proposed various attack and defence strategies; however, they often lack a theoretical foundation and a standardized evaluation criterion for their efficacy. In response, a framework called "Model Privacy" is presented, providing a foundation for comprehensively analyzing model stealing attacks and defences. In particular, a rigorous formulation for the threat model and objectives are established, approaches that quantify the goodness of attack and defense strategies are proposed, and fundamental tradeoffs regarding the utility and privacy of ML models are analyzed. The developed theory offers valuable insights for enhancing the security of ML models, illustrated by various regression learning scenarios, including those based on k-nearest neighbors, polynomials, reproducing kernels, and neural networks. The importance of breaking data

independence is also highlighted in devising powerful defenses. Moreover, this framework exhibits intimate connections to other critical AI areas, such as teacher-student learning.

E0514: **Word-level maximum mean discrepancy regularization for word embedding**
Presenter: **Youqian Gao**, The Chinese University of Hong Kong, Hong Kong
Co-authors: Ben Dai
The technique of word embedding is widely used in natural language processing (NLP) to represent words as numerical vectors in textual datasets. However, the estimation of word embedding may suffer from severe overfitting due to the huge variety of words. To address the issue, a novel regularization framework is proposed that recognizes and accounts for the "word-level distribution discrepancy", a common phenomenon in a range of NLP tasks where word distributions are noticeably disparate under different labels. The proposed regularization, referred to as word-level MMD (wMMD), is a variant of maximum mean discrepancy (MMD) that serves a specific purpose: to enhance/preserve the distribution discrepancies within word embedding numerical vectors and thus prevent overfitting. The theoretical analysis illustrates that wMMD can effectively operate as a dimension-reduction technique of word embedding, thereby significantly improving the robustness and generalization of NLP models. The numerical effectiveness of wMMD is demonstrated in various simulated examples, such as Chile Earthquake T1 and BBC News datasets with state-of-the-art NLP deep learning architectures.

E0615: **Inferring causal direction between two traits using R-squared with application to transcriptome-wide association studies**
Presenter: **Haoran Xue**, City University of Hong Kong, Hong Kong
Co-authors: Huiling Liao, Wei Pan
In the framework of Mendelian randomization, two single SNP-trait Pearsons correlation-based methods have been developed to infer the causal direction between an exposure (e.g. a gene) and an outcome (e.g. a trait), including the widely used MR Steiger's method and its recent extension called causal direction-ratio (CD-Ratio). Steiger's method uses a single SNP as an instrumental variable (IV) for inference, while CD-Ratio combines the results from each of multiple SNPs. An approach is proposed based on R-squared, the coefficient of determination, to simultaneously combine information from multiple SNPs to infer the presence and direction of a causal relationship between an exposure and an outcome. The proposed method can be regarded as a generalization of Steiger's method from using a single SNP to multiple SNPs as IVs. It is especially useful in transcriptome-wide association studies (TWAS) with typically small sample sizes for gene expression data, providing a more flexible and powerful approach to inferring causal directions. It can be applied to GWAS summary data with a reference panel. Its potential robustness is also discussed to invalid IVs. The performance of TWAS, Steiger's method, CD-Ratio, and the new R-squared-based method is compared in simulations and real data analysis to demonstrate some advantages of the proposed method.

E0949: **Non-asymptotic bounds for adversarial excess risk under misspecified models**
Presenter: **Changyu Liu**, Department of Statistics, The Chinese University of Hong Kong, Hong Kong
The aim is to propose a general approach to evaluating the performance of robust estimators based on adversarial losses under misspecified models. It is first shown that adversarial risk is equivalent to the risk induced by a distributional adversarial attack under certain smoothness conditions. This ensures that the adversarial training procedure is well-defined. To evaluate the generalization performance of the adversarial estimator, the adversarial excess risk is studied. The proposed analysis method includes investigations on both generalization error and approximation error. Non-asymptotic upper bounds are then established for the adversarial excess risk associated with Lipschitz loss functions. In addition, the general results are applied to adversarial training for classification and regression problems. For the quadratic loss in nonparametric regression, it is shown that the adversarial excess risk bound can be improved over those for a general loss.

EO170 Room 110 RECENT DEVELOPMENTS IN DESIGN AND ANALYSIS OF EXPERIMENTS Chair: Chenlu Shi

E0159: **Subsampling and rare events data beyond binary responses**
Presenter: **HaiYing Wang**, University of Connecticut, United States
Rare event data occur when certain events occur with very small probabilities. Subsampling effectively reduces the computational cost of analyzing rare event data without losing significant estimation efficiency. Existing investigations on subsampling with rare event data focus on binary response models. Rare event data beyond binary responses are investigated. There will be no statistical efficiency loss if sufficient data points for the non-rare observations are sampled. In the scenario of estimation efficiency loss due to downsampling, an optimal sampling design is developed to minimize the information loss.

E0400: **Construction of orthogonal-MaxPro Latin hypercube designs**
Presenter: **Qian Xiao**, Shanghai Jiaotong University, China
Co-authors: Yaping Wang, Sixu Liu
Orthogonal Latin hypercube designs (LHDs) and maximum projection (MaxPro) LHDs are widely used in computer experiments. They are efficient for estimating the trend part and the Gaussian process part of the universal Kriging (i.e. the Gaussian process) model, respectively, especially when only some of the factors are active. Yet, the orthogonality and the MaxPro criteria often do not agree with each other. A new class of optimal designs, called orthogonal-MaxPro LHDs, is proposed to optimize a well-defined multi-objective criterion combining correlation and MaxPro metrics. An efficient parallel algorithm via level permutations and expansions is developed, and its efficiency is guaranteed by theories. Numerical results are presented to show that the construction is fast and the obtained designs are attractive, especially for large computer experiments.

E0535: **Design inspired Thompson sampling**
Presenter: **Wei Zheng**, University of Tennessee, United States
Thompson sampling is a popular algorithm for multi-armed bandit problems, but its Bayesian posterior update can be computationally expensive for complex reward distributions. Recently, prior discretization has been proposed to address this issue. A new prior discretization method is proposed that guarantees the same regret rate without requiring the unreasonable assumption that the true value of the parameter is one of the discrete points. Additionally, a modified posterior update approach is introduced that further improves the performance of discrete prior Thompson sampling. It is proven that the accumulated regret has $O(\log(T))$ convergence rate with high probability. In addition, numerical experiments are conducted to validate the theoretical analysis and demonstrate that the proposed algorithm outperforms both the standard discrete prior method and the Laplace approximation approach for the continuous prior.

E0689: **Kernel discrepancy-based rerandomization for controlled experiments**
Presenter: **Yiou Li**, DePaul University, United States
Co-authors: Lulu Kang
Controlled experiments have been widely used in various disciplines for causal inference. Rerandomization has been proposed and advocated to improve the covariates balance. One key component in rerandomization is the balance criterion. The kernel discrepancy is introduced between the empirical distributions of the covariates in different treatment groups, and the upper bound of the variance of the difference-in-mean estimator of the treatment effects is shown to be regulated. Accordingly, using kernel discrepancy is proposed as a balance criterion. Using the linear kernel function, the distribution of the kernel discrepancy is obtained for finite samples, which provides the critical value for an acceptable rerandomization. For

more complicated kernel functions, empirical distributions are proposed using the kernel discrepancy to obtain the critical value. The discrepancy-based criterion is model-free and thus makes the estimation of the treatment effect(s) robust to the model assumptions. More importantly, the proposed design is applicable to both continuous and categorical response measurements. Through simulation study and a real example, it is shown that the proposed design approach achieves accurate estimation even if the model assumption is not correct.

EO054 Room 111 ADVANCES IN STATISTICAL METHODS FOR BIOLOGICAL DATA AND HEALTH INFORMATICS Chair: Peijun Sang

E0270: **Functional principal component analysis under informative sampling**
Presenter: **Peijun Sang**, University of Waterloo, Canada
Co-authors: Dehan Kong, Shu Yang
Functional principal component analysis has been shown to be invaluable for revealing variation modes of longitudinal outcomes, which serve as important building blocks for forecasting and model building. Decades of research have advanced methods for functional principal component analysis, often assuming independence between the observation times and longitudinal outcomes. Yet such assumptions are fragile in real-world settings where observation times may be driven by outcome-related reasons. Rather than ignoring the informative observation time process, the observational times are explicitly modeled by a counting process dependent on time-varying prognostic factors. Identification of the mean, covariance function, and functional principal components ensue via inverse intensity weighting. The use of weighted penalized splines is proposed for estimation, and consistency and convergence rates are established for the weighted estimators. Simulation studies demonstrate that the proposed estimators are substantially more accurate than the existing ones in the presence of a correlation between the observation time process and the longitudinal outcome process. The finite-sample performance of the proposed method is further examined using the acute infection and early disease research program study.

E0274: **Joint model for survival and multivariate sparse functional data for Alzheimer's disease**
Presenter: **Luo Xiao**, North Carolina State University, United States
Studies of Alzheimer's disease (AD) often collect multiple longitudinal clinical outcomes, which are correlated and predictive of AD progression. It is of great scientific interest to investigate the association between the outcomes and time to AD onset. The multiple longitudinal outcomes are modeled as multivariate sparse functional data, and a functional joint model linking multivariate functional data to event time data is proposed. In particular, a multivariate functional mixed model is proposed to identify the shared progression pattern and outcome-specific progression patterns of the outcomes, which enables more interpretable modelling of associations between outcomes and AD onset. The proposed method is applied to the Alzheimer's Disease Neuroimaging Initiative study (ADNI), and the functional joint model sheds new light on the inference of five longitudinal outcomes and their associations with AD onset. Simulation studies also confirm the validity of the proposed model. In addition, the model is extended and applied to multiple cohorts of AD studies.

E0281: **Analysis of microbiome differential abundance by pooling Tobit models**
Presenter: **Gen Li**, University of Michigan Ann Arbor, United States
Microbiome differential abundance analysis (DAA) is pivotal in identifying microbial features associated with various disease conditions. However, the inherent complexities of metagenomics sequencing data, including compositionality and sparsity, challenge the accuracy and effectiveness of current methodologies, resulting in inflated false discovery rates and diminished statistical power. Addressing this gap, a novel approach called ADAPT (analysis of differential abundance by pooling Tobit models) is presented. Explicit assumptions are first established for DAA to elucidate the relationship between relative and absolute abundances. Leveraging this insight, ADAPT strategically identifies a subset of reference taxa based on the ordered list of fold changes in relative abundances. Subsequently, it conducts hypothesis testing on the fold change of count ratios between each taxon and the reference set. ADAPT employs Tobit models to effectively estimate fold changes, treating zero values as partially observed values left censored at the detection limit. Through extensive simulation studies and real data analysis, ADAPT is demonstrated to surpass existing methods by better controlling false discovery rates and exhibiting higher power.

E0640: **Semiparametric joint modeling for biomarker trajectory before disease onset**
Presenter: **Yifei Sun**, Columbia University, United States
Understanding the dynamics of biomarkers prior to disease onset is a critical topic in biomedical research. A semiparametric joint model is proposed to analyze the temporal evolution of biomarkers, allowing for a flexible biomarker trajectory shape that depends on two-time scales: a natural time scale such as age and the time relative to disease onset. An additional complication arises because the natural time scale often differs from the time of the study, leading to analytical challenges such as left-truncation bias and irregular measurements. To address these issues, a profile kernel estimating equation approach is introduced to estimate regression coefficients and unspecified baseline mean trajectory functions. The large-sample properties are established of the proposed estimators and conduct simulation studies to evaluate their finite sample performances. The method is applied to investigate the brain biomarker trajectory before the onset of preclinical Alzheimer's disease.

EO311 Room 212 ADVANCES IN BIG DATA ANALYSIS AND DIMENSION ASYMPTOTICS (VIRTUAL) Chair: Fei Tan

E0759: **The A-optimal subsampling approach to the analysis of count data of massive size**
Presenter: **Fei Tan**, Indiana University-Purdue University Indianapolis, United States
Co-authors: Xiaofeng Zhao, Hanxiang Peng
Uniform and statistical leverage-scores-based (nonuniform) distributions are frequently used in the analysis of massive data. Both distributions, however, are not effective in the extraction of important information in data. The A-optimal subsampling estimators of parameters are constructed in generalized linear models (GLM) to approximate the full-data estimators and derive the A-optimal distributions based on the criterion of minimizing the sum of the component variances of the subsampling estimators. As the distributions have the same running time as the full-data estimator, the scoring algorithm introduced in a recent study is generalized in a big data linear model to GLM using the iterative weighted least squares. The purpose is to present a comprehensive numerical evaluation of the approach using the simulated and real data by comparing its performance with the uniform and the leverage-scores subsampling. The results exhibited that the approach substantially outperformed the uniform and the leverage-scores subsampling, and the algorithm significantly reduced the computing time required for implementing the full-data estimator.

E0655: **Predictive model degrees of freedom in linear regression**
Presenter: **Yunzhang Zhu**, Ohio State University, United States
Overparametrized interpolating models have drawn increasing attention from machine learning. Some recent studies suggest that regularized interpolating models can generalize well. This phenomenon seemingly contradicts the conventional wisdom that interpolation tends to overfit the data and performs poorly on test data. Further, it appears to defy the bias-variance trade-off. As one of the shortcomings of the existing theory, the classical notion of model degrees of freedom fails to explain the intrinsic difference among the interpolating models since it focuses on the estimation of in-sample prediction error. This motivates an alternative measure of model complexity, which can differentiate those interpolating models and take different test points into account. In particular, a measure with a proper adjustment is proposed based on the squared covariance between the predictions and observations. The analysis with the least squares method reveals some interesting properties of the measure, which

can reconcile the "double descent" phenomenon with the classical theory. This opens doors to an extended definition of model degrees of freedom in modern predictive settings.

E0763: **The A-optimal subsampling for big data penalized spline single index models**
Presenter: **Hanxiang Peng**, IUPUI, United States
Co-authors: Fang Li, Haixia Smithson
Motivated by the computational burden in fitting single index models caused by high parameter dimensionality and possibly compounded by data of massive size, the A-optimal subsampling estimators are constructed to approximate the full data estimators. The A-optimal sampling distribution is derived by minimizing the sum of the component variances of the subsampling estimator. For an arbitrary distribution (π_i) on the n data points with its minimum $\pi_{\min}$ satisfying $n\pi_{\min} \geq l_0 > 0$ for some constant l_0, asymptotic normality of the subsampling estimator is proven for either fixed or growing sum $p+d$ of the number p of the index parameters and the number d of basis functions as the subsample size r tends to infinity such that $p+d$ grows slowly at the rate $p+d = o(r^{1/5})$ under suitable conditions. An unweighted subsampling estimator is also constructed; its asymptotic normality is proven for growing dimension without the foregoing assumption on (π_i) and establishes its higher efficiency than the weighted estimator. The analytic formulas of the first-order bias are provided for both estimators and explore how the estimators and their biases are affected by the penalty λ, $p+d$, (π_i) and r. A fast algorithm having running time $O(r^2(p+d))$ is constructed with r far less than n, and the numerical behavior of the Subsampling approach is studied using both simulated and real data.

E0777: **Subsample size determination with different approaches**
Presenter: **Sheng Zhang**, Indiana University Purdue University at Indanapolis, United States
Motivated by subsampling in the analysis of big data and by data-splitting in machine learning, sample size determination for multidimensional parameters is studied with the traditional normal approximation approach. A novel approach is also proposed to the construction of confidence intervals based on concentration inequalities with the missing factors, and by applying reversely, the approach can be used to determine the subsample size for big data analysis. Improved concentration inequalities are derived by providing the missing factors, and the results are applied to estimate the tail probability of certain random sums. The formula for confidence interval is provided, and the simulation results are reported.

EO159 Room 204 RECENT ADVANCEMENTS IN BAYESIAN MODELING Chair: Xiaojing Wang

E1104: **Shape-based clustering method for dynamic latent abilities in item response theory models**
Presenter: **Xiaojing Wang**, University of Connecticut, United States
Co-authors: Jingyu Sun
With recent advances in computerized testing, the collection of longitudinal data become easier and more prominent in educational testing. However, classic item response theory (IRT) models show limitations in dealing with individually varying and irregularly spaced longitudinal dichotomous responses often collected from the computerized testing or learning platform. Following prior pioneering work, there is limited literature focusing on modelling longitudinal dichotomous response data to learn the dynamic changes of latent abilities, not to mention finding distinct patterns for the trajectories of latent abilities. The aim is to introduce a semi-parametric method using B-splines to estimate and cluster the trajectory of latent abilities simultaneously. Moreover, the semi-parametric method can consider the monotone shape constraints on the trajectory for growth. To demonstrate the usage of the method, the proposed model is applied to a real dataset collected from a personalized literacy learning platform called EdSphere. As a result, it has discovered some interesting group patterns of individuals based on their trajectory patterns of reading abilities with a monotone increasing assumption on the ability. The identified group patterns for learning trajectories may help educators facilitate tailored education.

E1107: **Parallel computing methods for Bayesian analysis of big data sets**
Presenter: **Erin Conlon**, University of Massachusetts Amherst, United States
Co-authors: Zheng David Wei
Recently, new parallel Bayesian Markov Chain Monte Carlo (MCMC) methods have been developed for massive data sets that are too large for traditional statistical analysis. These methods partition big data sets by observations into subsets. The purpose is to discuss the alternative parallel Bayesian MCMC computing algorithm that partitions big data sets by groups rather than observations. This two-stage approach analyzes groups independently in parallel in stage one; the posteriors from stage one are used as proposal distributions in stage two, which estimates the complete data model. The method is illustrated with both three-level and four-level models, and improvements are shown in computation time as well as MCMC efficiency versus the complete data evaluation.

E1113: **A uniform shrinkage prior in spatiotemporal Poisson models for count data**
Presenter: **Yisheng Li**, The University of Texas MD Anderson Cancer Center, United States
Bayesian inference in a Poisson generalized linear mixed model for spatiotemporal data is considered. Normal random effects are used to model the within-area correlation over time, and spatial effects represented by a proper conditional autoregressive (CAR) model 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 is proper, and the resulting posterior is 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 is 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. 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. We report the top ten estimated leptospirosis morbidity rates (per 100,000 population) across the provinces and quarters.

E1101: **Does the adjustment cost of future resource expansion affect labor cost stickiness?**
Presenter: **Wuqing Wu**, Renmin University of China, China
Co-authors: Guoliang Chen
Cost stickiness is an important topic in management accounting, and the cost of adjusting resources is one of the main reasons for cost stickiness. However, existing studies have mainly explored the adjustment cost related to the current resource reduction and paid less attention to the adjustment cost related to the future resource expansion. Based on the background of China's subway construction, it is explored whether the opening of subway stations affects firms' labor cost stickiness. It is found that the opening of subway stations can significantly reduce the cost of hiring employees, thus reducing firms' labor cost stickiness. The cross-sectional test also shows that the effect of subway station opening on labor cost stickiness is more pronounced for firms with high labor demand and low labor supply.

EO029 Room 207 STATISTICAL INFERENCE FOR HIGH-DIMENSIONAL DATA Chair: Alfonso Landeros

E0265: **Unconditional quantile regression for streaming data sets**
Presenter: **Rong Jiang**, Shanghai Polytechnic University, China
The unconditional quantile regression (UQR) method, initially introduced by another study, has gained significant traction as a popular approach for modeling and analyzing data. However, much like conditional quantile regression (CQR), UQR encounters computational challenges when it comes to obtaining parameter estimates for streaming data sets. This is attributed to the involvement of unknown parameters in the logistic regression loss function utilized in UQR, which presents obstacles in both computational execution and theoretical development. To address this, a novel approach is presented involving smoothing logistic regression estimation. Subsequently, a renewable estimator is proposed tailored for UQR with streaming data, relying exclusively on current data and summary statistics derived from historical data. Theoretically, the proposed estimators exhibit equivalent asymptotic properties to the standard version computed directly on the entire dataset without any additional constraints. Both simulations and real data analysis are conducted to illustrate the finite sample performance of the proposed methods.

E0282: **Detecting hub variables in large Gaussian graphical models**
Presenter: **Jose Sanchez Gomez**, UC-Riverside, United States
In modern scientific applications, identifying small sets of variables in a dataset with a strong influence over the rest is often vital. For example, when studying the gene expression levels of cancer patients, estimating the most influential genes can be a first step towards understanding underlying gene dynamics and proposing new treatments. A popular approach for representing variable influence is through a Gaussian graphical model (GGM), where each variable corresponds to a node, and a link between two nodes represents relationships among pairs of variables. In a GGM, influential variables correspond to nodes with a high degree of connectivity, also known as hub variables. A new method is presented for estimating hub variables in GGMs. To this end, a connection is established between the presence of hubs in a GGM and the concentration of principal component vectors on the hub variables. Probabilistic guarantees of convergence for the method are provided, even in high-dimensional data where the number of variables can be arbitrarily large. An application of this new method is also discussed in a prostate cancer gene-expression dataset, through which several hub genes are detected with close connections to tumor development.

E1022: **Simultaneous quantile regression models: Homogeneity, sparsity, and efficiency**
Presenter: **Zhen Zeng**, Nanjing University of Finance and Economics, China
Quantile regression analyzes the impact of predictors on the conditional distribution of the response by focusing on a collection of conditional quantiles rather than a single conditional mean. The main goal is to pursue homogeneity in simultaneous quantile regression (PHISQ) to get more interpretable and efficient quantile regression estimation. The new method reveals not only the predictors that are associated with the outcome but also the true sources of homogeneity, the specific predictors that only have homogeneous effects on the response at the quantiles of interest, and the true sources of heterogeneity, the predictors that have heterogeneous effects across quantile levels of interest. Therefore, the new method may eventually result in a model that is considerably more parsimonious and interpretable. In addition, the new method can pool/borrow information across different quantiles to estimate the homogeneous regression parameters and thus improve the estimation efficiency, especially at the tails. It is demonstrated that the penalized PHISQ method exhibits desirable properties under mild regularity conditions. Simulation results and a real data application further validate the effectiveness of PHISQ.

E1036: **The proximal distance principle: Algorithms and applications**
Presenter: **Alfonso Landeros**, University of California, Riverside, United States
Statistical methods often involve solving optimization problems. The addition of constraints, either to enforce a hard requirement in estimation or to regularize solutions, complicates matters. Fortunately, the rich theory of convex optimization provides ample tools for devising novel methods, especially when there is tension between theory and computational demands in a high-dimensional setting. A distance-to-set penalty strategy is discussed as a general approach to solving constrained estimation problems. Special emphasis is given to sparsity constraints for variable selection, which compromise between exhaustive combinatorial searches and shrinkage penalties. Examples drawn from life science applications vividly illustrate the ease of incorporating structure into estimators within the proximal distance framework.

EO225 Room 209 ADVANCED STATISTICAL METHODS FOR ANALYZING COMPLEX DATA Chair: Fengqing Zhang

E0183: **A Bayesian regression model with misreported response**
Presenter: **Yuan Wang**, Washington State University, United States
The main objective is to identify the risk factors associated with adolescent marijuana use in Washington State, utilizing data from the 2021 Healthy Youth Survey (HYS). While the survey guarantees anonymity, the possibility of over- or under-reporting exists for various reasons, such as fear of being exposed, social stigma, peer pressure, and so on. The interest is in identifying factors that are associated with true marijuana use as well as the occurrence of misreport. A full Bayesian framework is developed with a two-level latent linear regression model. The top level is for the true Marijuana use response, and the second level is for the occurrence of misreporting. A partially collapsed Gibbs sampling algorithm is proposed to sample the regression coefficients. Intensive Monte Carlo simulation is used to demonstrate the performance of the proposed methods. The analysis of HYS data discovers multiple factors for identifying at-risk adolescents and informing future prevention efforts.

E0236: **Bayesian fixed-domain asymptotics for covariance parameters in spatial Gaussian process models**
Presenter: **Cheng Li**, National University of Singapore, Singapore
Co-authors: Saifei Sun, Yichen Zhu
Gaussian process models typically contain finite dimensional parameters in the covariance function that need to be estimated from the data. The Bayesian fixed-domain asymptotics is studied for the covariance parameters in spatial Gaussian process regression models with an isotropic Matern covariance function, which has many applications in spatial statistics. For the model without nugget, it is shown that when the dimension of the domain is less than or equal to three, the microergodic parameter and the range parameter are asymptotically independent in the posterior. While the posterior of the microergodic parameter is asymptotically close in total variation distance to a normal distribution with shrinking variance, the posterior distribution of the range parameter does not converge to any point mass distribution in general. New evidence for the model with nugget is derived from lower bound and consistent higher-order quadratic variation estimators, which lead to explicit posterior contraction rates for both the micro ergodic and nugget parameters. The asymptotic efficiency and convergence rates of Bayesian kriging prediction are further studied. All the new theoretical results are verified in numerical experiments and real data analysis.

E0394: **Link prediction problems in functional brain networks**
Presenter: **Panpan Zhang**, Vanderbilt University Medical Center, United States
Functional magnetic resonance imaging (fMRI) has been widely used to discover the neural underpinnings of cognition decline caused by neurological disorders. Graph-theory-based methods are prevalent for analyzing brain networks constructed from fMRI data. Due to data usability and data noise, the constructed brain networks may contain false positive and/or false negative links. Using graph-based measures extracted from such brain networks as predictors in downstream analyses may cause inference bias. A Bayesian approach is introduced to functional brain net-

work analysis with the presence of false positive and false negative links. The proposed method is applied to investigate the association between functional connectivity and cognitive changes in Alzheimer's disease.

E0211: **Graphical approaches and reverse graphical approaches in clinical studies with multiple endpoints**
Presenter: **Jiangtao Gou**, Villanova University, United States
Multiple test problems in clinical trials are common and diverse. The graphical approach has been proposed as a general framework for clinical trial designs with multiple hypotheses, where decisions are based on the observed marginal p values. It begins with a graph comprising all hypotheses as vertices, then progressively removes vertices when their corresponding hypotheses are rejected. In contrast, the reverse graphical approach starts with singleton graphs and incrementally adds vertices until a hypothesis set is rejected. Some properties of the reverse graphical approach are further evaluated and discussed.

EO155 Room 210 STATISTICAL METHODS FOR BIOLOGICAL DATA ANALYSIS AND BIOINFORMATICS Chair: Zeny Feng

E0356: **Z-residual diagnostics for Bayesian hurdle models**
Presenter: **Longhai Li**, University of Saskatchewan, Canada
Residual diagnostics is important for frequentist normal regression modelling. A fitted model's overall goodness-of-fit (GOF) is inspected by checking the residuals' normality with QQ plots and statistical tests. However, residual diagnostic tools are not available for Bayesian models. The method of Z-residual is proposed to check the adequacy of Bayesian models. The Z-residual is transformed from the cross-validatory randomized predictive p-values (RPP) with the normal quantile function. It is shown that the RPP has a uniform distribution on $(0,1)$ when the likelihood and the prior are correctly specified. Due to the uniformity of the RPPs, Z-residuals are normally distributed under the true model. Therefore, the Z-residual is used to conduct residual diagnostics for Bayesian models as for normal regression. Applying Z-residual diagnostics to Bayesian hurdle models, the graphical and numerical diagnostics are demonstrated based on Z-residuals, which can effectively identify the misspecification in the distributional family for the response variable and the misspecified covariate functional form. The sizes and powers of statistical tests are also investigated based on the Z-residual with large-scale simulation studies.

E0491: **Statistical methods applied in microbial metagenomics**
Presenter: **Yunliang Li**, University of Saskatchewan, Canada
The microbial community plays a pivotal role in influencing human health, environmental sustainability, and ecosystem resilience. Advances in next-generation sequencing techniques have enabled the capture of vast metagenomic information on uncultured microbiota. Leveraging metagenomic sequencing data, researchers can delve into the complex composition of microbial communities and explore the functions of community members in relation to host and environment. However, microbiome data exhibits distinct characteristics, including high dimensionality, sparsity with numerous zero counts, and compositional nature, which present substantial challenges for the application of traditional statistical methods. To tackle these issues, novel statistical approaches are under development to facilitate more effective data analysis and the generation of reproducible and stable conclusions. Statistical methods tailored to various analytical objectives are introduced within microbiome data analysis, encompassing microbial diversity analysis, differentially abundant feature analysis, and microbial interaction analysis. Insights are gained into the strengths and limitations of these methods, empowering them to effectively employ these methods in microbiome data analysis.

E0493: **Randomized quantile residuals for diagnosing zero-inflated models with applications to microbiome count data**
Presenter: **Cindy Feng**, Dalhousie University, Canada
Zero-inflated generalized linear models, particularly zero-inflated negative binomial models, are commonly used for differential abundance analysis of microbiome and other sequencing count data. When estimating the false discovery rate (FDR), it's assumed that p-values follow a uniform distribution under the null hypothesis. Ensuring that the chosen model adequately fits the count data is crucial to control FDR and avoid excess false discoveries. Model checking is, therefore, essential in this analysis. The randomized quantile residual (RQR) method has been shown to effectively diagnose the count regression models. However, its performance in diagnosing zero-inflated generalized linear mixed models (GLMMs) for sequencing count data has not been extensively studied. Large-scale simulation studies are conducted to assess the performance of RQRs for zero-inflated GLMMs. The simulations demonstrate that the type I error rates of the goodness-of-fit tests with RQRs closely match the nominal level. Scatter plots and QQ plots of RQRs are valuable for distinguishing between good and bad models. RQRs are also applied to diagnose six GLMMs in a real microbiome dataset, finding that RQR is an excellent tool for diagnosing GLMMs for zero-inflated count data, particularly in microbiome studies.

E0741: **Optimization of the regularized Dirichlet multinomial regression and its application in compositional data analysis**
Presenter: **Zeny Feng**, University of Guelph, Canada
Co-authors: Alysha Cooper, Ayesha Ali, Lorna Deeth, Tim Arciszewski
Compositional data measured as taxonomical counts are prevalent in many biological fields, including ecology and microbiology. In ecology, samples of benthic macroinvertebrates taken from different aquatic sites were classified into taxonomic ranks based on phylum, class, order, family, genus, and species. At a given rank, the taxa counts of species conditional on the total counts can be modelled by the Dirichlet multinomial (DM) distribution, which can accommodate the multinomial over-dispersion. The model fitting in the presence of covariates can be challenging because the DM distribution falls outside the exponential family, and the number of parameters is as high as pxD, where p is the number of covariates and D is the number of taxa. With these challenges, a sparse group LASSO is proposed in the regularized DM regression. An MM-algorithm is formulated to optimize the penalized DM regression likelihood. The proposed method will be applied to identify the associations between water variables and the composition of benthic macroinvertebrates using the data collected from the Oil Sand Region in Alberta, Canada.

EO020 Room 307 FRONTIERS IN NONPARAMETRIC STATISTICS AND FUNCTIONAL DATA ANALYSIS Chair: Xinyi Li

E0473: **A sparse empirical Bayes approach to high-dimensional Gaussian process-based varying coefficient models**
Presenter: **Myungjin Kim**, Kyungpook National University, Korea, South
Co-authors: Gyuhyeong Goh
Despite the increasing importance of high-dimensional varying coefficient models, the study of their Bayesian versions is still in its infancy. The contribution to the literature is by developing a sparse empirical Bayes formulation that addresses the problem of high-dimensional model selection in the framework of Bayesian varying coefficient modeling under Gaussian process (GP) priors. To break the computational bottleneck of GP-based varying coefficient modeling, a low-cost computation strategy that incorporates linear algebra techniques and the Laplace approximation into the evaluation of the high-dimensional posterior model distribution is introduced. A numerical study is conducted to demonstrate the superiority of the proposed Bayesian method compared to an existing high-dimensional varying coefficient modeling approach.

E0520: **Spatially-varying coefficient models with structure identification**
Presenter: **Guannan Wang**, College of William & Mary, United States
Co-authors: Zhiling Gu, Xinyi Li, Lily Wang

A general framework of spatially varying coefficient models with structure identification is introduced, which involves automatic detection of the significance and types of effects (spatially varying or constant) of various factors on the response of interest. The proposed method can efficiently identify spatially varying coefficient components, enhancing computational efficiency and statistical power for downstream analysis. To provide a solid theoretical foundation for the proposed method, the consistency and asymptotic normality of the constant estimators are rigorously established. Moreover, extensive Monte Carlo simulations are conducted to examine the efficacy of the proposed method in identifying the true model structure and improving estimation and prediction accuracy. Additionally, the practical utility of the framework is illustrated through an analysis of particulate matter (PM), which provides valuable insights into the influence of environmental factors on observed PM.

E0561: **Regularizing BELIEF for smooth dependency**
Presenter: **Wan Zhang**, UNC Chapel Hill, United States
As the complexity of models and the volumes of data increase, interpretable methods for modeling complicated dependence are in great need. A recent framework of binary expansion linear effect (BELIEF) provides a "divide and conquer" approach to decompose any complex form of dependency into small linear regressions over data bits. Although BELIEF can be used to approximate any relationship, it faces an important challenge of high dimensionality. To overcome this obstacle, a novel definition of smoothness is proposed for binary interactions, and a regularization of BELIEF is created under smoothness interpretations. It has been proven that there is a one-on-one correspondence between each marginal binary interaction and the smoothness defined. Additionally, it is shown that in higher dimensions, the smoothness can be expressed as a product of marginal binary interactions. Based on these observations, it is proposed to model the smooth form of dependency with a generalized LASSO model with a larger penalty on less smooth terms. The numerical studies show that the smooth LASSO takes advantage of clear interpretability and effectiveness for nonlinear and high-dimensional data.

E0727: **Fast multilevel functional principal component analysis**
Presenter: **Erjia Cui**, University of Minnesota, United States
Fast multilevel functional principal component analysis (fast MFPCA) is introduced, which scales up to high dimensional functional data measured at multiple visits. The new approach is orders of magnitude faster than the original MFPCA and achieves comparable estimation accuracy. Methods are motivated by the National Health and Nutritional Examination Survey (NHANES), which contains minute-level physical activity information of more than 10,000 participants over multiple days and 1440 observations per day. While MFPCA takes more than five days to analyze these data, fast MFPCA takes less than five minutes. A theoretical study of the proposed method is also provided. The associated function mfpca.face() is available in the R package refund.

EO259 Room 313 MODERN DEVELOPMENTS IN SPACE-TIME MODELING Chair: Yawen Guan

E1086: **Distributed heterogeneity learning for generalized partially linear models with spatially varying coefficients**
Presenter: **Shan Yu**, University of Virginia, United States
Co-authors: Guannan Wang, Lily Wang
Spatial heterogeneity is of great importance in social, economic, and environmental science studies. The spatially varying coefficient model is a popular and effective spatial regression technique to address spatial heterogeneity. However, accounting for heterogeneity comes at the cost of reducing model parsimony. To balance flexibility and parsimony, the purpose is to develop a class of generalized partially linear spatially varying coefficient models which allow the inclusion of both constant and spatially varying effects of covariates. Another significant challenge in many applications comes from the enormous size of the spatial datasets collected from modern technologies. To tackle this challenge, a novel distributed heterogeneity learning (DHL) method is designed based on bivariate spline smoothing over a triangulation of the domain. The proposed DHL algorithm has a simple, scalable, and communication-efficient implementation scheme that can almost achieve linear speedup. In addition, rigorous theoretical support is provided for the DHL framework. It is proven that the DHL constant coefficient estimators are asymptotic normal and the DHL spline estimators reach the same convergence rate as the global spline estimators obtained using the entire dataset. The proposed DHL method is evaluated through extensive simulation studies and analyses of the U.S. loan application data.

E1089: **Theoretical and computational challenges for space-time models**
Presenter: **Hao Zhang**, Michigan State University, United States
Spatiotemporal data are usually huge in size, and the corresponding covariance matrix is of high dimension. Theoretically, it is shown that this covariance matrix is ill-conditioned if the number of spatial locations is huge and constrained in a bounded domain. In that case, it is forced to seek an approximation of the likelihood and the best linear unbiased prediction. Some approximation methods and a new attempt to apply deep learning to the prediction of spatiotemporal data are discussed.

E1092: **Spatial-temporal data integration to improve the assessment of population exposure and health risks to PM2.5**
Presenter: **Zhengyuan Zhu**, Iowa State University, United States
Monitoring and forecasting PM2.5 is important for countries where air pollution is a serious public health issue. Current PM2.5 forecasts are mostly based on observations from monitoring stations, which have high temporal frequency but sparse and uneven spatial distribution. Aerosol Optical Depth (AOD) data from satellites such as MODIS has better spatial coverage but low temporal frequency. The fusion of PM2.5 data from stations and AOD data from satellites to provide hourly high-resolution PM2.5 data is useful for forecasting and epidemiology studies. The aim is to propose a novel data fusion framework using spatial functional data analysis tools. Efficient algorithms are developed to estimate the non-stationary mean and covariance structure using Bivariate spline and PACE. Estimates from the AOD data are then used to improve the spatial prediction of PM2.5. The proposed method is applied to data in the Beijing area in China, and the proposed approach outperforms several existing data fusion methods.

E1100: **A spatiotemporal model for arctic sea ice trajectories**
Presenter: **Yawen Guan**, Colorado State University, United States
Arctic sea ice is a barrier between the warm air of the ocean and the colder atmosphere. Thus, it plays an important role in the climate. When narrow linear cracks (leads) form in the sea ice, the heat from the ocean is released into the atmosphere. Motion data from the RADARSAT geophysical processing system (RGPS) is analyzed to estimate where leads may form. RGPS provides a set of trajectories of points on an ice sheet to trace the displacement of sea ice; however, chunks of data are missing because data is collected by satellite. A spatiotemporal clustering and interpolation method is proposed that estimates where a lead may form and allows the inference of missing observations.

EO035 Room 405 ADVANCING STATISTICAL INFERENCE FOR COMPLEX DATA Chair: Meimei Liu

E0472: **Controlling false discovery rate in high-dimensional linear regression: The Gaussian mirror approach**
Presenter: **Xin Xing**, Virginia Tech University, United States
Identifying key variables that influence outcomes in linear regression models while also maintaining control over the false discovery rate (FDR) poses a significant challenge in statistical analysis. The aim is to introduce the Gaussian Mirror (GM) method, a novel approach for identifying crucial variables in linear regression models while controlling the false discovery rate (FDR). By creating a pair of mirror variables for each predictor using Gaussian perturbations, the GM method improves variable selection. It's compatible with standard regression techniques like ordinary least squares and Lasso and offers the flexibility of applying mirror variables pre or post-selection. The key advancement of the GM method lies in its capacity to generate test statistics that effectively maintain the FDR at a predefined level under realistic covariate dependence assumptions. The analysis showcases the GM method's superiority in managing FDR constraints, especially in situations with high covariate correlation and a dense array of influential variables. The GM method's innovative approach is covered, theoretical foundation, and empirical efficacy, offering attendees valuable insights into tackling complex statistical challenges.

E0683: **CLT for generalized linear spectral statistics of large-dimensional sample covariance matrices and its application**
Presenter: **Qing Yang**, University of Science and Technology of China, China
The generalized linear spectral statistics (GLSS) for high dimensional sample covariance matrices is introduced. The joint asymptotic normality of this statistic associated with various test functions is established when the dimension and the sample size are comparable under weak assumptions. As a natural application of the theory, a novel statistic is proposed based on GLSS to conduct hypothesis testing for spiked covariance matrices. Simulations indicate the accuracy and enhanced power of the proposed statistics and illustrate considerably better performance compared to the existing methods on various occasions.

E0651: **Network tight community detection**
Presenter: **Huimin Cheng**, Boston University, United States
Conventional community detection methods often categorize all nodes into clusters. However, the presumed community structure of interest may only be valid for a subset of nodes (named tight nodes), while the rest of the network may consist of noninformative scattered nodes. For example, a protein-protein network often contains proteins that do not belong to specific biological functional modules but are involved in more general processes or act as bridges between different functional modules. Forcing each of these proteins into a single cluster introduces unwanted biases and obscures the underlying biological implication. To address this issue, a tight community detection (TCD) method is proposed to identify tight communities excluding scattered nodes. The algorithm enjoys a strong theoretical guarantee of tight node identification accuracy and is scalable for large networks. The superiority of the proposed method is demonstrated by various synthetic and real experiments.

E0243: **Saddle point approximations for the tests of covariance matrices from decomposable Gaussian graphical models**
Presenter: **Yanyan Wu**, University of Hawaii at Manoa, United States
The aim is to consider a classical testing problem of equal covariance matrices from Gaussian models with conditional independences and is Markov with respect to a decomposable graph. Under these conditional independences, the dimension of the parameter space is then reduced significantly and can be represented by a directed acyclic graph (DAG). Such models are important for high-dimensional or sparse data in many fields, such as finance, marketing or genomics. Under the null hypothesis, the likelihood ratio test statistics (LRT), derived from the hyper Wishart distribution according to the DAG, follows a chi-square distribution and has the first-order accuracy. The proposed saddle point approximation method had a third-order of accuracy. Briefly, the derivation of the method involved (a) computation of the modified LRT, Bartlett Box M-statistic, which improved the accuracy of the test to the 2nd-order, (b) derivation of the cumulant generating function of the M-statistic, and (c) application of the Lugannani-Rice formula to the cumulative generating function. Simulation studies show that the proposed method has extremely accurate tail coverages even when the sample size is small.

EO212 Room 406 RECENT DEVELOPMENTS IN RELIABILITY ANALYSIS Chair: Man Ho Ling

E1048: **Imputations in one-shot devices data using machine learning algorithms**
Presenter: **Hon Yiu So**, Oakland University, United States
One-shot devices are products that will be destroyed immediately after use. Most of them have multiple components. Malfunctioning any one of the components will result in the device's failure. The one-shot devices are often tested under constant stress, accelerated life-test, or collect data from users or surveys to assess such devices. A link function relating to stress levels and lifetime is then applied to extrapolate the lifetimes of units from accelerated conditions to normal operating conditions. However, missing data often occurs during the data collection, and imputation is a popular way to analyze this data. The aim is to explore imputation performance using machine learning algorithms on one-shot datasets and compare them to traditional imputation methods.

E1039: **Reliability analysis of load-sharing systems using a flexible model with piecewise linear functions**
Presenter: **Debanjan Mitra**, Indian Institute of Management Udaipur, India
Co-authors: Ayon Ganguly
A flexible model to analyse data from load-sharing systems is constructed by approximating the cumulative hazard functions of component lifetimes using piecewise linear functions. The advantages of the resulting model are that it is data-driven and does not use strong assumptions on the underlying component lifetimes. Due to its flexible nature, the model is capable of providing a good fit to data obtained from load-sharing systems in general, thus resulting in an accurate estimation of important reliability characteristics. Estimates of reliability at a mission time, quantile function, mean time to failure and mean residual time for load-sharing systems are developed under the proposed model involving piecewise linear functions. Maximum likelihood estimation and construction of confidence intervals for the proposed model are discussed in detail. The performance of the proposed model is observed to be quite satisfactory through a detailed Monte Carlo simulation study. Analysis of load-sharing data pertaining to the lives of a two-motor load-sharing system is provided as an illustrative example. A comprehensive discussion is presented on a flexible model that can be used for load-sharing systems efficiently.

E0800: **On parameter estimation for generalized inverse Gaussian distribution**
Presenter: **Hideki Nagatsuka**, Chuo University, Japan
Co-authors: Shunsuke Kaneko
The generalized inverse Gaussian (GIG) distribution, provided by Halphen in 1941 and then developed by Barndorff-Nielsen, is a generalized distribution of the inverse Gaussian and gamma distributions. The GIG distribution has some desired properties. For example, any GIG distribution with a non-positive power parameter is the first hitting time to level 0 for a time-homogeneous diffusion process, which implies the potential use of this distribution as a lifetime distribution. The GIG distribution has infinite divisibility, which suggests that a Levy process can be constructed based on the GIG distribution. Some challenging problems are addressed in parameter estimation for the GIG distribution, and some applications of this distribution are introduced.

E0174: **Extended gamma process model for accelerated destructive degradation analysis**
Presenter: **Man Ho Ling**, The Education University of Hong Kong, Hong Kong

Accelerated degradation destructive testing (ADDT) has emerged as a valuable technique for analyzing highly reliable products. This approach has garnered significant attention in the field of reliability research. A common characteristic observed in many degradation studies is the presence of randomness in the initial degradation levels of tested units. Products with lower initial degradation levels tend to experience failure at an earlier stage. In light of this, an extended gamma process model specifically designed to analyze ADDT data with random initial degradation levels is presented. Approximations for the conditional mean-time-to-failure and the variance of failure time for products exhibiting higher degradation levels are also presented. These approximations provide valuable insights to practitioners in evaluating the impacts of different degradation levels on product reliability. Finally, an ADDT dataset of return springs illustrates the proposed model and methodologies for making informed decisions regarding quality management and product performance.

EO082 Room 408 STATISTICAL DEMOGRAPHY Chair: Zehang Li

E0675: **Using Bayesian methods and the singular value decomposition for fast, scalable demographic estimation and forecasting**
Presenter: **Junni Zhang**, National School of Development, Peking University, China

Statistical demographers estimate and forecast detailed age-sex profiles for quantities such as mortality, fertility, migration, health expenditure, or labor force participation. Increasingly, profiles are required for many combinations of geography, ethnicity, education status, or other stratifying variables. The dimensions of the resulting models can easily become large. However, demographic processes often have highly regular age-sex patterns. The number of parameters required to accurately represent age-sex patterns is typically much smaller than the number of age-sex categories. Statistical methods are developed that take advantage of these regularities. Singular value decompositions are applied to high-quality data from international databases, and the results are used to formulate informative prior distributions for age-sex profiles. These prior distributions are then embedded into a larger hierarchical model. Inference is done using a Laplace approximation, as implemented in R package TMB, and is extremely fast, even with thousands of parameters. It is illustrated using examples from analyses of mortality rates and labor force participation. The methods are implemented in an open-source R package.

E0421: **Obtaining population-based estimates for survey data using Bayesian hierarchical models with poststratification**
Presenter: **Emma Zang**, Yale University, United States

For large-scale surveys such as the National Health and Aging Trends Study (NHATS), investigators may wish to combine data from two (or more) cohorts in a single analysis to obtain larger sample sizes. Unfortunately, it is not possible to combine the 2011 and 2015 NHATS cohorts while retaining the sample weights. Bayesian hierarchical models are applied with poststratification as an alternative strategy for obtaining population-based estimates from NHATS. As proof of principle, prevalence estimates of frailty obtained from the Bayesian approach are compared with those obtained from the 2011 and 2015 cohorts using the NHATS sample weights. Once validated, the strategy is applied to combine the cohorts into a single analytical dataset without overlapping participants and generate Bayesian estimates of frailty for the combined cohort. The Bayesian models were validated within each cohort, producing nearly identical results to those using NHATS sample weights. The Bayesian estimates for the combined cohort were similar to cohort-specific estimates but were more precise. The ability to combine cohorts while generating population-based estimates will permit investigators to not only produce more precise estimates but also address questions that require larger sample sizes and, in turn, increase the value of NHATS to the scientific community.

E0716: **Dynamic models augmented by hierarchical data**
Presenter: **Yifan Jiang**, Pennsylvania State University, United States
Co-authors: Le Bao

Dynamic models have been successfully used in producing estimates of HIV epidemics at the national level due to their epidemiological nature and their ability to estimate prevalence, incidence, and mortality rates simultaneously. Recently, HIV interventions and policies have required more information at sub-national levels to support local planning, decision making and resource allocation. Unfortunately, many areas lack sufficient data for deriving stable and reliable results, and this is a critical technical barrier to more stratified estimates. One solution is to borrow information from other areas within the same country. However, directly assuming hierarchical structures within the HIV dynamic models is complicated and computationally time-consuming. A simple and innovative way is proposed to incorporate hierarchical information into dynamic systems by using auxiliary data. The proposed method efficiently uses information from multiple areas within each country without increasing the computational burden. As a result, the new model improves predictive ability and uncertainty assessment.

E0601: **Domain adaptive cause-of-death assignment using verbal autopsies under distribution shift**
Presenter: **Zehang Li**, University of California, Santa Cruz, United States

Understanding cause-specific mortality rates is crucial for monitoring population health and designing public health interventions. Worldwide, two-thirds of deaths do not have a cause assigned. Verbal autopsy (VA) is a well-established tool to collect information describing deaths outside of hospitals by conducting surveys to caregivers of a deceased person. It is routinely implemented in many low- and middle-income countries. Statistical algorithms to assign the cause of death using VAs are typically vulnerable to the distribution shift between the data used to train the model and the target population. This presents a major challenge for analyzing VAs, as labeled data are usually unavailable in the target population. The purpose is to discuss a latent class model framework for VA data that jointly models VAs collected over heterogeneous domains, such as multiple study sites, different time periods, or distinct subpopulations. A parsimonious representation of the joint distribution of the collected symptoms is introduced, and a computationally efficient algorithm is developed to generate posterior inference and out-of-domain cause-of-death assignment. The importance of accounting for data shift is also discussed in other related decision-making problems in VA studies.

EO191 Room 411 (Virtual sessions) RECENT ADVANCES IN BAYESIAN METHODS AND APPLICATIONS Chair: Dongu Han

E0899: **Semiparametric Bayesian two-stage meta-analysis between ambient temperature and daily confirmed cases of COVID-19**
Presenter: **Dongu Han**, Korea University, Korea, South
Co-authors: Kiljae Lee, Yeonseung Chung, Genya Kobayashi, Taeryon Choi

Environmental epidemiological studies often use a two-stage meta-analysis to explore the short-term link between environmental exposure and health outcomes across various locations. Initially, location-specific exposure-response relationships are estimated using a generalized linear model with splines and lag structures. Then, these location-specific associations are combined in a second stage, alongside location-level predictors, to identify factors contributing to location differences. While traditional methods use frequentist frameworks, our study introduces a Bayesian approach, improving both stages' models. The first stage employs a Bayesian distributed lag nonlinear model accommodating interactive nonlinearities and decaying lag effects. The second stage utilizes a nonparametric Bayesian kernel mixture multivariate meta-regression, explaining association parameters with meta-predictors and addressing violations of linearity, normality, and homoscedasticity assumptions. Markov Chain Monte Carlo and variational Bayes algorithms are developed for estimation. To validate, these methods are applied to study the short-term relationship between ambient temperature and COVID-19 incidence in the United States. Results show superior accuracy of the first-stage model with

small sample sizes and decaying lag effects, while the second-stage model captures distributional structures and nonlinear relationships effectively, outperforming conventional methods.

E0919: **Bayesian nonparametric model of marked Hawkes processes, with application to earthquake occurrences**
Presenter: **Hyotae Kim**, Duke University, United States
Co-authors: Athanasios Kottas
The aim is to propose a Bayesian nonparametric model for marked Hawkes processes (MHPs). The processes' conditional intensity function is decomposed into the ground process intensity and the mark density function. The primary concentration is modelling the process intensity, but it offers several choices for the mark density. The prior probability model for intensity has been carefully designed to provide model flexibility and tractable posterior inference using a novel mixture modeling method. This model was motivated by seismology applications, where magnitude is regarded as a mark associated with a time point for earthquake occurrence. Accordingly, the mixture model basis is a function of occurrence time and magnitude, with its functional form selected considering not only model flexibility but also earthquake data characteristics, for example, the fact that earthquakes of greater magnitude cause more subsequent shocks than earthquakes of smaller magnitude. The model is illustrated with an earthquake occurrence dataset and several synthetic examples.

E0926: **Predicting COVID-19 hospitalization using a mixture of Bayesian predictive syntheses**
Presenter: **Genya Kobayashi**, School of Commerce, Meiji University, Japan
Co-authors: Shonosuke Sugasawa, Yuki Kawakubo, Taeryon Choi, Dongu Han
A novel methodology is proposed, called the mixture of Bayesian predictive syntheses (MBPS), for multiple time series count data for the challenging task of predicting the numbers of COVID-19 inpatients and isolated cases in Japan and Korea at the subnational level. MBPS combines a set of predictive models and partitions the multiple time series into clusters based on their contribution to predicting the outcome. In this way, MBPS leverages the shared information within each cluster and is suitable for predicting COVID-19 inpatients since the data exhibit similar dynamics over multiple areas. Also, MBPS avoids using a multivariate count model, which is generally cumbersome to develop and implement. The Japanese and Korean data analyses demonstrate that the proposed MBPS methodology has improved predictive accuracy and uncertainty quantification.

E0972: **Robust Bayesian change point detection**
Presenter: **Daewon Yang**, Chungnam National University, Korea, South
Co-authors: Taeryon Choi
A new Bayesian approach for robust change point detection is proposed. The model utilizes the Dirichlet process hidden Markov model for multiple change point detection, which has the advantage of not requiring a predetermined number of change points. Furthermore, for robust estimation, a heavy tail error assumption is introduced based on the Student's t distribution. The model employs a mixture error assumption with Gaussian and Student's t errors to perform robust estimation, which aids in clearer change point detection. The proposed model is applied to an environmental epidemiology application based on a two-stage meta-analysis, analyzing the association between temperature and mortality in Japan from 1974 to 2015. In the first stage, the relationship between temperature and mortality is analyzed across each prefecture of Japan for four-year non-overlapping sub-periods using a distributed lag nonlinear model. In the second stage, the robust Bayesian change point detection methodology is applied to examine how the association between temperature and mortality rates in Japan has changed over time.

EC301 Room 202 HIGH-DIMENSIONAL STATISTICS Chair: Runmin Wang

E1037: **Vertex cover matroid variable selection**
Presenter: **Toby Kenney**, Dalhousie University, Canada
Co-authors: Hong Gu, Sarah Organ
Medium-to-high dimensional variable selection is plagued by the issue of correlation. When predictors are highly correlated, it is often impossible to tell which is "true". For prediction purposes, this is not a serious issue. However, when the interest is in controlling the false discovery rate, it becomes impossible to achieve good variable selection. In order to overcome this issue, a new paradigm is developed for variable selection, where instead of selecting a single set of variables, a list is provided of possible sets of true variables. This allows for making selections of the form "one of this pair of variables" when appropriate. Vertex cover matroids of graphs are found to be an effective structure for selecting variables in this paradigm. A challenge is defining the false positive and true positive rates when we are not selecting individual variables. By viewing variable selection in the right way, there is a very natural extension of the usual definitions to the current case, and while computation of these true positive and false positive rates is theoretically NP-hard, in practice, it is usually fairly easy to compute them. Through simulation studies, the new paradigm is shown to control the false discovery rate at the desired level while greatly increasing the true discovery rate compared with state-of-the-art methods. It is also shown that the selected variables have better predictive ability than the variables selected by other methods.

E1060: **Empirical priors inference in sparse high-dimensional generalized linear models**
Presenter: **Yiqi Tang**, Colby College, United States
Co-authors: Ryan Martin
High-dimensional linear models have been widely studied, but the developments in high-dimensional generalized linear models, or GLMs, have been slower. The focus is on the novel empirical or data-driven prior framework for inference on the coefficient vector and for variable selection in high-dimensional GLM. In this framework, data is used to appropriately center the prior distribution, leading to an empirical Bayes posterior distribution. The proposed posterior distribution is shown to concentrate around the true/sparse coefficient vector at the optimal rate, and conditions under which the posterior can achieve variable selection consistency are provided. Computation of the proposed empirical Bayes posterior is simple and efficient and is shown to perform well in simulations compared to existing methods in terms of estimation and variable selection.

E1033: **Poisson principal component analysis and its ensemble approaches for cross-study analyses**
Presenter: **Hong Gu**, Dalhousie University, Canada
Co-authors: Toby Kenney, Molly Hayes, Tianshu Huang
High-dimensional count data are ubiquitous. Parametric methods based on log-normal Poisson distribution assumptions for principal component analysis (PCA) are typically sensitive to outliers. The aim is to first present a semi-parametric PCA (Poisson PCA) method generally applicable to count data for dimension reduction and data exploration, then further present a family of Poisson PCA ensemble methods for common principal component or common factor analysis approaches to cross-study analyses on multiple data sets. Applications using microbiome count data to find microbial communities are used to demonstrate the methods.

E0563: **Testing for the equality of distributions in high dimension**
Presenter: **Xu Li**, Shanxi Normal University, China
A new homogeneous test is proposed for two high-dimensional random vectors. The test is built on a new measure, the so-called characteristic distance, which can completely characterize the homogeneity of two distributions. The newly proposed metric has some desirable properties; for example, it possesses a clear and intuitive probabilistic interpretation and could be used to address the high-dimensional distance inference.

Theoretically, the limiting behaviors under the conventional fixed dimension and high-dimensional distance inference are thoroughly investigated. Simulation studies and real data analysis are presented to illustrate the finite-sample performance of the proposed test statistic.

Wednesday 17.07.2024 13:30 - 15:10 Parallel Session C – EcoSta2024

EO043 Room 102 NEW ADVANCES IN TIME SERIES ANALYSIS AND ECONOMETRICS Chair: Kun Chen

E0915: **Sparse matrix estimation based on greedy algorithms and information criteria**
Presenter: **Hsueh-Han Huang**, Academia Sinica, Taiwan
The problem of estimating the covariance matrix of serially correlated vectors whose dimension is allowed to be much larger than the sample size is considered. It is proposed using the orthogonal greedy algorithm (OGA) and a high-dimensional Akaike information criterion (HDAIC) to estimate the matrix, showing that the proposed estimate is rate optimal under a sparsity condition more flexible than the existing literature. When the covariance matrix is bandable, a banding/tapering estimate whose parameters are chosen by a novel information criterion is introduced. The rate optimality of the latter estimate is also established.

E1004: **Predictive subgroup logistic regression: A new approach in customer churn modeling with unobserved heterogeneity**
Presenter: **Rui Huang**, Nanjing University, China
Co-authors: Kun Chen, Zhiwei Tong
Modeling customer churn has become increasingly vital in the competitive landscape of today's markets, as it helps businesses understand customer churn behavior and develop tailored marketing strategies. Traditional techniques, such as decision trees and logistic regression, often overlook the heterogeneity in the latent factors driving customer churn. The aim is to introduce a novel predictive subgroup logistic regression (PSLR) model designed to identify unobserved subgroup structures among existing customers, accurately classify new customers into these subgroups and subsequently generate churn predictions. A penalized likelihood function is derived and optimized for estimating this model, addressing the challenges associated with optimization through the development of an alternating direction method of multipliers (ADMM) algorithm, for which convergence is proven. Extensive simulation studies confirm the PSLR model's efficacy in inferential and predictive tasks. An empirical study of a telecommunication dataset demonstrates that the PSLR model identifies the presence of unobserved heterogeneity among customers, even after initial segmentation by decision trees. Moreover, compared to selected benchmark models, the PSLR model achieves more balanced performance in identifying both positives and negatives while also achieving better or at least comparable results in terms of various aggregate accuracy metrics.

E1054: **Asymptotic and bootstrap inference for change-points in time series**
Presenter: **Xinyi Tang**, The Hang Seng University of Hong Kong, Hong Kong
The asymptotic distribution of a change-point estimator is studied for piecewise stationary time series under various break sizes. In particular, the break sizes $\|d_n\| = O(1/n^{\alpha})$ is considered for $0 < \alpha < 1/2$, $\alpha = 1/2$ and $\alpha > 1/2$, which represent large, moderate and small break sizes respectively, where n is the sample size. It is shown that the asymptotic distributions in all three cases are different but are related to the maximizer of a two-sided drifted Brownian motion. Also, the distribution is pivotal for the cases $\alpha = 1/2$ and $\alpha > 1/2$, for which the analytical densities are derived for conducting statistical inference. In addition, a modified parametric bootstrap (MPB) and a modified block bootstrap (MBB) procedure are proposed to approximate the finite sample distribution of the change-point estimator, which are shown to work well under any break sizes. Extensive simulation studies are provided to demonstrate the promising performance of the proposed asymptotic and bootstrap distributions. Applications to financial time series are also illustrated.

E1041: **A frequency domain functional approach for time series classification with application to epileptic seizure**
Presenter: **Kun Chen**, Southwestern University of Finance and Economics, China
Co-authors: Rui Huang, Xingzuo He
The automated diagnosis of epileptic seizures via electroencephalogram (EEG) signals poses significant challenges due to their high-dimensional nature and inherent temporal dependencies. Addressing these issues is crucial for improving diagnostic accuracy in clinical settings. A novel frequency domain functional approach is introduced, targeted at enhancing the classification of time series data for such applications. The method centers on the spectral density function, which captures the second-order dynamics of general stationary processes, utilizing log-periodogram ordinates as asymptotically unbiased estimators of the log-spectral density functions. These ordinates, approximately independent across different frequencies, are transformed into smooth curves to facilitate the application of functional principal component analysis combined with a distance-based classification rule. It is shown that the misclassification rates tend to be zero under mild conditions. Based on extensive simulations in various scenarios and a real application to EEG signals of epilepsy seizures, the efficacy of the proposed method is proven.

EO184 Room 103 RECENT ADVANCES IN STATISTICS Chair: Le Zhou

E0189: **Gradient synchronization for multivariate functional data, with application to brain connectivity**
Presenter: **Yaqing Chen**, Rutgers University, United States
Co-authors: Shu-Chin Lin, Yang Zhou, Owen Carmichael, Jane-Ling Wang, Hans-Georg Mueller
Quantifying the association between components of multivariate random curves is of general interest and is a ubiquitous and basic problem that can be addressed with functional data analysis. An important application is the problem of assessing functional connectivity based on functional magnetic resonance imaging (fMRI), where one aims to determine the similarity of fMRI time courses that are recorded on anatomically separated brain regions. In the functional brain connectivity literature, the static temporal Pearson correlation has been the prevailing measure of functional connectivity. However, recent research has revealed temporally changing patterns of functional connectivity, leading to the study of dynamic functional connectivity. This motivates new similarity measures for pairs of random curves that reflect the dynamic features of functional similarity. Specifically, gradient synchronization measures are introduced in a general setting. These similarity measures are based on the concordance and discordance of the gradients between paired smooth random functions. The asymptotic normality of the proposed estimates is obtained under regularity conditions. The proposed synchronization measures are illustrated via simulations and an application to resting-state fMRI signals from the Alzheimer's disease neuroimaging initiative (ADNI), and they are found to improve discrimination between subjects with different disease statuses.

E0679: **Inference for changing periodicity, smooth trend and covariate effects in nonstationary time series**
Presenter: **Lucy Xia**, The Hong Kong University of Science and Technology, Hong Kong
Co-authors: Ming-Yen Cheng, David Siegmund, Shouxia Wang
Traditional analysis of a periodic time series assumes its pattern remains the same over the entire time range. However, some recent empirical studies in climatology and other fields find that the amplitude may change over time, and this has important implications. A formal procedure is developed to detect and estimate change points in the periodic pattern. Often, there is also a smooth trend, and sometimes, the period is unknown, with potential other covariate effects. Based on a new model that takes all of these factors into account, a three-step estimation procedure is proposed to accurately estimate the unknown period, change- points, and varying amplitude in the periodic component, as well as the trend and the covariate effects. First, penalized segmented least squares estimation is adopted for the unknown period, with the trend and covariate effects approximated by B-splines. Then, given the period estimate, a novel SupF statistic is constructed, and it is used in binary segmentation to estimate

change points in the periodic component. Finally, given the period and change-point estimates, the entire periodic component, trend, and covariate are estimated effects using B-splines. Asymptotic results for the proposed estimators are derived, including consistency of the period and change-point estimators and the asymptotic normality of the estimated periodic sequence, trend and covariate effects. Simulation results demonstrate the appealing performance of the new method.

E0995: **Flexible regularized estimating equations: Some new perspectives**
Presenter: **Archer Yang**, McGill University, Canada
The focus is on observations about the equivalences between regularized estimating equations, fixed-point problems and variational inequalities: (a) A regularized estimating equation is equivalent to a fixed-point problem, specified via the proximal operator of the corresponding penalty. (b) A regularized estimating equation is equivalent to a (generalized) variational inequality. Both equivalences extend to any estimating equations with convex penalty functions. To solve large-scale regularized estimating equations, it is worth pursuing computation by exploiting these connections. While fast computational algorithms are less developed for regularized estimating equations, there are many efficient solvers for fixed-point problems and variational inequalities. In this regard, some efficient and scalable solvers are applied which can deliver a hundred-fold speed improvement. These connections can lead to further research in both computational and theoretical aspects of the regularized estimating equations.

EO145 Room 104 NEW ADVANCES IN STATISTICAL LEARNING Chair: Di He

E0346: **Bayesian edge regression: Characterizing observation-specific heterogeneity in estimating undirected graphical models**
Presenter: **Zeya Wang**, University of Kentucky, United States
Bayesian edge regression is a novel edge regression model for undirected graphs, which estimates conditional dependencies as a function of subject-level covariates. By doing so, this model allows for the accounting of observation-specific heterogeneity in estimating networks. Two case studies are presented using the proposed model: one is a set of simulation studies focused on comparing tumor and normal networks while adjusting for tumor purity; the other is an application to a dataset of proteomic measurements on plasma samples from patients with hepatocellular carcinoma (HCC), in which the variation in blood protein networks with disease severity is ascertained.

E0401: **Additive-effect assisted learning**
Presenter: **Jiawei Zhang**, University of Kentucky, United States
Co-authors: Yuhong Yang, Jie Ding
In an increasing number of machine learning applications, multiple learning agents hold datasets that can be collated by a particular identifier and have different features. These agents are often decentralized in nature and may appropriately assist each other in improving modeling performance. A two-stage assisted learning architecture is developed for an agent, Alice, to seek assistance from another agent, Bob, without sharing data. In the first stage, a privacy-aware hypothesis testing-based screening method is proposed for Alice to decide on the usefulness of the data from Bob in a way that only requires Bob to transmit sketchy data. Once Alice recognizes Bob's usefulness, Alice and Bob move to the second stage, where they jointly apply a synergistic model training procedure. Nontrivial theoretical analyses are provided to show that Alice can asymptotically achieve the oracle performance as if the training were from centralized data under appropriate settings. Simulation studies and real data demonstrations, including health condition prediction, image classification, and internet attack detection, show the encouraging performance of the proposed approach.

E0740: **Information theoretic learning meets deep neural networks**
Presenter: **Jun Fan**, Hong Kong Baptist University, Hong Kong
Information theoretic learning, a machine learning approach that incorporates ideas from information theory, offers a family of supervised learning algorithms based on the principle of minimum error entropy (MEE). These algorithms provide an alternative to traditional least squares methods, particularly effective when dealing with heavy-tailed noises or outliers. The integration of information-theoretic learning with deep learning has garnered significant attention in addressing the evolving challenges of modern machine learning. The theoretical exploration of MEE algorithms generated by deep neural networks is delved into in the context of regression tasks. The focus is on establishing fast learning rates for these algorithms when the noise satisfies weak moment conditions.

E0851: **Enhancing the power of OOD detection via sample-aware model selection**
Presenter: **Chuanlong Xie**, Beijing Normal University, China
A novel perspective is presented on detecting out-of-distribution (OOD) samples and an algorithm is proposed for sample-aware model selection to enhance the effectiveness of OOD detection. The algorithm determines, for each test input, which pre-trained models in the model zoo are capable of identifying the test input as an OOD sample. If no such models exist in the model zoo, the test input is classified as an in-distribution (ID) sample. It is theoretically demonstrated that the method maintains the true positive rate of ID samples and accurately identifies OOD samples with high probability when there are a sufficient number of diverse pre-trained models in the model zoo. Extensive experiments were conducted to validate the method, demonstrating that it leverages the complementarity among single-model detectors to consistently improve the effectiveness of OOD sample identification. Compared to baseline methods, the approach improved the relative performance by 65.40% and 37.25% on the CIFAR10 and ImageNet benchmarks, respectively.

EO325 Room 105 AGRICULTURAL ECONOMICS IN CHINA Chair: Shangpu Li

E0723: **Does township-town merger affect the county land supply for polluting industries?**
Presenter: **Xiaodan Zheng**, South China Agricultural University, China
Land resource allocation is closely related to economic activities, and it is impacted by the administrative division adjustment (ADA). Considering China's special land system, local government is of great significance to understanding land allocation. Exploiting the township-town merger (TTM), a difference-in-difference (DID) model is constructed to investigate the impact of ADA on the land transfer of local government in China. It was found that the reform intensified the preference of local governments for land transfer for polluting industries and passed the robustness test. Besides, the reform raises fiscal decentralization at the district/county levels, impacts local governments' horizontal competition, and ultimately increases local governments' preference for the supply of polluting industries' land. And the degree of top-down regulation affects this course. Further evidence indicates that the impact varies depending on the period, place, and economic foundation. The aim is to re-examine the land use effect of the administration-oriented urbanization model, which helps provide new policy ideas for implementing environmental governance in China. Local governments can employ land supply tools to implement environmental governance.

E0958: **Effects of agricultural machinery purchase subsidy on the market structure of machinery operations**
Presenter: **YunQi Wu**, South China Agricultural University, China
Cross-regional outsourced machinery service is an innovation explored in China's agricultural mechanization practices. However, with the implementation of the agricultural machinery purchase subsidy, local machinery operations began to occupy a dominant position, resulting in market share and operation scope shrinkage for COMS. To the best of knowledge, there is no previous literature empirically testing this phenomenon. Hence, the aim is to address this research gap in the literature. To explain the evolutionary law of the machinery operation market structure and the

role of AMPS, the theoretical framework is first explored, including deductive induction reasoning and mathematical analysis, and then an empirical analysis is based on the panel data of 31 provinces of China over the 2000-2019 period. The results show that: First, since the implementation of AMPS, the relationship between COMS and LMO has changed from complements to substitutes. Second, an increase in the AMPS amount directly leads to a proportional decrease in COMS of the according province. Finally, heterogeneity analysis reveals that AMPS only affects the market structure of agricultural machinery operation in the tilling and sowing phases but not in the harvesting phase.

E0960: **The impact of digital technology application on employment quality in China: A perspective on employment inequality**
Presenter: **Lin Xie**, South China Agricultural University, China

The purpose is to utilize data from three waves of the China family panel studies (CFPS) in 2014, 2016, and 2018 to examine the impact of digital technology applications on the employment quality of different groups from the perspective of employment inequality. Empirical results indicate that the application of digital technology has a significant positive effect on enhancing employment quality, income, stability, and the environment. These results remain robust after considering endogeneity and a series of robustness tests. However, the empirical findings also reveal a "double-edged sword" effect of digital technology applications on the quality of employment in China's labor market: on the one hand, digital technology application helps increase the overall income level of women and rural laborers, narrow income disparities, and increases employment opportunities for laborers in central and western regions. Therefore, the application of digital technology is crucial for reducing employment inequality. On the other hand, low-skilled laborers and the elderly face challenges in income growth and development opportunities in the digital wave, and the application of digital technology leads to decreased employment stability for rural laborers.

E1006: **Effect of the rural collective-owned commercial construction land marketization reform on land misallocation**
Presenter: **Dan Cheng**, University of Electronic Science and Technology of China, China

To integrate the rural-urban land market, a pilot reform of rural collectively-owned commercial construction land (RCOCCL) circulation was implemented in 33 counties in 2015. The aim is to explore the impact of the reform of RCOCCL circulation on land misallocation to provide a reference for the implementation of a new round of pilot. The conclusion shows that, firstly, the reform increased the county's land resource mismatch degree. Secondly, the direction of land misallocation, pilot time, and location of the pilot county affected the policy effect of the reform on land resource mismatches. Thirdly, the reform improved the degree of marketization, but it didn't change the path dependence of local governments on land-driven development. Through the expansion of urban investment bonds and discounted transfers for industrially occupied land, the reform increased the degree of land misallocation. In summary, the reform of RCOCCL trade increased the degree of county land misallocation. This is primarily due to the imperfect land market institutions and the path dependence of the traditional economic development model. Policy implications to perfect the financing system, move away from land dependence and foster a new development mode are concluded with.

EO183 Room 106 RECENT ADVANCES IN HIGH-DIMENSIONAL CHANGE POINT INFERENCE Chair: Runmin Wang

E0444: **Adaptive matrix change point detection: Leveraging structured mean shifts**
Presenter: **Xinyu Zhang**, Univerisity of Iowa, United States
Co-authors: Kung-Sik Chan

In high-dimensional time series, the component processes are often assembled into a matrix to display their interrelationship. The focus is on detecting mean shifts with unknown change point locations in these matrix time series. Series that are activated by a change may cluster along certain rows (columns), which forms mode-specific change point alignment. Leveraging mode-specific change point alignments may substantially enhance the power for change point detection. Yet, there may be no mode-specific alignments in the change point structure. A powerful test is proposed to detect mode-specific change points, yet robust to non-mode-specific changes. It shows the validity of using the multiplier bootstrap to compute the p-value of the proposed methods and derive non-asymptotic bounds based on the size and power of the tests. A parallel bootstrap is also proposed as a computationally efficient approach for computing the p-value of the proposed adaptive test. In particular, the consistency of the proposed test is shown under mild regularity conditions. To obtain the theoretical results, new, sharp bounds on Gaussian approximation and multiplier bootstrap approximation are derived, which are of independent interest for high dimensional problems with diverging sparsity.

E0853: **Change point detection for high-dimensional linear models: A general tail-adaptive approach**
Presenter: **Bin Liu**, School of Management at Fudan University, China

The change point detection problem is studied for high-dimensional linear regression models. The existing literature mainly focused on the change point estimation with stringent sub-Gaussian assumptions on the errors. In practice, however, there is no prior knowledge about the existence of a change point or the tail structures of errors. To address these issues, a novel tail-adaptive approach is proposed for simultaneous change point testing and estimation. The method is built on a new loss function, which is a weighted combination between the composite quantile and least squared losses, allowing the borrowing of information on the possible change points from both the conditional mean and quantiles. Under some mild conditions, the validity of the new tests is justified in terms of size and power under the high-dimensional setup. The corresponding change point estimators are shown to be rate optimal up to a logarithm factor. Moreover, combined with the wild binary segmentation technique, a new algorithm is proposed to detect multiple change points in a tail-adaptive manner. Extensive numerical results are conducted to illustrate the appealing performance of the proposed method.

E1063: **Exact and assumption-lean change-point detection with applications in post-detection inference**
Presenter: **Guanghui Wang**, East China Normal University, China

A novel framework for exact change-point detection that ensures valid Type I error rate control in finite samples is introduced. This versatile framework is applicable across a range of change-point models, including high-dimensional contexts. Additionally, its use is explored in quantifying uncertainty in detected change-points, enhancing the reliability of post-detection inference.

E1110: **Controlling FDR of change points in structural break time series**
Presenter: **Wei Zhia Kua**, The Chinese University of Hong Kong, Hong Kong
Co-authors: Chun Yip Yau

ScoreFDR is proposed, a procedure using global segmentation in a multiscale way with score statistics to find multiple change points in an autoregressive (AR) process. In contrast to traditional methods that are focused on achieving consistency on estimated changepoints, our approach is specifically designed for false discovery rate (FDR) control, i.e., controlling the expected proportion of falsely detected change points against all identified change points. The key to controlling the FDR involves constructing a multiscale quantile constraint that enables the control of local errors within individual segments. We proved that ScoreFDR is able to control the FDR asymptotically. The performance of FDR control in finite samples is illustrated via extensive simulation studies. Applications to real data examples are also illustrated.

EO030 Room 108 STATISTICAL LEARNING ON DATA WITH SOPHISTICATED STRUCTURES AND DEPENDENCE Chair: Wen Zhou

E0443: **BELIEF in dependence: Leveraging atomic linearity in data bits for rethinking generalized linear models**
Presenter: **Kai Zhang**, University of North Carolina at Chapel Hill, United States
Co-authors: Xiao-Li Meng, Benjamin Brown

Two linearly uncorrelated binary variables must also be independent because non-linear dependence cannot manifest with only two possible states. This inherent linearity is the atom of dependency constituting any complex form of relationship. Inspired by this observation, we develop a framework called binary expansion linear effect (BELIEF) for understanding arbitrary relationships with a binary outcome. Models from the BELIEF framework are easily interpretable because they describe the association of binary variables in the language of linear models, yielding convenient theoretical insight and striking Gaussian parallels. With BELIEF, one may study generalized linear models (GLM) through transparent linear models, providing insight into how the choice of link affects modelling. For example, setting a GLM interaction coefficient to zero does not necessarily lead to the kind of no-interaction model assumption as understood under their linear model counterparts. Furthermore, for a binary response, maximum likelihood estimation for GLMs paradoxically fails under complete separation when the data are most discriminative, whereas BELIEF estimation automatically reveals the perfect predictor in the data that is responsible for complete separation. These phenomena are explored, and related theoretical results are provided. A preliminary empirical demonstration of some theoretical results is also provided.

E0457: **Neyman-Pearson and equal opportunity: When efficiency meets fairness in classification**
Presenter: **Xin Tong**, University of Southern California, United States
Co-authors: Jianqing Fan, Shunan Yao, Yanhui Wu

Organizations often rely on statistical algorithms to make socially and economically impactful decisions. The fairness issues in these important automated decisions are addressed. On the other hand, economic efficiency remains instrumental in organizations survival and success. Therefore, a proper dual focus on fairness and efficiency is essential in promoting fairness in real-world data science solutions. Among the first efforts towards this dual focus, the equal opportunity (EO) constraint is incorporated into the Neyman-Pearson (NP) classification paradigm. Under this new NP-EO framework, the oracle classifier is derived, finite-sample-based classifiers are proposed that satisfy population-level fairness and efficiency constraints with high probability, and the statistical and social effectiveness of the algorithms is demonstrated on simulated and real datasets.

E0486: **On the testing of multiple hypothesis in sliced inverse regression**
Presenter: **Zhigen Zhao**, Temple University, United States
Co-authors: Xin Xing

Multiple tests of the general regression framework are considered, with the aim of studying the relationship between a univariate response and a p-dimensional predictor. To test the hypothesis of the effect of each predictor, an angular balanced statistic (ABS) is constructed based on the estimator of the sliced inverse regression without assuming a model of the conditional distribution of the response. According to the developed limiting distribution results, ABS is shown to be asymptotically symmetric with respect to zero under the null hypothesis. A model-free multiple testing procedure is proposed using angular balanced statistics (MTA), and it is theoretically shown that the false discovery rate of this method is less than or equal to a designated level asymptotically. Numerical evidence has shown that the MTA method is much more powerful than its alternatives, subject to the control of the false discovery rate.

E0522: **Sparse heteroskedastic PCA in high dimensions**
Presenter: **Zhao Ren**, University of Pittsburgh, United States

Principal component analysis (PCA) is one of the most commonly used techniques for dimension reduction and feature extraction. Though it has been well-studied for high-dimensional sparse PCA, little is known when the noise is heteroskedastic, which turns out to be ubiquitous in many scenarios, like single-cell RNA sequencing (scRNA-seq) data and information network data. An iterative algorithm is proposed for sparse PCA in the presence of heteroskedastic noise, which alternatively updates the estimates of the sparse eigenvectors using orthogonal iteration with adaptive thresholding in one step and imputes the diagonal values of the sample covariance matrix to reduce the estimation bias due to heteroskedasticity in the other step. The procedure is computationally fast and provably optimal under the generalized spiked covariance model, assuming the leading eigenvectors are sparse. A comprehensive simulation study shows its robustness and effectiveness in various settings. Additionally, the application of the new method to two high-dimensional genomics datasets, i.e., microarray and scRNA-seq data, demonstrates its ability to preserve inherent cluster structures in downstream analyses.

EO232 Room 109 STATISTICAL/MACHINE LEARNING AND APPLICATIONS Chair: Jingjing Wu

E0867: **Analysis of the impact mechanism of talent agglomeration on the high quality development of Chinese urban economy**
Presenter: **Shengchun Ma**, Minzu University of China, China
Co-authors: Yuzhu Zheng

Exploring the impact mechanism of talent agglomeration on promoting high-quality economic development has important theoretical and practical significance for promoting high-quality economic development in Chinese cities and also has certain reference value for the study of the relationship between talent agglomeration and economic development in other countries or regions. The article is based on China's population census data from 2010 and 2020, as well as China's urban statistical yearbook data from 2010 to 2020. Using the spatial panel data regression analysis method, it deeply analyzes the impact mechanism of talent agglomeration on the high-quality economic development of 281 prefecture-level and above cities in China. Research has shown that urban talent agglomeration has a positive promoting effect on high-quality economic development; Technological innovation has a partial mediating effect on the high-quality development of regional economy in talent aggregation; Talent aggregation has a positive promoting effect on the high-quality development of the local economy, while also having a positive promoting effect on the high-quality development of neighboring areas; The improvement of local talent aggregation level also has a significant spatial spillover effect on neighboring areas.

E0707: **Exploring the evolution law of intelligent voice technology using text mining**
Presenter: **Jinluan Ren**, Communication University of China, China
Co-authors: Rui Huang, Junjiang Liu, Huiwen Deng, Bo Li

Intelligent voice technology (IVT), as a 'main artery' for advancing human-computer interaction, has a profound impact on the development of artificial intelligence technology. Studying the competitive advantages and development trends of IVT and summarizing its evolutionary law has theoretical significance for clarifying the development trends of IVT and exploring strategies to advance its development. Several methods are comprehensively used to lock the IVT keyword set, ultimately identifying 22,355 patents related to IVT. Moreover, this research employs the economic fitness-complexity (EFC) method to calculate national fitness and technological complexity and analyze the global evolution patterns of IVT. It is found that the leapfrog development of IVT occurred between 2011 and 2020, and the development of IVT subfields has its own characteristics. A country with high adaptability holds an absolute dominant position in the IVT field. If there is no significant technological barrier in the development of IVT, then it is possible for relatively lagging countries to achieve leapfrogging in the field of IVT.

E0999: **Reinforcement learning in credit risk management**
Presenter: **Jorge C-Rella**, University of A Coruna, Spain
Co-authors: David Martinez Rego, Juan Vilar Fernandez
Credit risk problems are dynamic because customer behavior is not stable, and they are cost-sensitive in the sense that a decision's impact depends on the loan amount. In addressing these challenges, online learning algorithms serve as valuable tools, adapting in real-time as new data becomes available. However, relying solely on approved transactional data introduces potential unfair biases and opportunity costs. Within reinforcement learning, bandit algorithms offer a solution by effectively balancing the trade-off between exploiting current models and exploring actions with limited information to improve predictions. Novel dynamic learning strategies are presented, which extend online learning and bandit algorithms to cost-sensitive classification. Empirical evaluations conducted on benchmark datasets and through extensive simulation studies corroborate the effectiveness and efficiency of the proposed methodologies.

E0234: **The impact of news-based and Twitter-based economic uncertainty on realized volatility**
Presenter: **Shaonan Tian**, San Jose State University, United States
Co-authors: Qing Bai, Cathy W-S Chen
Employing a two-regime threshold quantile autoregressive model with exogenous variables and GARCH specification, the focus is on looking into the dynamic relationship between perceived uncertainty and realized equity volatility using Twitter-based and newspaper-based economic uncertainty measures across diverse market conditions. Findings reveal that these indicators capture various facets of uncertainty and display diverse behavior in stable and volatile markets. Specifically, Twitter-based indicators show a pronounced positive impact on realized volatility during market upturns, while the newspaper-based indicator becomes more relevant in highly volatile conditions. Factors highlighted that drive aggregate market volatility can shift between different regimes, even under the same market conditions.

EO111 Room 110 RECENT ADVANCEMENTS IN EXPERIMENTAL DESIGN AND ITS APPLICATION Chair: Fasheng Sun

E0420: **Uniform designs of experiments with mixtures under the criterion mean L1-distance**
Presenter: **Yaping Wang**, East China Normal University, China
Mixture experiments analyze how changes in component proportions impact the response variable within the experimental region of a simplex. A new criterion, named the mean L1-distance (ML1D) criterion is introduced, for constructing uniform designs in mixture experiments. This criterion allows flexibility in point size and showcases a more uniform pattern within the experimental region. The optimal Scheffe-type simplex-lattice designs are also explored under the ML1D criterion. An interesting discovery is that the uniform mixture designs and the optimal Scheffe-type simplex-lattice designs are connected. For a two-component mixture design, these two types of designs are proven to be equivalent. For more than two-component mixture designs, numerical equivalences between the two designs are observed. These findings strengthen the rationale for users to adopt these designs in mixture experiments for modeling and prediction.

E0488: **Theory and construction of complex experimental designs**
Presenter: **Chunyan Wang**, Renmin University, China
Fractional factorial designs are important designs for experiments. They can be divided into regular designs and nonregular designs. As a special kind of nonregular design, the parallel flats design has received more and more attention. Parallel flat designs (PFDs) consisting of three parallel flats (3-PFDs) are the most frequently utilized PFDs due to their simple structure. Generalizing to f-PFD with $f > 3$ is more challenging. A method for obtaining the confounding frequency vectors is proposed for all nonequivalent f-PFDs to find the least G-aberration f-PFD from any single flat. PFDs are particularly useful for constructing nonregular fractions, split plots, or randomized block designs. The quaternary code design series is also characterized as PFDs. Finally, it shows how designs constructed by concatenating regular fractions from different families may also have a parallel flat structure. Moreover, a different approach is pursued, applying coordinate exchange optimization to the structure of the parallel flat and employing an efficient computation for the confounding frequency vector. Beginning with any single flat design, the proposed algorithm can construct an efficient f-PFD in terms of G-aberration. A user-friendly software function called GMAPFDace has also been developed in MATLAB and GUN Octave to implement this algorithm, which allows one to obtain low G-aberration PFDs with the required sizes easily and quickly.

E0714: **Generalized uniform projection designs**
Presenter: **Yishan Zhou**, Qingdao University of Science & Technology, China
Uniform projection designs (UPDs) are an important class of computer experiment designs that address design space-filling properties for low-dimensional projections. Especially in most computer experimental designs, only a few factors are active. In this case, UPDs are very suitable. However, the existing UPDs only care about the uniformity of designs corresponding to the main effects model. In more complex practical models that include second-order interaction effects, it is not sufficient to consider linear main effects alone. To reduce the impact of the second-order interaction effect on the main effect, the 3-dimensional uniform projection designs are proposed, which can realize the 3-dimensional projection uniformity in all projection spaces. In addition, some theoretical connections are established between the 3-dimensional uniform projection criterion and the criteria of maximin distance and 3-orthogonality, which help find better space-filling designs under multiple criteria. Moreover, the 2-dimensional and 3-dimensional projection uniformity of a design is also comprehensively considered, and the generalized uniform projection criterion is proposed. The corresponding optimal design is called the minimum uniform projection design and provides the optimization search algorithm. It can simultaneously achieve 2-dimensional and 3-dimensional projection uniformity in all projection spaces. Numerical simulations confirm the rationality of the proposed designs.

E0699: **An efficient global optimization method for sequential order-of-addition experiments**
Presenter: **Jianbin Chen**, Beijing Institute of Technology, China
The order-of-addition (OofA) experiment has received a great deal of attention in the recent literature. The primary goal of the OofA experiment is to identify the optimal order in a sequence of m components. Existing methods require pre-specifying a model and cannot flexibly adjust the model according to the progress of the experiment. Moreover, these methods are not applicable to deal with large m (e.g., m 7). With this in mind, this article proposes an efficient global optimization algorithm based on the sequential experiment and variable selection, utilizing information gained during the iterative process. Moreover, an efficient construction method for the optimal initial design, which achieves space-filling and balance properties for OofA components, is proposed. Theoretical supports are given to illustrate the effectiveness of the proposed method. The proposed method is able to obtain the optimal order for large m efficiently. Numerical experiments are used to demonstrate the effectiveness of the proposed method.

EO193 Room 212 MODERN SEMIPARAMETRIC METHODS WITH APPLICATIONS Chair: Myungjin Kim

E0289: **Amenity alchemy: Unveiling the dual nature of amenities in shaping regional futures**
Presenter: **Yong Chen**, Oregon State University, United States
Co-authors: Myungjin Kim
The dual nature of amenities is explored: quality-of-life effects on households and productivity effects on firms. It reintroduces the notion that amenities can influence production costs, worker productivity, and the business environment. These features are integrated into a quantitative spatial equilibrium model to reveal the complex dynamics in amenity-driven growth. The model results suggest that the net impacts of amenity-driven migration hinge on both the qualify-of-life effects and the productivity effects of amenities. Empirical evidence highlights the interplay between the dual natures of amenities. By revealing the prominence of productivity aspects, the research shows that amenity-driven growth is not a panacea for the development of rural communities.

E0383: **Data integration with nonprobability sample: Semiparametric model-assisted approach**
Presenter: **Danhyang Lee**, Baylor University, United States
Co-authors: Sixia Chen
A novel semiparametric model-assisted estimation method is introduced that integrates data from both probability and nonprobability samples, thereby facilitating robust and efficient inferences regarding finite population parameters. To mitigate selection bias, whether ignorable or nonignorable, associated with the nonprobability sample, a flexible semiparametric propensity score model that extends beyond the missing at-random assumption is proposed. The approach employs a pseudo-profile-likelihood method to estimate the propensity score model. Subsequently, a difference estimator is constructed utilizing the probability sample as a foundation, where the proxy values of the study variable for the finite population are derived from the nonprobability sample using the estimated propensity score model. The asymptotic properties of the proposed estimators are presented, and formulae for variance estimation are provided. Through a series of simulations and a real data application, the proposed estimation procedure is validated, and its superiority over some existing estimators is demonstrated.

E0595: **Variable selection for ultra-high-dimensional generalized spatial partial varying coefficient models**
Presenter: **Jingru Mu**, Kansas State University, United States
The authors propose a generalized partially linear spatially varying coefficient model (GPLSVCM) to accommodate different data types and allow more flexibility simultaneously. The purpose is to study the estimation and structure identification for ultra-high-dimensional generalized partially linear spatial varying coefficient models. A fast and efficient procedure is proposed for identifying model structure via the group adaptive lasso approach and estimating models via spline approaches. The method is shown to be consistent for model structure selection and estimation. The asymptotic normality for the linear components has also been constructed. Simulation studies are conducted to evaluate the performance and use a real spatial dataset to illustrate the application of the proposed method.

E1064: **A fast and flexible space-time varying coefficient model selection**
Presenter: **Daisuke Murakami**, The Institute of Statistical Mathematics, Japan
Co-authors: Shinichiro Shirota, Mami Kajita, Seiji Kajita
The space-time varying coefficient (STVC) model attracts attention these days as a flexible tool to explore the spatiotemporal patterns in regression coefficients. However, the model tends to suffer from difficulty in balancing computational efficiency and model flexibility. A fast and flexible STVC modeling method has been developed to break the bottleneck. For flexible modeling, multiple processes are assumed in each varying coefficient, including purely spatial, purely temporal, and space-time interaction processes with/without time cyclicity. While consideration of multiple processes can be time-consuming, a pre-conditioning method is combined, and a model selection procedure inspired by reluctant interaction modelling to select/specify the latent space-time structure computationally efficiently. Monte Carlo experiments show that the proposed method outperforms alternatives in terms of coefficient estimation accuracy and computational efficiency. Finally, the proposed method is applied to a crime analysis with a sample size of 279,360 and confirmed that the proposed method provides reasonably varying coefficient estimates.

EO041 Room 202 RECENT DEVELOPMENT OF DIMENSION REDUCTION AND SEMIPARAMETRIC REGRESSION Chair: Jing Zeng

E0280: **Detecting influential observations in single-index models with metric-valued response objects**
Presenter: **Abdul-Nasah Soale**, Case Western Reserve University, United States
Regression with random data objects is becoming increasingly common in modern data analysis. Unfortunately, like the traditional regression setting with Euclidean data, random response regression is not immune to the trouble caused by unusual observations. A metric Cook's distance extending the classical Cook's distances to general metric-valued response objects is proposed. The performance of the metric Cook's distance in both Euclidean and non-Euclidean response regression with Euclidean predictors is demonstrated in an extensive experimental study. A real data analysis of county-level COVID-19 transmission in the United States also illustrates the usefulness of this method in practice.

E0406: **High-dimensional differential networks with sparsity and reduced-rank**
Presenter: **Cheng Wang**, Shanghai Jiao Tong University, China
Differential network analysis plays a crucial role in capturing nuanced changes in conditional correlations between two samples. Under the high dimensional setting, the differential network, i.e., the difference between the two precision matrices, is usually stylized with sparse signals and some low-rank latent factors. Recognizing the distinctions inherent in the precision matrices of such networks, a novel approach is introduced, termed "SR-Network", for the estimation of sparse and reduced-rank differential networks. This method directly assesses the differential network by formulating a convex empirical loss function with ℓ_1-norm and nuclear norm penalties. Finite-sample error bounds are established for parameter estimation and highlight the superior performance of the proposed method through extensive simulations and real data studies. The significant contribution is to the advancement of methodologies for accurate analysis of differential networks, particularly in the context of structures characterized by sparsity and low-rank features.

E0461: **Optimal sparse sliced inverse regression via random projection**
Presenter: **Jia Zhang**, Southwestern University of Finance and Economics, China
A novel simple sparse sliced inverse regression method is proposed based on random projections in a large p small n setting. Embedded in a generalized eigenvalue framework, the proposed approach finally reduces to parallel execution of low-dimensional (generalized) eigenvalue decompositions, which facilitates high computational efficiency. Theoretically, it is proven that this method achieves the minimax optimal rate of convergence under suitable assumptions. Furthermore, the algorithm involves a delicate reweighting scheme, which can significantly enhance the identifiability of the active set of covariates. Extensive numerical experiments demonstrate the high superiority of the proposed algorithm in comparison to competing methods.

E0871: **Sliced average variance estimation for tensor data**
Presenter: **Chuanquan Li**, Jiangxi University of Finance and Economics, China

Tensor data have been widely used in many fields, e.g., modern biomedical imaging, chemometrics, and economics, but suffer from some common issues, such as high dimensional statistics. How to find their low-dimensional latent structure is of great interest. To this end, two efficient tensor-sufficient dimension reduction methods are developed based on the sliced average variance estimation (SAVE) to estimate the corresponding dimension reduction subspaces. The first one, entitled tensor sliced average variance estimation (TSAVE), works well when the response is discrete or takes finite values but is not $\sqrt{n}$ consistent for continuous response; the second one, named bias-correction tensor sliced average variance estimation (CTSAVE), is a de-biased version of the TSAVE method. The asymptotic properties of both methods are derived under mild conditions. Simulations and real data examples are also provided to show the superiority of the efficiency of the developed methods.

EO213 Room 207 HIERARCHICAL AND JOINT STATISTICAL MODELS IN HEALTH AND APPLICATIONS Chair: Gang Han

E0168: **Multi-way overlapping clustering by Bayesian tensor decomposition**
Presenter: **Zhuofan Wang**, Institute of Statistics and Big Data, Renmin University of China, China

The development of modern sequencing technologies provides great opportunities to measure gene expression of multiple tissues from different individuals. The three-way variation across genes, tissues, and individuals makes statistical inference challenging. A Bayesian multi-way clustering approach is proposed to cluster genes, tissues, and individuals simultaneously. The proposed model adaptively trichotomizes the observed data into three latent categories and uses a Bayesian hierarchical construction to further decompose the latent variables into lower-dimensional features, which can be interpreted as overlapping clusters. With a Bayesian nonparametric prior, i.e., the Indian buffet process, the method determines the cluster number automatically. The utility of the approach is demonstrated through simulation studies and an application to the genotype-tissue expression (GTEx) RNA-seq data. The clustering result reveals some interesting findings about depression-related genes in the human brain, which are also consistent with biological domain knowledge.

E0205: **A meta-analysis based hierarchical variance model for powering one and two-sample t-tests**
Presenter: **Xinlei Wang**, University of Texas at Arlington, United States
Co-authors: Jackson Barth

Sample size determination (SSD) is essential in statistical inference and hypothesis testing, as it directly affects the accuracy and power of the analysis. An SSD methodology is proposed for one and two-sample t-tests that ensure clinical relevance using a pre-determined unstandardized effect size. The novel approach leverages Bayesian meta-analysis to account for the uncertainty surrounding the variance, a common issue in SSD. By incorporating prior knowledge from related studies via a Bayesian gamma-inverse gamma model, an informative posterior predictive distribution is obtained for the variance that leads to better decisions about sample size. An empirical Bayes approach is proposed for efficient posterior sampling, which is further combined with a discretized simulation approach to facilitate computation. Simulations and empirical studies demonstrate that the methodology outperforms other aggregate approaches (simple average, weighted average, median) in variance estimation for SSD, especially in meta-analyses with large disparities in sample size and moderate variance. Thus, it offers a robust and practical solution for sample size determination in t-tests.

E0324: **Designing a reinforcement learning agent to facilitate consensus among human doctors in rare disease treatment**
Presenter: **Yinghao Fu**, Chinese university of hongkong, ShenZhen, China
Co-authors: Shuang Li

Diagnosing and treating rare diseases pose distinctive challenges, mainly due to the intricate nature of patient symptoms, limited awareness among physicians, and the infrequency of individual cases. These challenges often make it difficult for clinicians to reach a diagnostic consensus, leading to suboptimal treatment strategies. To address these challenges and elevate diagnostic accuracy for rare disease patients, a reinforcement learning framework is introduced, that combines artificial intelligence with the expertise of medical professionals. The AI agent model is built upon a joint optimization framework that seamlessly integrates three pivotal modules: active perception, human doctors' opinion aggregation, and probabilistic treatment prediction. Through iterative optimization, these modules collaborate to fine-tune the diagnostic process until a consensus is achieved among expert doctors, ultimately enhancing patient outcomes. The integrated framework adeptly navigates the intricacies of rare disease diagnosis, fostering more precise diagnoses and improved patient results.

E0385: **Estimation and sequential forecast of disease progression in the absence of true disease state process**
Presenter: **Zexi Cai**, Columbia University, United States
Co-authors: Yuanjia Wang

Forecasting future disease progression based on patients' evolving health information is challenging when limited by diagnostic capabilities. For example, the absence of gold-standard neurological diagnoses due to a lack of use of objective biomarkers hinders distinguishing Alzheimer's disease (AD) from related conditions such as AD-related dementias (ADRDs) and Lewy body disease (LBD). Despite the increasing use of biomarkers, not everyone has access to them, and some practitioners may not utilize them, resulting in less precise diagnoses. Borrowing information from a series of temporally dependent surrogate labels and health markers may improve the accuracy of future disease prediction. Integrating the hidden Markov model is proposed as a generative model to handle erroneous clinical diagnoses with a time-varying multinomial logistic regression as a discriminative model to identify features of disease progression. An adaptive forward-backwards algorithm is developed to facilitate parameter estimation with pseudo-expectation maximization, with a penalty introduced when many feature variables are present. Furthermore, the posterior rule and the Viterbi algorithm are developed to forecast disease progression. Asymptotic properties are established, and performance with finite samples is demonstrated via simulation studies. Analysis of the neuropathological dataset of the National Alzheimer's Coordinating Center (NACC) shows improved accuracy in distinguishing LBD from AD.

EO238 Room 209 FAST DENOISING TECHNIQUES FOR COMPLEX DATA STRUCTURES Chair: Weixing Song

E0589: **Goodness-of-fit tests in functional-coefficient autoregressive models with measurement error**
Presenter: **Pei Geng**, University of New Hampshire, United States

The functional-coefficient autoregressive (FAR) models are flexible to fit the nonlinear patterns in time series data. The aim is to introduce a goodness-of-fit test for the FAR models when the time series is observed with measurement error. The calibrated autoregressive model based on the observed time series is represented, and the test for the parametric functional coefficients is constructed based on a marked empirical process with the residuals and the covariate. The asymptotic property shows that the proposed test is asymptotically distribution-free. A finite-sample simulation study is conducted to demonstrate the empirical level and power under certain alternatives.

E0782: **High-dimensional regression adjustment estimation for average treatment effect with highly correlated covariates**
Presenter: **Lili Yue**, Nanjing Audit University, China

Regression adjustment is often used to estimate the average treatment effect (ATE) in randomized experiments. Recently, some penalty-based regression adjustment methods have been proposed to handle the high-dimensional problem. However, these existing high-dimensional regression

adjustment methods may fail to achieve satisfactory performance when the covariates are highly correlated. A novel adjustment estimation method is proposed for ATE by combining the semi-standard partial covariance (SPAC) and regression adjustment methods. Under some regularity conditions, the asymptotic normality of the proposed SPAC adjustment ATE estimator is shown. Some simulation studies and an analysis of HER2 breast cancer data are carried out to illustrate the advantage of the proposed SPAC adjustment method in addressing the highly correlated problem of the Rubin causal model.

E1098: **Distance-based clustering of functional data with derivative principal component analysis**
Presenter: **Jianhong Shi**, Shanxi Normal University, China
Functional data analysis (FDA) is an important modern paradigm for handling infinite-dimensional data. An important task in FDA is clustering, which identifies subgroups based on the shapes of measured curves. Considering that derivatives can provide additional useful information about the shapes of functionals, we propose a novel L^2 distance between two random functions by incorporating the functions and their derivative information to determine the dissimilarity of curves under a unified scheme for dense observations. The Karhunen-Loeve expansion is utilized to approximate the curves and their derivatives. Cluster membership prediction for each curve intends to minimize the new distances between the observed and predicted curves through subspace projection among all possible clusters. Consistent estimators for the curves, curve derivatives, and the proposed distance are provided. Identifiability issues of the clustering procedure are also discussed. The utility of the proposed method is illustrated via simulation studies and applications to two real datasets. The proposed method can considerably improve cluster performance compared with existing functional clustering methods.

E1097: **Composite expectile estimation in partial functional linear regression model**
Presenter: **Ping Yu**, Shanxi Normal University, China
Recent research and substantive studies have shown a growing interest in expectile regression (ER) procedures. Similar to quantile regression, ER concerning different expectile levels can provide a comprehensive picture of the conditional distribution of a response variable given predictors. Three composite-type ER estimators are proposed to improve estimation accuracy. The proposed ER estimators are the composite estimator, which minimizes the composite expectile objective function across expectiles; the weighted expectile average estimator, which takes the weighted average of expectile-specific estimators; and the weighted composite estimator, which minimizes the weighted composite expectile objective function across expectiles. Under certain regularity conditions, the convergence rate of the slope function is derived, the mean squared prediction error is obtained, and the asymptotic normality of the slope vector is established. Simulations are conducted to assess the empirical performance of various estimators. An application to the analysis of capital bike share data is presented. The numerical evidence endorses the theoretical results and confirms the superiority of the composite-type ER estimators to the conventional least squares and single ER estimators.

EO044 Room 210 SPATIAL STATISTICS Chair: Pei-Sheng Lin

E0365: **Estimation and selection for spatial zero-inflated count models**
Presenter: **Chun-Shu Chen**, National Central University, Taiwan
Co-authors: Chung-Wei Shen
The count data arise in many scientific areas. The focus is on spatial count responses with an excessive number of zeros and a set of available covariates. How to estimate model parameters, as well as a selection of important covariates for spatial zero-inflated count models, are both essential. Importantly, to alleviate deviations from model assumptions, a spatial zero-inflated Poisson-like methodology is proposed to model this type of data, which just relies on assumptions for the first two moments of spatial count responses. An effectively iterative estimation procedure is then designed between the generalized estimating equation and the weighted least squares method to estimate the regression coefficients and the variogram of the data model, respectively. Moreover, the stabilization of estimators is evaluated via a block jackknife technique. Further, a distribution-free model selection criterion based on an estimate of the mean squared error of the estimated mean structure is proposed to select the best subset of covariates. Numerical results show the effectiveness of the proposed methods.

E0733: **Instrumental variable estimation and inference for spatial autoregressive geographically weighted quantile regression**
Presenter: **Vivian Yi-Ju Chen**, National Chengchi University, Taiwan
Past years have witnessed significant advancements in spatial modelling techniques that allow simultaneously dealing with spatial heterogeneity in the regression coefficients and the spatial autoregressive lag in the response variable. Spatial autoregressive geographically weighted quantile regression (GWQR-SAR) is one such technique that has recently been devoted to the literature for conducting spatial quantile-based analysis. GWQR-SAR is proposed as a new estimation method, termed instrumental variable quantile estimation. The associated inference properties are also derived, which offer a covariance matrix estimate that is simpler to construct compared to the existing method. To strengthen the theoretical framework of GWQR-SAR, bootstrap tests are further developed to identify constant parameters, as well as the semiparametric modelling framework. The proposed methodologies are then evaluated through simulations. Lastly, an empirical example is given to illustrate the application of the approach.

E0908: **Association of human mobility and weather conditions with dengue mosquito abundance in three areas in Hong Kong**
Presenter: **Hsiang-Yu Yuan**, City University of Hong Kong, Hong Kong
Co-authors: Yufan Zheng, Keqi Yue, Eric Wing Ming Wong
While Aedes mosquitoes, the Dengue vectors, are expected to expand their spread due to international travel and climate change recently, the effects of human mobility and low rainfall conditions on them remain largely unknown. The aim is to assess these influences during the COVID-19 pandemic in Hong Kong, characterized by varying levels of human mobility. Mobility indices (including residential, parks, and workplaces), weather conditions (total rainfall and mean temperature), and measurements of Aedes albopictus' abundance are obtained using Gravid traps between April 2020 and August 2022. The analysis revealed that both low rainfall (<50 mm) after 4.5 months and heavy rainfall (>500 mm) within 3 months were both associated with higher relative risks (RRs), 1.73 and 1.41, of mosquito abundance, compared to 300 mm. Warmer conditions (21-30 C, compared with 20 C) were associated with a higher RR (1.47) after half a month. Furthermore, residential mobility showed a negative association with mosquito abundance. The model projected that if residential mobility in 2022 returned to pre-pandemic levels, mosquito abundance would increase by an average of 80.49% compared to observed levels.

E0942: **Local linear estimation for covariate-dependent coefficients model in disease mapping**
Presenter: **Feng-Chang Lin**, University of North Carolina at Chapel Hill, United States
Co-authors: Pei-Sheng Lin, Jun Zhu, Yexuan Jiang
Spatial regression effects may depend on covariates in the disease mapping modeling. Taking infectious disease, for example, the association between incidence and risk factors may vary by climatic factors such as season and temperature. The aim is to build a model that can simultaneously study spatial and varying effects of covariates that are deemed to modulate the spatial association. A local linear estimation method is employed for the covariate-dependent coefficients in a spatial model dealing with excess zero counts. The local linear estimator is designed to smooth time-dependent coefficient estimation effectively. Comprehensive simulation studies were used to demonstrate the performance of our local linear estimators under various scenarios. Dengue incidences in the villages using weekly reported dengue cases in Kaohsiung City from January 2014

to December 2015 are used to offer insights into the proposed method for the practical use of real-world applications.

EO012 Room 307 RECENT ADVANCES IN FUNCTIONAL DATA ANALYSIS (VIRTUAL) Chair: Eftychia Solea

E0478: **Sparse modeling and non-asymptotic bounds via correlation operator of multivariate functional data**
Presenter: **Jun Song**, Korea University, Korea, South

Functional data analysis has gained significant attention in recent years, particularly in the context of scalar-on-function regression problems involving multivariate functional predictors. However, existing methods for predictor selection in this domain often lack theoretical validity or rely on overly stringent assumptions that may not hold in practice. A novel approach is presented to functional predictor selection that addresses these limitations. Correlation operators are defined for multivariate functional data and identify the core characteristics that form the foundation for theoretical assumptions in sparse methods for multivariate functional data analysis. Building upon this theoretical framework, a novel penalty scheme is introduced for functional regression, which enables deriving superior asymptotic properties under more relaxed and reasonable assumptions. Furthermore, the non-asymptotic behavior of the proposed method is investigated under a finite-sample design, and non-asymptotic bounds are derived to demonstrate selection and estimation consistency. Simulation studies and a real-world application to human brain datasets highlight the effective performance of the method, showcasing its potential to enhance numerous penalized methods for functional data analysis.

E0646: **Robust statistical process monitoring of multivariate profiles**
Presenter: **Christian Capezza**, University of Naples Federico II, Italy
Co-authors: Fabio Centofanti, Antonio Lepore, Biagio Palumbo

In modern Industry 4.0 quality control applications, manufacturing processes generate large amounts of data, which often include outliers adversely affecting traditional control chart methods, especially in high-dimensional contexts. To face these challenges, the research introduces a novel framework, named robust multivariate functional control chart (RoMFCC), specifically designed to monitor multivariate functional quality characteristics while neutralizing the influence of both functional casewise and componentwise outliers. This innovation is important in multivariate profile monitoring, where outliers can affect an entire vector of functions or only a few components. The RoMFCC framework contains four main elements: a functional filter to detect functional component-wise outliers, a robust imputation of missing components in multivariate functional data, a robust dimension reduction that deals with functional case-wise outliers, and a procedure for prospective process monitoring. The RoMFCC's superior performance is assessed through a wide Monte Carlo simulation in comparison to competing monitoring schemes that have already appeared in the literature. The practical applicability of the RoMFCC is then demonstrated in monitoring a resistance spot welding process in automotive body-in-white manufacturing. The RoMFCC is implemented in the R package fun charts, which are available on CRAN.

E1017: **Sufficient dimension reduction for conditional quantiles for functional data**
Presenter: **Eliana Christou**, University of North Carolina at Charlotte, United States
Co-authors: Eftychia Solea, Shanshan Wang, Jun Song

Functional data analysis is an important research area with the potential to transform numerous fields. However, existing work predominantly relies on the more traditional mean regression methods, with surprisingly limited research focusing on quantile regression. Furthermore, the infinite-dimensional nature of the functional predictors necessitates the use of dimension-reduction techniques. Therefore, this gap is addressed by developing dimension-reduction techniques for the conditional quantiles of functional data. The convergence rates of the proposed estimators are derived, and their finite sample performance is demonstrated using simulation examples and a real dataset from fMRI studies.

E1081: **Motif discovery driven forecasting for functional data**
Presenter: **Jacopo Di Iorio**, Penn State University, United States

Forecasting has always been a major goal of functional data analysis, involving the prediction of future values and/or the evolution of functional observations. Given the increasing attention in the field of functional motif discovery, functional forecasting is performed through the identification of functional motifs. Functional motifs represent typical "shapes" or "patterns" recurring multiple times within a single curve and/or across misaligned portions of multiple curves. Portions characterized by the same motif are hypothesized to be more likely to evolve similarly. Extensive diagnostics can guide the user not only in tuning parameters but also in validating the aforementioned hypothesis, thus ensuring the applicability of the method. Method performance is assessed through simulations and is applied to a real-data case study.

EO015 Room 313 RECENT ADVANCES IN NONLINEAR TIME SERIES Chair: Guodong Li

E0217: **On buffered moving average models**
Presenter: **Philip Yu**, The Education University of Hong Kong, Hong Kong
Co-authors: Yipeng Zhuang, Dong Li, Wai-keung Li

There has been growing interest in extending the popular threshold time series models to include a buffer zone for regime transition. However, almost all attention has been paid to buffered autoregressive models. Note that the classical moving average (MA) model plays an equally important role as the autoregressive model in classical time series analysis. It is, therefore, natural to extend the investigation to the buffered MA (BMA) model. The focus is on the first-order BMA model while extending to a more general MA model should be direct in principle. The proposed model shares the piecewise linear structure of the threshold model but has a more flexible regime-switching mechanism. Its probabilistic structure is studied to some extent. A nonlinear least squares estimation procedure is proposed. Under some standard regularity conditions, the estimator is strongly consistent, and the estimator of the coefficients is asymptotically normal when the parameter of the boundary of the buffer zone is known. A portmanteau goodness-of-fit test is derived. Simulation results and empirical examples are carried out and lend further support to the usefulness of the BMA model and the asymptotic results.

E0337: **Functional threshold autoregressive model**
Presenter: **Yuanbo Li**, University of International Business and Economics, China
Co-authors: Kun Chen, Chun Yip Yau

A functional threshold autoregressive model is proposed for flexible functional time series modeling. In particular, the behavior of a function at a given time point can be described by different autoregressive mechanisms depending on the values of a threshold variable at a past time point. Sufficient conditions for the strict stationarity and ergodicity of the functional threshold autoregressive process are investigated. A novel criterion-based method is developed simultaneously, conducting dimension reduction and estimating the thresholds, autoregressive orders, and model parameters. The consistency and asymptotic distributions of the estimators of both thresholds and the underlying autoregressive models are also established. Simulation studies and an application to U.S. Treasury zero-coupon yield rates are provided to illustrate the effectiveness and usefulness of the proposed methodology.

E0503: **Uncover networks of R&D activities by a two-way constrained MAR model**
Presenter: **Xiaohang Wang**, Shenzhen Technology University, China
Co-authors: Philip Yu, Ling Xin
High-dimensional time series are traditionally monitored by vector time series models that have dimensionality and interpretation issues. Recently, an inspiring idea has been proposed to formulate the high dimensional data into a matrix or tensor time series structure when the variables have multi-way classifications. The new structures lead to a substantial dimensional reduction by ignoring the networks between classifications and focusing on within-class connections. It admits explicit interpretations of the within-class networks and inspires many applications. A matrix autoregressive model is formulated with two-way constraints (MAR-2C) to monitor R&D activities at the firm or regional level. To impose low-rank constraints on the network among different R&D activities and impose sparsity constraints on the network among different firms/regions, the model can highlight important information by the two matrix networks. A reduced-rank shrinkage-thresholding (RR-ST) algorithm is adopted in the model estimation, and a bootstrapping approach is used to make statistical inferences. The simulation results show that the RR-ST algorithm can achieve good accuracy. In real data analysis, R&D activities are monitored in 31 regions of China during the past 15 years by the MAR-2C model. Meaningful information like lead-lag relationships among different R&D activities, as well as the R&D spillovers among different regions, are uncovered from the estimated networks.

E0433: **Supervised factor modeling for high-dimensional linear time series**
Presenter: **Guodong Li**, University of Hong Kong, Hong Kong
Motivated by Tucker tensor decomposition, low-rank structures are imposed to the column and row spaces of coefficient matrices in a multivariate infinite-order vector autoregression (VAR), which leads to a supervised factor model with two-factor modelings being conducted to responses and predictors simultaneously. Interestingly, the stationarity condition implies an intrinsic weak group sparsity mechanism of infinite-order VAR, and hence, a rank-constrained group Lasso estimation is considered for high-dimensional linear time series. Its non-asymptotic properties are discussed thoughtfully by balancing the estimation, approximation and truncation errors. Moreover, an alternating gradient descent algorithm with thresholding is designed to search for high-dimensional estimates, and its theoretical justifications, including statistical and convergence analysis, are also provided. The theoretical and computational properties of the proposed methodology are verified by simulation experiments, and the advantages over existing methods are demonstrated by two real examples.

EO031 Room 405 NEW DEVELOPMENTS IN STATISTICAL INFERENCE FOR NON-GAUSSIAN DATA Chair: Zheng Wei

E0887: **Extending APP for skew normal distributions using Bonferroni method**
Presenter: **Ziyuan Wang**, University of Wisconsin Oshkosh, United States
Recently, researchers have become increasingly concerned with estimating the minimum sample sizes needed to provide good estimates of corresponding population parameters. The already large and ever-increasing literature on the a priori procedure (APP) is an outgrowth of this concern. APP equations and online calculators provide researchers with the ability to determine the minimum sample sizes needed to provide good estimates of corresponding population parameters. However, an APP limitation is that, until now, each advance concerned one population parameter at a time. The present contribution is the first to consider two parameters at once: locations and scales under skew normal populations, using the Bonferroni method. In addition to the underlying mathematical derivations, a link to an online calculator is provided that confers upon researchers the ability to determine the minimum sample size necessary to obtain good estimates of both locations and scales simultaneously.

E1077: **Probabilistic loss reserving prediction via denoising diffusion model**
Presenter: **Shiying Gao**, The University of Sydney, Australia
Co-authors: Boris Choy, Yuning Zhang, Junbin Gao, Ruikun Li
The aim is to propose a novel approach for predicting loss reserves within the insurance industry using a revised diffusion model. This approach considers the run-off triangles of claim data as graphical representations to elucidate connections among data points within the triangle. In contrast to the traditional cross-classified over-dispersed Poisson (ccODP) model, the diffusion model not only exhibits enhanced accuracy and efficiency but also provides probabilistic forecasts. Through a thorough analysis encompassing both simulation and empirical studies, the superior forecast accuracy of the diffusion model is demonstrated compared to existing methodologies. These results suggest that harnessing network-based interactions within run-off triangles holds promise for improving loss reserve forecasting.

E0879: **Robust Bayesian A/B testing**
Presenter: **Boris Choy**, University of Sydney, Australia
Co-authors: Xuan Li
Bayesian A/B testing has been widely adopted in industry practice. Most practitioners typically focus on large-scale inference from A/B testing and assume collected data follows a Gaussian sampling distribution by applying the central limit theorem (CLT). However, not every company can afford to have a large amount of data for analysis. In this case, the normality assumption held by CLT may not be applicable. To address this issue, the aim is to relax the normality assumption and provide robust Bayesian A/B testing models that can be used for both large and small datasets. By applying the heavy-tailed distributions under the Bayesian hierarchical framework, the proposed models stand out from other existing models on both large- and small-scale A/B testing.

E0890: **Bayesian stochastic frontier models under the skew normal settings**
Presenter: **Zheng Wei**, Texas A&M University, United States
Recently, a skew normal-based stochastic frontier model has emerged as a promising tool for efficiency analysis. A Bayesian framework for statistical inference is presented, incorporating both informative and non-informative prior knowledge. The efficacy of the Bayesian approach is evaluated through a rigorous examination using both simulation data and real data from a manufacturing productivity study. A comprehensive comparison with the conventional maximum likelihood approach is conducted. Results from both simulated and empirical investigations unequivocally demonstrate the superior performance of the Bayesian methodology.

EO229 Room 406 STATISTICAL LEARNING WITH APPLICATIONS Chair: Shengtong Han

E0434: **Subgroup detection based on partially linear additive individualized model with missing data in response**
Presenter: **Tingting Cai**, Capital Normal University, China

Based on a partially linear additive individualized model, a fusion-penalized inverse probability weighted least squares method is proposed to detect the subgroup for missing data in response. Firstly, the B-spline technique is used to approximate the unknown additive individualized functions, and then, an inverse probability weighted quadratic loss function is established with a fusion penalty on the difference in subject-wise B-spline coefficients. Secondly, minimization of such quadratic loss function leads to the estimation of linear regression parameters and individualized B spline coefficients. With a proper tuning parameter, some differences in penalty terms are shrunk into zero, and thus, the corresponding subjects will be clustered into the same subgroup. Thirdly, a clustering method is developed to automatically determine the subgroup membership for the subjects with missing data. Fourthly, large sample properties of resulting estimates are given under some regular conditions. Finally, numerical studies are presented to illustrate the performance of the proposed subgroup detection method.

E0485: **Confidence interval and hypothesis testing for high-dimensional quantile regression: Convolution smoothing and debiasing**
Presenter: **Yibo Yan**, East China Normal University, China

L1-penalized quantile regression (L1-QR) is a useful tool for modeling the relationship between input and output variables when detecting heterogeneous effects in a high-dimensional setting. Hypothesis tests can then be formulated based on the debiased L1-QR estimator that reduces the bias induced by the Lasso penalty. However, the non-smoothness of the quantile loss brings great challenges to the computation, especially when the data dimension is high. Recently, the convolution-type smoothed quantile regression (SQR) model has been proposed to overcome such shortcomings, and people developed the theory of estimation and variable selection therein. The debiased method is combined with the SQR model and comes up with the debiased L1-SQR estimator, based on which confidence intervals and hypothesis testing are then established in the high-dimensional setup. Theoretically, the non-asymptotic Bahadur representation is provided for the proposed estimator and also the Berry-Esseen bound, which implies the empirical coverage rates for the studentized confidence intervals. Furthermore, the theory of hypothesis testing is built on both a single variable and a group of variables. Finally, extensive numerical experiments are exhibited on both simulated and real data to demonstrate the method's good performance.

E0521: **A structural Mallows model for ranked data aggregation**
Presenter: **Han Li**, Shenzhen University, China

The rank aggregation problem is studied, which aims to find a consensus ranking by aggregating multiple ranking lists. To tackle the problem probabilistically, an elaborate ranking model is formulated by generalizing the traditional Mallows model. The original model assumes a uniform pair preference structure, which imposes a strict condition on the data. The attempt is to relax this condition and propose a new model that allows the pair preference to vary structurally. The model is quite flexible and has a closed-form expression for complete rankings as well as top-k rankings. Several useful theoretical properties of the model are investigated, and efficient algorithms are proposed to infer the model structure and parameters. Through extensive simulation studies and real applications, the new model is demonstrated to have satisfactory performance in different scenarios.

E0397: **Bayesian rare variant analysis identifies novel Schizophrenia putative risk genes from SCHEMA case control sample**
Presenter: **Shengtong Han**, Marquette University School of Dentistry, United States

The genetics of schizophrenia is so complex that it involves both common variants and rare variants. Rare variant association studies of schizophrenia are challenging because statistical methods for rare variant analysis are underpowered due to the rarity of rare variants. A recent schizophrenia exome meta-analysis (SCHEMA) consortium, the largest consortium to date, has successfully identified ten schizophrenia risk genes from ultra-rare variants by burden test. In contrast, more risk genes remain to be discovered by more powerful rare variant association test methods. A recently developed Bayesian rare variant association method is used, which is powerful for detecting sparse rare variants that implicate new candidate risk genes associated with schizophrenia from the SCHEMA case-control sample. These newly identified genes are significantly enriched in autism risk genes, and GO enrichment analysis indicates that new candidate risk genes are involved in mechanosensory behavior, regulation of cell size, neuron projection morphogenesis, and plasma membrane bounded cell projection morphogenesis, that may provide new insights on etiology of schizophrenia.

EO247 Room 408 SUFFICIENT DIMENSION REDUCTION (VIRTUAL) Chair: Kyongwon Kim

E0475: **On sparse directional regression**
Presenter: **Gayun Kwon**, Ewha Womans University, Korea, South

Sufficient dimension reduction has developed as a powerful tool to extract the core information hidden in high-dimensional data in the past decades. However, as most of the sufficient dimension reduction methods provide linear combinations of the original predictors, interpreting the extracted components becomes challenging. Sparse sufficient dimension reduction was introduced in a past study to facilitate simpler interpretation by producing sparse estimates. Sparse directional regression is introduced by extending the proposed methods of another study. To demonstrate the competitiveness of the method, the performance of sparse directional regression is compared with that of sparse sliced inverse regression and sliced average variance estimation through numerical experiments. The method is further applied to the large-scale wave energy farm dataset.

E0523: **On a SAVE and DR for large scale dataset**
Presenter: **Chaehyun Ryu**, Ewha Womans University, Korea, South

Directional regression is an effective dimension reduction approach for capturing inherent characteristics in regression problems. The idea of sliced average variance estimation and directional regression is extended to handle massive datasets. In particular, a "divide and conquer" strategy is adopted, breaking down the dataset into manageable chunks, and subsequently merging the results based on the proximity between dimension reduction subspaces. The capabilities of capturing distance is further harnessed to significantly enhance computational efficiency and optimize memory usage. The competitiveness of the approach is demonstrated through a comprehensive numerical study, and its application to a real-world dataset is demonstrated. In both simulation and application, R packages "foreach" and "bigmemory" are utilized for optimizing the execution speed and managing the memory when dealing with a massive dataset. The comparison between the proposed methodology, BIG-SAVE and BIG-DR, and the existing method, namely BIG-SIR, was conducted with a focus on computational speed and accuracy. The application of the methods to real datasets demonstrates its practical applicability and versatility.

E0609: **Multivariate response directional regression: An approach via projective resampling method**
Presenter: **Ah Reum Lee**, Ewha Womans University, Korea, South

Directional regression is a widely used tool to implement sufficient linear dimension reduction, which is useful for extracting core information from a high-dimensional dataset. The integration of a projective resampling method with sufficient dimension reduction, as outlined by another study, facilitates the recovery of dimension reduction subspaces in multivariate response regression. A novel approach is presented to multivariate response dimension reduction by combining projective resampling with directional regression. The proposed method is demonstrated to offer a

practical alternative to existing methods. Through both simulation studies and real data analysis, the method is illustrated to outperform other dimension reduction methods in a number of scenarios.

E0746: **A novel basis expansion for functional sliced inverse regression**
Presenter: **Harris Quach**, University of Pennsylvania, United States
Co-authors: Wensheng Guo, Wei Yang
An alternative basis expansion is considered for functional sliced inverse regression that leads to a novel estimator for the functional central subspace. The estimator provides some improvements over conventional functional sliced inverse regression in terms of simplicity of implementation and recovery of less smooth effective directions. Some theoretical results, numerical analyses and an application to the Chronic Renal Insufficiency Cohort study are provided.

EO202 Room 411 (Virtual sessions) RECENT DEVELOPMENT OF SPATIAL DATA AND TIME SERIES ANALYSIS Chair: Takaki Sato

E0177: **Grouped GEE for heterogeneous longitudinal data**
Presenter: **Tsubasa Ito**, Hokkaido University, Japan
Co-authors: Shonosuke Sugasawa
A generalized estimating equation is widely adopted for regression modelling for longitudinal data, taking account of potential correlations within the same subjects. However, since the standard GEE assumes common regression coefficients among all the subjects, such an assumption is not reasonable when there are potential heterogeneities in regression coefficients among subjects. Then, the method called grouped GEE analysis, which is more flexible and interpretable, is proposed to model longitudinal data by allowing heterogeneity in regression coefficients. The proposed method assumes that the subjects are divided into a finite number of groups and that subjects within the same group share the same regression coefficient. A simple algorithm for grouping subjects and estimating the regression coefficients simultaneously is proposed, and the asymptotic properties of the proposed estimator are shown. Finally, the finite sample performances of the proposed methods are demonstrated through simulation studies and real data analysis using health data.

E0344: **Powerful multiple test with a fixed sample size**
Presenter: **Masaki Toyoda**, Hitotsubashi University, Japan
Co-authors: Yoshimasa Uematsu
Modern data analyses often encounter challenges due to very small sample sizes despite high dimensionality. The focus is on statistical inference in such a situation, where the number of hypotheses is very large relative to a limited sample size, and a novel method of FDR-controlled multiple test is proposed. The key idea is to use an accumulation test and data fission, which have recently been developed in the literature. Remarkably, it is shown that the power can tend to unity as the number of hypotheses increases, even though the sample size is fixed. The validity of the method is also confirmed by numerical experiments and a real data analysis.

E0530: **GMM estimation of spatial autoregressive models with cluster-dependent errors**
Presenter: **Takaki Sato**, Musashi University, Japan
The generalized method of moment (GMM) estimation of spatial autoregressive (SAR) models is considered, with unknown cluster correlations among error terms. In the presence of cluster correlations within errors, nonlinear moment conditions suitable for independent errors lose their validity, and GMM estimators obtained from the moment conditions are inconsistent. A GMM estimator obtained from another nonlinear moment condition is proposed, suitable for cluster-dependent error terms, and its asymptotic properties are shown. Because the asymptotic variance of the GMM estimator depends on the choice of a weight matrix for GMM estimation, an optimal weight is also discussed that minimizes the asymptotic variance, and a feasible optimal GMM estimator is introduced based on a consistent estimator of the weight. Monte Carlo experiments indicate that the proposed GMM estimator has a small bias and root mean squared errors when error terms have cluster correlation compared to two-stage least squares estimators and GMM estimators for independent errors.

E0806: **Likelihood-based analysis of general Gaussian processes having scaling properties**
Presenter: **Tetsuya Takabatake**, Hiroshima University, Japan
Recent studies in mathematical finance and financial econometrics suggest that properties of the roughness of the sample path and the persistency of the auto-covariance function would be important factors in constructing better forecasting models of the volatility of asset prices. The log-volatility process is modeled as a general Gaussian process having scaling properties that capture the roughness and long-memory properties simultaneously and then discuss asymptotic properties of likelihood-based estimators for the log-volatility.

EC270 Room 111 SURVIVAL ANALYSIS Chair: Matias Quiroz

E0850: **Accelerated failure time model under dependent truncated data**
Presenter: **Jin-Jian Hsieh**, Department of Mathematics, National Chung Cheng University, Taiwan
The purpose is to delve into the accelerated failure time model within the framework of dependent truncation data and leverage the copula model to establish correlations within the dataset. Building upon a past work that utilized the copula-graphic method to estimate survival functions and proposed an approach for estimating correlation parameters, the methodology is further extended by introducing two distinct estimation techniques for regression parameters. The first method involves parameter evaluation through the calculation of the area between survival curves, while the second method employs the weight of survival jump in conjunction with the least squares approach to estimate regression parameters. The efficacy of these proposed estimation procedures is evaluated through simulation studies, and a comparative analysis is conducted between the two approaches. Furthermore, these methodologies are applied to two real-world datasets, providing insights into their practical applicability. Through this analysis, a deeper understanding of how these approaches can be effectively utilized in real-world scenarios is gained.

E0976: **A presmoothed estimator for the cure probability: An application to a cardiotoxicity dataset**
Presenter: **Ana Lopez-Cheda**, University of A Coruna, Spain
Co-authors: Samuel Saavedra, M Amalia Jacome
Current cancer treatments have caused an increased ratio of cured patients or, at least, long-term survival. To accommodate the insusceptible proportion of subjects, a cure fraction can be explicitly incorporated into survival models, and as a consequence, cure models will arise. The goals in cure models are usually to estimate the cure rate and the probability of survival of the uncured patients up to a given point in time (latency). Although, in the literature, parametric and semiparametric models have been considered, nonparametric estimation methods for cure models have attracted much attention in the last few years. A presmoothed nonparametric estimator is proposed for the probability of cure in mixture cure models. Specifically, the methodology in a prior study is considered to improve the cure rate estimator of a subsequent study. The introduced nonparametric estimator is compared with existing approaches in a simulation study. Finally, the proposed methodology is applied to a study of cardiotoxicity in breast cancer patients.

E0979: **High dimensional single-index mixture cure models in cardio-oncology**

Presenter: **Beatriz Pineiro-Lamas**, Universidade da Coruna, Spain

Co-authors: Ana Lopez-Cheda, Ricardo Cao

In survival analysis, there are situations in which not all subjects are susceptible to the final event. For example, if the event is a cancer therapy-related adverse effect, there will be a fraction of patients (considered as cured) that will never experience it. Mixture cure models allow to estimate the probability of cure and the survival function for the uncured subjects. In the literature, nonparametric estimation of both functions is limited to continuous univariate covariates. This important gap is filled by proposing single-index mixture cure models. They allow working with a vector covariate and assume that the survival function depends on it through an unknown linear combination, that can be estimated by maximum likelihood. The proposed models are extended to functional covariates and a preprocessing algorithm is implemented to deal with medical images. The methodology is applied to a cardiotoxicity dataset. The goal is to determine whether (and how) certain factors affect the probability of experiencing the cardiovascular problem and the amount of time it takes for it to manifest. Understanding risk factors may lead to a personalized preventive medicine.

E1045: **Testing covariate effects in the mixture cure model using distance correlation**

Presenter: **Maria Amalia Jacome Pumar**, Universidade da Coruna, Spain

Co-authors: Blanca Estela Monroy-Castillo, Ricardo Cao

One of the goals of cure models is to test whether a covariate influences the cure rate. Distance correlation is a novel class of multivariate dependence coefficients with advantages over classical correlation coefficients: it is applicable to random vectors of arbitrary dimensions not necessarily equal, and it is zero if and only if the vectors are independent. Distance correlation has been applied in a standard survival model without cure based on the distance covariance between covariates and the survival times. But to the best of knowledge, distance correlation has not been applied yet in the presence of a cure fraction. A method is proposed to study the effect of a covariate on the probability of cure by means of the distance correlation with a procedure that overcomes the challenge of handling the missingness of the cure indicator.

Wednesday 17.07.2024 16:40 - 18:20 Parallel Session E – EcoSta2024

EI008 Room 106 RECENT DEVELOPMENT BASED ON STOCHASTIC PROCESSES Chair: Catherine Liu

E0180: **Functional principal component analysis of spatially and temporally indexed point processes**
Presenter: **Yehua Li**, University of California at Riverside, United States
Spatially and temporally indexed point process data is modeled as a multi-level log-Gaussian Cox process where the log intensity function depends on a partially linear single-index structure of spatio-temporal covariates and three latent functional random effects representing the spatial and temporal random effects as well as their interactions. It is assumed that the latent functional effects are Gaussian processes with Karhunen-Loeve representations, and the unknown link function of the single-index as well as the covariance functions of the latent functional effects as splines, are modeled. The proposal is to estimate the partially linear coefficients and the single-index link function using a Poisson maximum likelihood method and the covariance functions of the latent processes using maximum composite likelihood methods. Approaches to predict the functional principal component scores are also proposed. Under the multi-level dependence structure and allowing the spatiotemporal covariates to be non-stationary, the proposed estimators follow rather unconventional convergence rates, which depend on both the number of locations and the number of repeated measures in time. The proposed method is illustrated through a simulation study and a real-data application in modeling bike-sharing events.

E0152: **Jump-size-based Bayesian detection of multiple change-points**
Presenter: **Catherine Liu**, The Hong Kong Polytechnic University, Hong Kong
An original and general NOn-SEgmental (NOSE) approach is proposed for the detection of multiple change-points. NOSE identifies change-points by the non-negligibility of posterior estimates of the jump heights. Specifically, under the Bayesian paradigm, NOSE treats the step-wise signal as a global infinite dimensional parameter drawn from a proposed process of truncated atomic representation. The random jump heights are further modeled by a Gamma-Indian buffet process shrinkage prior under the form of discrete spike-and-slab. Under the mean-shifted model, the proposed prior elicitation successfully achieves the minimax optimal posterior contraction rate regarding prediction loss. For a bounded number of change-points, NOSE enjoys a lower localization error compared with existing approaches, even in addressing more difficult problems that have a lower signal-to-noise ratio. NOSE is applicable and effective in detecting scale shifts, mean shifts, and structural changes in regression coefficients under linear or autoregression models. Comprehensive simulations and several real-world examples demonstrate the superiority of NOSE in detecting abrupt changes under various data settings. Next, we introduce SBPCPM, an extension of the NOSE approach, which deploys a hypothesis test rather than tackle sparsity and discovers new change-points in a dataset of the London House Index between 2000 and 2022.

E0179: **Bayesian analysis of nonlinear structured latent factor models using a Gaussian process prior**
Presenter: **Jian Qing Shi**, SUSTECH, China
Co-authors: Yimang Zhang, Xiaorui Wang
Factor analysis models are widely utilized in social and behavioral sciences, such as psychology, education, and marketing, to measure unobservable latent traits. A nonlinear structured factor analysis (FA) model is introduced, which is more flexible in characterizing the relationship between manifest variables and latent factors, and then the confirmatory identifiability of the latent factor is given to ensure the substantive interpretation of the latent factors. A Bayesian approach with a Gaussian process prior is proposed to estimate the unknown nonlinear function. Asymptotic results are established, including structural identifiability of the latent factors, consistency of all parameters and the unknown nonlinear function. Simulation studies and real data analysis are conducted to investigate the performance of the proposed method. Simulation studies and real data analysis show the proposed method performs well in handling nonlinear model and successfully identifies the latent factors.

EO257 Room 102 STRUCTURAL MACROECONOMETRICS Chair: Roberto Leon-Gonzalez

E0313: **Estimating medium-scale new Keynesian model under the zero lower bound for Japan**
Presenter: **Hirokuni Iiboshi**, Nihon University, Japan
The Japanese economy has been caught in a liquidity trap where the nominal interest rate is on the zero lower bound (ZLB) since the late 1990s and has experienced prolonged deflation and stagnation. The impact of the ZLB on deflation and stagnation in Japan is examined by estimating a medium-scale New Keynesian model with the ZLB. To this end, OccBin and inversion filter are incorporated into the SMC with model tempering proposed by a recent study and efficient Bayesian inference of the model from 1980 to 2016, before the negative interest rate policy. From these estimates, estimates of structural shocks are calculated as historical decompositions and impulse response functions. In addition, by computing counterfactuals that do not impose the ZLB constraint, the extent to which the ZLB constraint contributes to the deterioration of the economy is quantified.

E0533: **Posterior inferences on incomplete structural models: The minimal econometric interpretation**
Presenter: **Takashi Kano**, Hitotsubashi University, Japan
The minimal econometric interpretation (MEI) of DSGE models provides a formal model evaluation and comparison of misspecified nonlinear dynamic stochastic general equilibrium (DSGE) models based on atheoretical reference models. The MEI approach recognizes DSGE models as incomplete econometric tools that provide only prior distributions of targeted population moments but have no implications for actual data and sample moments. The purpose is to develop a Bayesian posterior inference method based on the MEI approach. Prior distributions of targeted population moments simulated by the DSGE model restrict the hyperparameters of Dirichlet distributions. These are natural conjugate priors for multinomial distributions followed by corresponding posterior distributions estimated by the reference model. The Polya marginal likelihood of the resulting restricted Dirichlet-multinomial model has a tractive approximated log-linear representation of the Jensen-Shannon divergence, which the proposed distribution-matching posterior inference uses as the limited information likelihood function. Monte Carlo experiments indicate that the MEI posterior sampler correctly infers calibrated structural parameters of an equilibrium asset pricing model and detects the true model with posterior odds ratios.

E0508: **Accurate marginal likelihood estimation for point, weakly and partially identified models**
Presenter: **Nianling Wang**, Capital University of Economics and Business, China
Marginal likelihood is an important quantity that evaluates the fitting performance of a model to data and defines the Bayes factor for model comparison in Bayesian econometrics. However, an analytical form of marginal likelihood is often unavailable, and one can only resort to numerical estimation. What's worse, the marginal likelihood estimation for weakly and partially identified models is additionally subject to parameter identification problems. A general and accurate approach is proposed to estimating marginal likelihood. It is general in the sense that it is suitable for models where parameters can be point, weakly or partially identified. It can give an accurate estimation of marginal likelihood in a finite sample size. Particularly, a weighted power posterior is defined with an optimized reference distribution, and then the marginal likelihood is evaluated based on a series of weighted power posteriors. Three examples, including the linear regression model, DSGE model and entry game model, are used to illustrate and check the performance of the proposed approach.

E0317: **Likelihood based estimation of nonlinear dynamic stochastic general equilibrium models**
Presenter: **Roberto Leon-Gonzalez**, GRIPS, Japan
Co-authors: Elnura Baiaman

A new likelihood-based approach is proposed using perturbation methods to estimate nonlinear DSGE models. A nonlinear approximation is implicitly used for the policy function that is invertible with respect to the shocks, implying that shocks can be recovered uniquely from some of the control variables in the approximation. Based on this approximation, the likelihood can then be obtained by using a standard change of variables theorem and a Lagrange inversion formula. This technique is implemented to estimate the DSGE model. In contrast with previous likelihood-based approaches, this method allows for unobserved non-stochastic state variables and requires neither additional shocks nor simulation to evaluate the likelihood. Using US data, the proposed approach is demonstrated to the well-known neoclassical growth model of a prior study. In addition to the baseline model, versions of the model are also considered in which the structural shocks have time-varying variances. It is found that a nonlinear heteroscedastic model has much better empirical performance. It is a much better fit for the observed data than the linearized model. In addition, the monetary policy shock is found to primarily drive the time changes in the uncertainty in the economy.

EO105 Room 103 ADVANCEMENTS IN LATENT VARIABLE MODELING Chair: Shiyu Wang

E0332: **Dynamic cognitive diagnostic frameworks: A general model for learning**
Presenter: **Zichu Liu**, Beijing Normal University, China
Co-authors: Shiyu Wang, Shumei Zhang, Tao Qiu, Houping Xiao

In education, understanding students' learning trajectories is essential for educators to monitor and enhance their progress effectively. With the advent and widespread use of computer-based testing, researchers now have access to rich and varied datasets that offer deeper insights into student performance. A novel general dynamic cognitive diagnostic model that integrates response accuracy and response times is introduced. The aim is to distinguish between different learning and testing behaviors, allowing for the estimation of students' learning trajectories concerning their proficiency levels in various assessed skills over time. Comprising two key components, a dynamic transition model assessing the transition probabilities of students' attributes and a mixture fluency model evaluating students' attribute profiles, the proposed model is rigorously validated through extensive simulation studies. These studies demonstrate the model's efficacy in providing valuable insights into students' learning trajectories. Furthermore, the proposed model is applied to a real dataset derived from a spatial rotation diagnostic test, further showcasing its practical utility in educational settings. Through its comprehensive approach and rigorous validation, this model emerges as a valuable tool for educators and researchers alike, offering nuanced insights into students' learning progress and behavior dynamics.

E0381: **Calibrating item response theory models with sparse data**
Presenter: **Shiyu Wang**, University of Georgia, United States
Co-authors: Yuan Ke, Cong Cheng

A new statistical methodology framework is presented, tailored to address the intricate challenge of calibrating item response theory (IRT) models under conditions of limited samples and sparse response data. The approach introduces an innovative item parameter estimation method that integrates change point detection techniques, aiming to enhance the robustness and accuracy of IRT model calibration in resource-constrained settings. Situated within the realm of statistical and machine learning methodologies, the proposed approach endeavours to distil valuable insights by unveiling lower-dimensional patterns inherent in the data. To evaluate the effectiveness of our proposed techniques, a series of simulation studies are conducted designed to mimic various characteristics of small-scale assessments. These simulations serve to validate the performance and robustness of our methodology across a range of scenarios commonly encountered in practical applications. Additionally, an in-depth analysis is conducted utilizing real-world data derived from a computer-based classroom assessment, providing empirical evidence of the efficacy and applicability of the approach in real-world educational settings. The outcomes of this research project hold significant promise in advancing the application of IRT in the context of small-scale assessments.

E0412: **Examining the usage of rating scale in subjective creativity assessments through partial credit model**
Presenter: **Sujie Yang**, University of Science and Technology of China, China
Co-authors: Jue Wang

The evaluation of subjective creativity assessments is challenging due to the involvement of raters and a lack of scoring criteria. Raters' cognitive process is viewed as a black box. A psychometric framework is introduced for examining subjective creativity assessments along with a set of guidelines for evaluating the usage of rating scales. Within this framework, the focus is on the selection of an appropriate measurement model for analyzing ratings that reflect quite different judgment decisions among raters. In particular, the use of a modified partial-credit Rasch model is proposed with a rater facet that allows for the estimation of a unique threshold structure for each rater. This model can also be viewed as a reduced version of the many-facet Rasch model that provides more information regarding the unique category usages by raters and examination of various rater effects. An empirical data analysis is conducted using rating scores obtained from a subjective creativity assessment based on science tasks. Results indicate very different usages of rating scales among raters and reveal certain effects on individual raters. Discussions on the uses of the partial credit model and implications for improving the rating procedure within the context of subjective creativity assessments are provided.

E0459: **Willing and able to fake: A new and flexible item response model for applicant faking**
Presenter: **Siwei Peng**, Jiangxi Normal University, China
Co-authors: Yan Cai, Dongbo Tu

Applicant faking (AF), the intentional distortion of responses on non-cognitive assessments to present oneself with a false self-image, poses a critical concern in human resource management by increasing the risk of selecting non-ideal candidates. A new psychometric model is introduced for faking behavior, termed AF-IRT, which translates existing faking theory into a quantitative model. The AF-IRT identifies faking behavior on the item-by-respondent level and decomposes the faking process into (a) applicant's willingness to fake and (b) their perceptions of desirable response options. It helps explore person and item characteristics associated with higher prevalence of faking behavior and examine which response categories are more desirable for a specific item. The simulation studies demonstrated that the proposed AF-IRT model exhibited reasonable parameter recovery. The AF-IRT exhibited better estimation accuracy and higher reliability than the model that ignores faking. Empirical findings further supported the practical advantages of the AF-IRT. To improve clarity and accessibility, a step-by-step tutorial has been included to assist novice or non-quantitative researchers in utilizing the AF-IRT to analyze their empirical data through the R language.

EO107 Room 104 RECENT DEVELOPMENTS IN NETWORK MODELING AND APPLICATIONS Chair: Dongxue Zhang

E0464: **High-dimensional-responses-assisted heterogeneous nodal influence analysis**
Presenter: **Dongxue Zhang**, Southwestern University of Finance and Economics, China

An $m \times n$ matrix network data is considered with m network nodes and n-dimensional responses for each node, where both m and n can diverge to infinity. The heterogeneity of network nodal influence is addressed by different influence parameters of each node, which are expressed through high-dimensional responses using a specific link function. By allowing heterogeneous error variances, a response-assisted network influence model is proposed to integrate information on the matrix response variable and network structures across both m network nodes and n dimensions of responses. Since the traditional maximum likelihood estimation method is invalid in this case, an optimal generalized method is built on the moment's estimation method to avoid estimating unknown error variances by restricting the diagonal of the weighting matrix in quadratic moments. The consistency and asymptotic normality of the estimator are established. In addition, a homogeneity test has also been developed to examine the influence of heterogeneity. Extensive simulation studies and an empirical study of fund and stock matrix network data are presented to demonstrate the usefulness of the proposed model.

E0468: **A Gaussian mixture model for multiple instance learning with partially subsampled instances**
Presenter: **Baichen Yu**, Peking University, China
Co-authors: Xuetong Li, Jing Zhou, Hansheng Wang

Multiple instance learning is a powerful machine learning technique, which is found useful when numerous instances can be naturally grouped into different bags. Accordingly, a bag-level label can be created for each bag according to whether the instances contained in the bag are all negative or not. Thereafter, how to train a statistical model with bag-level labels with/without partially labelled instances becomes a problem of great interest. To this end, a Gaussian mixture model (GMM) framework is developed to describe the stochastic behavior of the instance-level feature vectors. Both the instance-based maximum likelihood estimator (IMLE) and the bag-based maximum likelihood estimator (BMLE) are theoretically investigated. It is found that the statistical efficiency of the IMLE could be much better than that of the BMLE, if the instance-level labels are relatively hard to be predicted. To fix the problem, a subsampling-based maximum likelihood estimation (SMLE) approach is developed, where the instance-level labels are partially provided through careful subsampling. This leads to a significantly reduced labeling cost with little sacrifice in terms of statistical efficiency. Extensive simulation studies are presented to demonstrate the finite sample performance. A real data example using whole-slide images (WSIs) to diagnose metastatic breast cancer is illustrated.

E0471: **Network varying coefficient model**
Presenter: **Xinyan Fan**, Renmin University of China, China
Co-authors: Wei Lan, Kuangnan Fang

A novel network varying coefficient model (NVCM) is proposed that extends traditional varying coefficient models (VCM) to accommodate network data. The key idea is to model the regression coefficients as functions of the latent locations of network nodes that drive the formation of the network. To estimate the model, the latent locations are identified via the latent space model, and an iterative projected gradient descent algorithm is developed by maximizing the network parameters and regression coefficients alternately. The non-asymptotic bounds of the estimated coefficients matrix are obtained theoretically. Practically, the dimension of the latent space is chosen via a Bayesian information criterion (BIC)-type criterion. The method is further combined with a penalization procedure to select covariates with varying coefficients, as well as those that are significant to the response variable and derive the related theoretical properties. The utility of the model is further illustrated via simulation studies as well as a real-world application in the field of finance by analyzing the relationship between stock returns and firm characteristics from a network perspective. The results show that the proposed model outperforms most existing methods.

E0665: **Interpret how external shocks affect industrial chain using graph machine learning**
Presenter: **Bin Liu**, Southwestern University of Finance and Economics, China

A quantitative analysis of the development of the industry chain is conducted from the perspective of external shocks. Factors that may impact the performance of the industrial chain have been studied in the literature, such as government regulation, monetary policy, etc. The interest lies in how to quantify the impacts of these shocks on the industrial chain's performance. To achieve this goal, the industrial chain is modeled with a graph neural network (GNN) and node regression is conducted on some financial performance metrics. To capture the effects of external shocks, it is proposed to compute the interaction between shocks and industrial chain features with a cross-attention module and then filter the original node features in the graph regression. Experiments on two real datasets demonstrate that (i) there are significant effects of external shocks on the industrial chain, and (ii) model parameters, including regression coefficients and the attention map, can explain how external shocks affect the performance of the industrial chain.

EO196 Room 105 DATA TWINS: EFFICIENT DATA COLLECTION AND EFFECTIVE DATA ANALYSIS Chair: Xiaoning Kang

E0818: **An efficient approach for identifying important biomarkers for biomedical diagnosis**
Presenter: **Frederick Kin Hing Phoa**, Academia Sinica, Taiwan
Co-authors: Jing-Wen Huang, Yan-Hong Chen, Yan-Han Lin, Shau Ping Lin

The challenges associated with biomarker identification are explored for diagnosis purposes in biomedical experiments, and a novel approach is proposed to handle the above challenging scenario via the generalization of the Dantzig selector. To improve the efficiency of the regularization method, a transformation from an inherent nonlinear programming due to its nonlinear link function is introduced into a linear programming framework. The use of the method is illustrated in an experiment with binary response, showing superior performance on biomarker identification studies when compared to their conventional analysis. The proposed method does not merely serve as a variable/biomarker selection tool; its ranking of variable importance provides valuable reference information for practitioners to reach informed decisions regarding the prioritization of factors for further investigations.

E0840: **Change-point detection in generalized extreme value distribution via generalized fiducial inference**
Presenter: **Xia Cai**, Hebei University of Science and Technology, China

Generalized extreme value (GEV) distribution is used to analyze the maximum from a block of data. It is very useful to describe the unusual event rather than the usual event. A change-point detection procedure for GEV distribution based on generalized fiducial inference is proposed. The fiducial distribution of the change-point location is constructed. Meanwhile, Markov Chain Monte Carlo method combined with Gibbs sampling and the Metropolis-Hastings algorithm is utilized to estimate the location of the change point and its confidence interval. Simulation results show that the proposed generalized fiducial method performs better in accuracy, robustness and the length of confidence intervals. Finally, the proposed method is applied to the annual maximum rainfall data in Beijing.

E1084: **A discrepancy decomposition model for calibration experiments**
Presenter: **Shifeng Xiong**, Chinese Academy of Sciences, China

Computer simulation models are widely used to study complex physical systems. A related topic is the calibration problem, which aims at learning about the values of parameters in the model based on observations. In most real applications, the parameters have specific physical meanings and are called physical parameters. To understand the true underlying physical system, effectively estimating such parameters is needed. However, existing calibration methods have limitations in addressing this due to the model identifiability issue. The aim is to propose a method based on the discrepancy decomposition model to describe the discrepancy between the physical system and the computer model. The proposed model possesses a clear interpretation, and more importantly, it is identifiable under mild conditions. Under this model, estimators of the physical parameters and the discrepancy functions are presented, and their asymptotic properties are established. Numerical examples show that the proposed method is capable of accurately estimating the physical parameters.

E1085: **Risk-adjusted monitoring of online user-generated reviews via user preference learning**
Presenter: **Qiao Liang**, Southwestern University of Finance and Economics, China

Online customer reviews provide valuable information about product quality, and recently, some control chart-based schemes have been proposed to detect product performance anomalies from reviews. As a review outcome depends not only on inherent product quality but also on customer rating bias and latent preference, ignoring customers' latent factors may lead to misjudgments about online product performance. Therefore, a risk-adjusted control chart is proposed for monitoring the decrease in review rating scores by separating the personal risk factors of individual customers from the assignable causes with respect to online product performance. The proposed risk-adjustment model is fitted by a united latent factor model that learns user rating bias and preference factors by combining both review texts and corresponding ratings. According to the experimental results of a real-world case and extensive simulation studies, the proposed method shows superior performance in review shift detection, with good interpretation for explaining the reasons behind anomalies.

EO187 Room 109 STATISTICAL LEARNING METHODS FOR COMPLEX BIOMEDICAL DATA ANALYSIS Chair: Jian Kang

E0437: **Positive-definite regularized estimation for high dimensional covariance on scalar regression**
Presenter: **Jie He**, Nanjing University of Aeronautics and Astronautics, China

Covariance is an important measure of marginal dependence among variables. However, heterogeneity in subject covariances and regression models for high-dimensional covariance matrices has not been well studied. Compared to regression analysis for conditional means, modeling high-dimensional covariances is much more challenging due to the large set of free parameters and the intrinsic positive-definite property that puts constraints on the regression parameters. A regularized estimation method is proposed for the regression coefficients of covariances under sufficient and necessary constraints for the positive definiteness of the conditional average covariance matrices given covariates. The proposed estimator satisfies the sparsity and positive-definite properties simultaneously. An alternation direction method of multipliers (ADMM) algorithm is proposed to solve the constrained and regularized optimization problem. The convergence of the proposed ADMM algorithm is shown, and the convergence rates of the proposed estimators are derived for the regression coefficients and the heterogeneous covariances. The proposed method is evaluated by simulation studies, and its practical application is demonstrated by a case study on brain connectivity.

E0671: **Learning context-aware distributed gene representations in spatial transcriptomics with SpaCEX**
Presenter: **Hao Wu**, Shenzhen Institute of Advanced Technology, China

Distributed gene representations are pivotal in genomic research, offering a means to understand the complexities of genomic data and providing the foundation for various data analysis tasks. Current gene representation learning methods demand costly pretraining on heterogeneous transcriptomic corpora, making them less approachable and prone to over-generalization. For spatial transcriptomics (ST), there are many methods for learning spot embeddings but lacking methods for generating gene embeddings from spatial gene profiles. To fill the gap, SpaCEX is presented, a pioneer self-supervised learning model that generates context-aware, semantically rich gene embeddings (SpaCEX-generated-Gene-Embeddings, SGEs) from ST data through exploiting spatial genomic context (SGC) identified as spatially co-expressed gene modules. As a few-shot learning method focusing on targeted single datasets, SpaCEX is cost-effective, context-sensitive, and robust to cross-sample technical artefacts. Real data analyses reveal the biological relevance of SpaCEX-identified SGC and affirm the functional and relational semantics of SGEs. Based on the SGEs, novel computational methods are developed for key downstream objectives: identifying disease-associated genes and gene-gene interactions, enhancing transcriptomic coverage of FISH-based ST, detecting spatially variable genes, and enhancing spatial clustering. Extensive real data results demonstrate these methods' superior performance.

E0726: **Bayesian scalar-on-image regression with the spatially varying neural network prior**
Presenter: **Ben Wu**, Renmin University of China, China
Co-authors: Keru Wu, Jian Kang

Deep neural networks (DNN) have been adopted in the scalar-on-image regression, which predicts the outcome variable using image predictors. However, training DNN often requires a large sample size to achieve good prediction accuracy, and the model-fitting results can be difficult to interpret. A novel Bayesian non-linear scalar-on-image regression framework is proposed with a spatially varying neural network (SV-NN) prior. The SV-NN is constructed using a single hidden layer neural network with weights generated by the soft-thresholded Gaussian process. The framework is able to select interpretable image regions and to achieve high prediction accuracy with limited training samples. The SV-NN provides large prior support for the imaging effect function, enabling efficient posterior inference on image region selection and automatically determining the network structures. The posterior consistency of model parameters and selection consistency of image regions is established when the number of voxels/pixels grows much faster than the sample size. An efficient posterior computation algorithm is developed based on stochastic gradient Langevin dynamics (SGLD). The methods are compared with state-of-the-art deep learning methods via analyses of multiple real data sets, including task fMRI data from the Adolescent Brain Cognitive Development (ABCD) study.

E0927: **Spectral co-clustering in multi-layer Directed networks**
Presenter: **Wenqing Su**, Tsinghua University, China
Co-authors: Ying Yang

Multilayer network data analysis is one of the research hotspots in statistics and biomedical. Current literature on multilayer network data is mostly limited to undirected relations. However, direct relations are more common and may introduce extra information. The focus is on community detection in multilayer-directed networks. To account for the asymmetry, a novel spectral-co-clustering-based algorithm is developed to detect co-clusters, which capture the sending patterns and receiving patterns of nodes, respectively. Specifically, the eigen-decomposition of the debiased sum of Gram matrices over the layer-wise adjacency matrices is computed, followed by the k-means, where the sum of Gram matrices is used to avoid possible cancellation of clusters caused by direct summation. Theoretical analysis of the misclassification rates is derived, which shows that multilayers would benefit clustering performance. The experimental results of simulated data corroborate the theoretical predictions, and the analysis of a real-world trade network dataset provides interpretable results.

EO094 Room 110 DESIGN AND ANALYSIS OF COMPUTER EXPERIMENTS Chair: Wenlong Li

E0295: **Construction of orthogonal maximin distance designs**
Presenter: **Wenlong Li**, Beijing Jiaotong University, China
Co-authors: Yubin Tian, Min-Qian Liu

Maximin distance designs and orthogonal designs are two attractive classes of space-filling designs for computer experiments, but their theoretical constructions are challenging, especially the construction of optimal designs in terms of both the maximin distance and orthogonality criteria. A systematic method is presented for constructing orthogonal maximin distance designs with flexible numbers of runs and factors. The method is carried out by rotating the subarrays of a saturated two-level regular design in Yates order or its circular shifting version. The principal objective is to construct high-level designs from two-level designs, and the method is effective because the performance of high-level designs is determined by that of two-level designs under both the maximin distance and orthogonality criteria. The proposed method is also generalized by rotating the subarrays of a saturated two-level nonregular design such that the resulting designs have flexible run sizes. Comparison results reveal that the resulting orthogonal designs are well worthy of recommendation under the maximin distance criterion. An illustrative example is provided to show that the proposed designs have a good two-dimensional stratification property. An application is given to present the effectiveness of the proposed designs in building statistical surrogate models.

E0301: **Feature calibration for computer models**
Presenter: **Wenzhe Xu**, Beijing University of Posts and Telecommunications, China

Computer model calibration involves using partial and imperfect observations of the real world to learn which values of a models input parameters lead to outputs that are consistent with real-world observations. When calibrating models with high dimensional output (e.g. a spatial field), it is common to represent the output as a linear combination of a small set of basis vectors. Often, when trying to calibrate to such output, what is important to the credibility of the model is that key emergent physical phenomena are represented, even if not faithfully or in the right place. In these cases, a comparison of model output and data in a linear subspace is inappropriate and will usually lead to poor model calibration. To overcome this, kernel-based history matching (KHM) is presented, generalizing the meaning of the technique sufficiently to be able to project model outputs and observations into a higher-dimensional feature space, where patterns can be compared without their location necessarily being fixed. The technical methodology is developed, presenting an expert-driven kernel selection algorithm, and then the techniques are applied to the calibration of boundary layer clouds for the French climate model IPSL-CM.

E0537: **SIGMA: Stochastic differential equations informed Gaussian process model for parameter inference**
Presenter: **Zhaohui Li**, Academy of Mathematics and Systems Science,Chinese Academy of Sciences, China

Stochastic differential equations (SDEs) have been extensively applied in diverse fields such as systems biology, pandemic, financial engineering, physics, etc. However, due to their inherent stochasticity and the complexity of the underlying dynamic systems, parameter estimation and uncertainty quantification for SDEs pose significant challenges. A novel method called the SDE-informed Gaussian process model for parameter inference (SIGMA), is proposed to address these challenges. The new method employs a nonstationary Gaussian process (GP) model to approximate the solutions of SDEs. The approach incorporates a nonparametric mean function and a parametric variance function, providing flexibility in the approximation process. The Matern 1/2 kernel is employed for the GP prior, as it yields non-differentiable sample paths that resemble those of SDEs. A Kullback-Leibler (KL) divergence is designed as a metric to quantify the discrepancy between the GP and the SDE. SIGMA adopts a Bayesian paradigm to incorporate the KL divergence into the posterior density, which enables thorough uncertainty quantification for both the parameters and the unobserved SDE solutions.

E0897: **Sequential Latin hypercube design for two-layer computer simulators**
Presenter: **Yan Wang**, Beijing university of technology, China

The two-layer computer simulators are commonly used to mimic multi-physics phenomena or systems. Usually, the outputs of the first-layer simulator (also called the inner simulator) are partial inputs of the second-layer simulator (also called the outer simulator). How to design experiments by simultaneously considering the space-filling properties of inner and outer simulators is a significant challenge that has received scant attention in the literature. To address this problem, a new sequential optimal Latin hypercube design (LHD) is proposed by using the maximin integrating mixed distance criterion. A corresponding sequential algorithm for efficiently generating such designs has also been developed. Numerical simulation results show that the new method can effectively improve the space-filling property of the outer computer inputs. The case study about composite structures assembly simulation demonstrates that the proposed method can outperform the benchmark methods.

EO089 Room 111 NOVEL STATISTICAL MODELS AND METHODS WITH APPLICATIONS Chair: Yingying Ma

E0161: **Penalized sparse covariance regression with high dimensional covariates**
Presenter: **Yuan Gao**, Peking University, China
Co-authors: Zhiyuan Zhang, Xuening Zhu, Zhanrui Cai, Tao Zou, Hansheng Wang

Covariance regression offers an effective way to model the large covariance matrix with the auxiliary similarity matrices. A sparse covariance regression (SCR) approach is proposed to handle the potentially high-dimensional predictors (i.e., similarity matrices). Specifically, the penalization method is used to identify the informative predictors and estimate their associated coefficients simultaneously. The Lasso estimator is first investigated, and subsequently, the folded concave penalized estimation methods (e.g., SCAD and MCP) are considered. However, the theoretical analysis of the existing penalization methods is primarily based on i.i.d. data, which is not directly applicable to the scenario. To address this difficulty, the non-asymptotic error bounds are established by exploiting the spectral properties of the covariance matrix and similarity matrices. Then, the estimation error bound is derived for the Lasso estimator, and the desirable oracle property of the folded concave penalized estimator is established. Extensive simulation studies are conducted to corroborate the theoretical results. The usefulness of the proposed method is also illustrated by applying it to a Chinese stock market dataset.

E0164: **A communication efficient boosting method for distributed spectral clustering**
Presenter: **Yingqiu Zhu**, University of International Business and Economics, China

Spectral clustering is one of the most popular clustering techniques in statistical inference. For large-scale datasets, spectral clustering is typically implemented through distributed computing. However, existing distributed implementations face two major challenges. First, the clustering performance is negatively affected by distributed computing since the topological structure of all objects has to be divided into distributed parts. Second, computer communication within a distributed system results in high communication costs. A communication-efficient algorithm for distributed spectral clustering is proposed to address these issues. The motivation stems from a theoretical comparison between the conventional spectral clustering algorithm, which operates on the entire dataset, and the local spectral clustering, performed on a subsample using a single computer. The critical factor that leads to the difference between the performances of global spectral clustering and local spectral clustering is identified. Based on the findings, a novel approach is proposed that iteratively aggregates the intermediate results generated by local spectral clustering. In this process, only low-dimensional vectors are exchanged between computers. The simulations and real data analysis results demonstrate that the proposed method apparently enhances the performance of distributed spectral clustering with low communication costs.

E0376: **Subsampling and jackknifing: A convenient solution for large data analysis with limited computational resources**
Presenter: **Shuyuan Wu**, Shanghai University of Finance and Economics, China
Co-authors: Xuening Zhu, Hansheng Wang
Modern statistical analysis often involves large data sets, for which conventional estimation methods are not suitable owing to limited computational resources. To solve this problem, a novel subsampling-based method is proposed with jackknifing. The key idea is to treat the whole sample as if it were the population. Then, multiple subsamples are obtained with greatly reduced sizes using simple random sampling with replacement. Sampling methods are not recommended without replacement, because this would incur a significant data processing cost when the processing occurs on a hard drive. However, such a cost does not exist if the data are processed in memory. Because subsampled data have relatively small sizes, they can be comfortably read into computer memory and processed. Based on subsampled data sets, jackknife-debiased estimators can be obtained for the target parameter. The resulting estimators are statistically consistent, with an extremely small bias. Finally, the jackknife-debiased estimators from different subsamples are averaged to form the final estimator. It is shown theoretically that the final estimator is consistent and asymptotically normal. Furthermore, its asymptotic statistical efficiency can be as good as that of the whole sample estimator under very mild conditions. The proposed method is easily implemented on most computer systems and thus is widely applicable.

E0894: **Statistical analysis of fixed mini-batch gradient descent estimator**
Presenter: **Haobo Qi**, Beijing Normal University, China
Co-authors: Feifei Wang, Hansheng Wang
A fixed mini-batch gradient descent (FMGD) algorithm is studied to solve optimization problems with massive datasets. In FMGD, the whole sample is split into multiple non-overlapping partitions. Once the partitions are formed, they are then fixed throughout the rest of the algorithm. For convenience, the fixed partitions are referred toas fixed mini-batches. Then, for each computation iteration, the gradients are sequentially calculated for each fixed mini-batch. Because the size of fixed mini-batches is typically much smaller than the whole sample size, it can be easily computed. This leads to much-reduced computation costs for each computational iteration. It makes FMGD computationally efficient and practically more feasible. To demonstrate the theoretical properties of FMGD, it starts with a linear regression model with a constant learning rate. Its numerical convergence and statistical efficiency properties are studied. It is found that sufficiently small learning rates are necessary for both numerical convergence and statistical efficiency. Nevertheless, an extremely small learning rate might lead to painfully slow numerical convergence. A diminishing learning rate scheduling strategy can be used to solve the problem. This leads to the FMGD estimator with faster numerical convergence and better statistical efficiency. Finally, the FMGD algorithms with random shuffling and a general loss function are also studied.

EO070 Room 212 RECENT ADVANCES AND DEVELOPMENT IN STATISTICAL MODELING Chair: Li-Hsien Sun

E0230: **Increment degradation model: A Bayesian perspective**
Presenter: **I-Tang Yu**, Tunghai University, Taiwan
One frequently employed approach for describing the degradation phenomenon involves using a degradation model that relies on stochastic processes. In a stochastic-process-based degradation model, the increments are assumed to follow a distribution with the additivity property. This property makes further inferences mathematically and statistically tractable. However, it limits the choices of the distributions. The aim is to use those distributions without the additivity property to model the increments. Under the frame of Bayesian analysis, Markov Chain Monte Carlo algorithms are developed to execute the necessary computations. Given that the proposed degradation models do not adhere to the additivity property, the challenges involved in predicting the lifetime of both online and offline products are tackled. The suitability of the proposed model is finally validated through a simulation study.

E0492: **Adaptive change point estimation: Interval time series analysis for GBM models**
Presenter: **Li-Hsien Sun**, National Central University, Taiwan
Co-authors: Chi-Yang Chiu
A method for detecting structural shifts is proposed within time series data, and the change-point estimation is obtained. In the field of finance, most models are developed for the daily closing price. Nevertheless, based on the intra-daily information from the financial market, maximum and minimum prices can also be observed. Hence, instead of a one-dimensional time series, an interval time series model is proposed that includes the daily maximum, minimum, and closing prices based on the geometric Brownian motion (GBM) model. The likelihood function and the corresponding maximum likelihood estimates (MLEs) are obtained using the Girsanov theorem and the Newton-Raphson (NR) algorithm. The proposed approach is evaluated through simulations. In empirical studies, the performance relies on real stock return data (S&P 500 index) in two distinct periods: the 2008 financial crisis and the COVID-19 pandemic in 2020.

E0507: **A modified VAR-deGARCH model for asynchronous multivariate financial time series via variational Bayesian inference**
Presenter: **Shih-Feng Huang**, National Central University, Taiwan
Co-authors: Wei-Ting Lai, Ray-Bing Chen
A modified VAR-deGARCH model is proposed, denoted by M-VAR-deGARCH, for modeling asynchronous multivariate financial time series with GARCH effects and simultaneously accommodating the latest market information. A variational Bayesian (VB) procedure is developed to infer the M-VAR-deGARCH model for structure selection and parameter estimation. Extensive simulations and empirical studies are conducted to evaluate the fitting and forecasting performances of the M-VAR-deGARCH model. The simulation results reveal that the proposed VB procedure produces satisfactory selection performances. In addition, empirical studies find that the latest market information in Asia can provide helpful information for predicting market trends in Europe and South Africa, especially when momentous events occur.

E0633: **Estimation of threshold-boundary logistic regression models**
Presenter: **ChihHao Chang**, National Chengchi University, Taiwan
The threshold boundary logistic regression (TBLR) model analyses binary data. The TBLR model combines logistic regression and threshold boundary functions using explanatory variables, allowing for the construction of the threshold boundary function by multiple explanatory variables to create linear or nonlinear classifiers. These classifiers split the binary data into two groups, and logistic regression models are fitted separately to each group. An ordered iterative algorithm named the TBLR-WSVM algorithm, which integrates weighted support vector machine (WSVM) and maximum likelihood estimation methods to estimate the TBLR model, is introduced. Simulation studies and empirical analyses are conducted to evaluate the performance of the TBLR-WSVM algorithm. Numerical analysis results demonstrate that the TBLR-WSVM algorithm exhibits robust estimation and prediction capabilities for linear and nonlinear threshold boundary logistic models, particularly under finite sample conditions.

EO171 Room 202 RECENT ADVANCES IN STATISTICAL PROCESS MONITORING AND CHANGE POINT DETECTION Chair: Jun Li

E0971: **Spatiotemporal surveillance of infectious diseases by statistical process control charts**
Presenter: **Peihua Qiu**, University of Florida, United States
Online sequential monitoring of the incidence rates of infectious diseases is critically important for public health. Governments have spent many resources building global, national, and regional disease reporting and surveillance systems. In these systems, conventional control charts, such as the cumulative sum (CUSUM) and the exponentially weighted moving average (EWMA) charts, are routinely included for disease surveillance purposes. However, these charts require many assumptions on the observed data that are rarely valid in practice, making their results unreliable to use. The purpose is to present a recent sequential monitoring approach for spatiotemporal disease surveillance, which can accommodate the dynamic nature of the observed disease incidence rates, spatiotemporal data correlation, and nonparametric data distribution. It is shown that the new method is more reliable than the commonly used conventional control charts for spatiotemporal surveillance of infectious diseases.

E1083: **Higher-criticism for multi-sensor change-point detection**
Presenter: **Yao Xie**, Georgia Institute of Technology, United States
Co-authors: Alon Kipnis, Tingnan Gong
The focus is on the distribution-free procedure based on higher criticism to address the sparse multi-stream sequential change-point detection problem. Namely, detecting a change point co-occurring is needed in a few data streams out of potentially many, while those affected streams are unknown in advance. The procedure involves testing for a change point in individual streams and combining p-values using higher criticism. As a by-product, the procedure also indicates a set of streams suspected to be affected by the change. It is shown that the procedure attains the information-theoretic optimal detection performance under a sparse Gaussian mean shift when individual tests are based on the LR or GLR. The effectiveness of the method is compared to other procedures using numerical evaluations.

E0755: **Variance change point detection with credible sets**
Presenter: **Oscar Hernan Madrid Padilla**, UCLA, United States
Co-authors: Lorenzo Cappello
A novel Bayesian approach is introduced to detect changes in the variance of a Gaussian sequence model, focusing on quantifying the uncertainty in the change point locations and providing a scalable algorithm for inference. Such a measure of uncertainty is necessary when change point methods are deployed in sensitive applications, for example, when one is interested in determining whether an organ is viable for transplant. The focus is on framing the problem as a product of multiple single changes in the scale parameter. The model is fit through an iterative procedure similar to what is done for additive models. The novelty is that each iteration returns a probability distribution on time instances, which captures the uncertainty in the change point location. Leveraging a recent result in the literature, the proposal is shown as a variational approximation of the exact model posterior distribution. The algorithm's convergence and the change point localization rate are studied. Extensive experiments in simulation studies and real data illustrate the usefulness of the proposed approach.

E1074: **Nonparametric Shiryaev-Roberts change-point detection**
Presenter: **Wei Ning**, Bowling Green State University, United States
Sequential change-point analysis, which identifies a change of probability distribution in an infinite sequence of random observations, has important applications in many fields. A good method should detect a change point as soon as possible and keep a low amount of false alarms. As one of the most popular methods, the Shiryaev-Roberts (SR) procedure holds many optimalities. However, its implementation requires the pre-change and post-change distributions to be known, which is not achievable in practice. A nonparametric version of the SR procedure is constructed by embedding different versions of empirical likelihood, assuming two training samples, before and after the change, are available for parameter estimations. Simulations are conducted to compare the performance of the proposed method with that of existing methods. The results show that when the underlying distribution is unknown, and training sample sizes are small, the proposed modified procedure shows an advantage by reducing the delay of detection.

EO177 Room 204 ADVANCES IN RANDOM FORESTS AND CAUSAL INFERENCE (VIRTUAL) Chair: Hiroshi Shiraishi

E1028: **Data adaptive random forest kernels via dimension reduction**
Presenter: **Tomoshige Nakamura**, Juntendo University, Japan
Co-authors: Hiroshi Shiraishi
A past study proposed random forests as an ensemble method that uses regression trees/decision trees as weak learners, which has demonstrated high accuracy in various regression and classification problems across different fields. Another study provided a new interpretation of random forests as a method for estimating data-adaptive kernel weighting functions based on a loss function. It showed that functional parameters characterized by local estimating equations can be estimated using random forests. They also proved the asymptotic normality for those estimators. Building upon these results, the asymptotic behavior of the kernel weighting function generated by random forests is investigated. It is found that random forests kernels converge to the Laplace kernel when the features are one-dimensional, while in the multidimensional case, they become the product of Laplace kernels for each feature dimension. This fact suggests that the expressive power of data-adaptive kernels generated by random forests is limited, and there is a problem of significantly lower accuracy when estimating functionals that can be represented by the sum of features, for example. The problem is demonstrated to be resolved by modifying the splitting rule of the trees constituting the random forest from a threshold-based split on a single variable to a data-adaptive split obtained through sufficient dimension reduction.

E1049: **An application of random forests to estimate the reporting delay in COVID-19 cases**
Presenter: **Xuanan Lin**, Keio University, Japan
Co-authors: Hiroshi Shiraishi
Accurate forecasting of COVID-19 cases is paramount for effective public health planning and response. However, the presence of reporting delays complicates this task, leading to underestimation or misrepresentation of the true disease burden. A novel approach leveraging random forests is introduced to evaluate the effect of reporting delays on COVID-19 case counts. By integrating historical case data with features indicative of reporting lags, such as infection time, patient density, and development periods, the random forests model is adapted to capture complex relationships and nonlinear effects inherent in the reporting process. Model performance is rigorously evaluated using the mean squared error (MSE), providing a quantitative measure of predictive accuracy. The application of random forests unveils insights into the temporal dynamics of reporting delays and their implications for epidemic surveillance and control efforts. The utility of machine learning methodologies, particularly the random forests, is underscored in unravelling the intricacies of infectious disease surveillance and informing evidence-based public health policies to mitigate the impact of reporting delays on COVID-19 case estimation.

E0944: **Asymptotic property for generalized random forests**
Presenter: **Hiroshi Shiraishi**, Keio University, Japan
Co-authors: Tomoshige Nakamura, Ryuta Suzuki
The asymptotic property of the generalized random forests (GRF) estimator proposed in a prior study is discussed. It derives the asymptotic normality of the GRF estimator, but it does not explicitly derive the rate of convergence and asymptotic variance. The aim is to derive the rate of convergence and explicit form of the asymptotic variance of the GRF estimator. To overcome this problem, the objective function is approximated as a class of Nadraya-Watoson type statistics and its asymptotic normality is derived by some modification of other studies.

EO217 Room 207 RECENT ADVANCES IN CAUSAL INFERENCE AND ITS APPLICATIONS Chair: Yuexia Zhang

E0405: **Introducing the specificity score: A measure of causality beyond P value**
Presenter: **Wang Miao**, Peking University, China
There has been considerable debate and doubt about the use of the P value in scientific research in recent years, particularly after its use has been banned in several prestigious journals. Much scientific research is concerned with uncovering causal associations. However, the P value, by definition, is a measure of the significance of a statistical association, which could be biased from the causal association of interest and lead to false discoveries due to confounding. A score measuring the specificity of causal associations and a specificity score-based test are introduced about the existence of causal effects in the presence of unmeasured confounding. Under certain conditions, this approach has controlled type I error and power approaching unity for testing the null hypothesis of no causal effect. This approach is particularly suitable for joint causal discovery with multiple treatments and multiple outcomes, such as gene expression studies, Mendelian randomization and EHR studies. A visualization approach using a specificity map is proposed to communicate all specificity score/test information in a universal and effective manner. Identification and estimation are briefly covered. Simulations are used for illustration, and an application to a mouse obesity dataset detects potential active effects of genes on clinical traits that are relevant to metabolic syndrome.

E0417: **Efficient nonparametric inference of causal mediation effects with nonignorable missing confounders**
Presenter: **Wei Li**, Renmin University of China, China
Causal mediation analysis is considered with confounders subject to nonignorable missingness in a nonparametric framework. The approach relies on shadow variables that are associated with the missing confounders but independent of the missingness mechanism. The mediation effect of interest is shown to be a weighted average of an iterated conditional expectation, which motivates the sieve-based iterative outward (SIO) estimator. The rate of convergence and asymptotic normality of the SIO estimator are derived, which does not suffer from the ill-posed inverse problem. Essentially, it is shown that the asymptotic normality is not affected by the slow convergence rate of nonparametric estimators of nuisance functions. Moreover, the estimator is demonstrated to be locally efficient and attains the semiparametric efficiency bound under certain conditions. The efficiency loss attributable is accurately depicted to missingness, and scenarios are identified in which efficiency loss is absent. A stable and easy-to-implement approach is also proposed to estimate asymptotic variance and construct confidence intervals for the mediation effects. Finally, the finite-sample performance of the proposed approach is evaluated through simulation studies, and it is applied to the CFPS data to show its practical applicability.

E0525: **Proxy-aided demand learning with an application on various pricing problems**
Presenter: **Tao Shen**, National University of Singapore, China
Co-authors: Yifan Cui
In data-driven demand learning, understanding customer willingness to pay presents a significant challenge due to the complex interplay between various influencing factors. The multifaceted relationship between quantities like price and sales is addressed, highlighting the difficulties in identifying the causal effect with the existence of unmeasured confounders. To mitigate bias in evaluating pricing decisions, proxy variables are introduced into the demand learning process. Data-driven pricing challenges are explored within a confounded environment, showcasing the practical application of the proposed demand learning process.

E0558: **Statistical inference in high-dimensional regression with hidden confounders by double debiased LASSO estimator**
Presenter: **Guoyou Qin**, Fudan University, China
The statistical inference in the high-dimensional linear regression is considered to have hidden confounders. A double debiased LASSO estimator is proposed based on the spectral transformation and the approximately inverse empirical covariance matrix of the transformed design matrix. The proposed estimator corrects the bias from the estimation of the high-dimensional coefficients and the hidden confounders without the sparse assumption on the precision matrix of the component of covariates unaffected by confounders. The asymptotic properties of the estimator for the individual component and finite-dimensional subset of the coefficient vector are presented. The performance of the estimator is investigated through simulation experiments and a real dataset.

EO317 Room 209 CLUSTERING AND CLASSIFICATION FOR TIME SERIES Chair: Angel Lopez Oriona

E0241: **Fuzzy clustering of circular time series based on a new dependence measure with applications to wind data**
Presenter: **Angel Lopez Oriona**, King Abdullah University of Science and Technology (KAUST), Saudi Arabia
Time series clustering is an essential machine learning task with applications in many disciplines. While the majority of the methods focus on time series taking values on the real line, very few works consider time series defined on the unit circle, although the latter objects frequently arise in many applications. The problem of clustering circular time series is addressed. To this aim, a distance between circular series is introduced and used to construct a clustering procedure. The metric relies on a new measure of serial dependence considering circular arcs, thus taking advantage of the directional character inherent to the series range. Since the dynamics of the series may vary over time, a fuzzy approach is adopted, which enables the procedure to locate each series into several clusters with different membership degrees. The resulting clustering algorithm is able to group series generated from similar stochastic processes, reaching accurate results with series coming from a wide variety of models. A simulation study shows that the proposed method outperforms several alternative techniques besides being computationally efficient. An interesting application involving time series of wind direction in Saudi Arabia highlights the potential of the proposed approach.

E0255: **Dynamic clustering of multivariate time series using DTGARCH model and spectral clustering**
Presenter: **Sipan Aslan**, King Abdullah University of Science and Technology, Saudi Arabia
Co-authors: Ceylan Yozgatligil, Cem Iyigun
A novel dynamic clustering procedure is introduced for multivariate time series derived from complex systems such as brain circuitry or financial markets. The objective is to devise a clustering in which latent underlying data-generating mechanisms (DGMs) form the cluster centres. In other words, multivariate time series are treated as objects to be grouped according to their similarities of underlying DGMs. The proposed approach mainly leverages the distinguishable features extracted from a nonlinear time series model, such as the double threshold generalized auto-regressive conditional heteroskedasticity (DTGARCH) model, and groups them by spectral clustering methodology. Approximating the time series DGMs using a rich model, such as DTGARCH, is proposed to enable the comparisons of complex, nonlinear, time-dependent features of underlying data-

generating processes and effectively track dynamic cluster changes. The efficiency of the approach is validated through synthetic and real-world datasets. Clustering accuracies compared with several distance measures designed for multivariate time series clustering. The presented framework offers significant implications for fields ranging from economics to neuroscience by providing a more nuanced understanding and analysis of time series data.

E0546: **Robust linkage methods for functional data clustering**
Presenter: **Tianbo Chen**, Anhui University, China
Clustering, an essential component of data mining and machine learning, serves as a statistical tool for classifying the data unsupervised. Among hierarchical clustering techniques, Ward's linkage method measures the incremental sum of squares errors (SSE, or graphically, the diameter of a cluster) when two clusters are merged. However, traditional linkage methods exhibit limitations in handling outliers and contaminations within datasets, compromising their partitioning capabilities. The aim is to introduce two robust Ward-like linkage methods for functional data clustering by only taking the most central curves into account. The cluster diameter is defined as the width of the band delimited by the most central curves selected by modified band depth and magnitude-shape outlyingness measure. Results from simulations and EEG data analysis demonstrate the superior performance of the proposed methods over conventional Ward's linkage and centroid linkage, particularly when different types of outliers and contaminations are presented.

E0616: **Homogeneity pursuit in forecasting high-dimensional functional time series: Is clustering necessary**
Presenter: **Chen Tang**, The Australian National University, Australia
Co-authors: Han Lin Shang, Yanrong Yang, Yang Yang
Joint modelling and forecasting high-dimensional functional time series (HDFTS) has begun to gain popularity in the literature. However, heterogeneity would deteriorate prediction accuracy. In this vein, pursuing homogeneity within the HDFTS is the key to improving forecast accuracy. Two methods are compared in the effort to pursue homogeneity: one is through the clustering framework based on a functional panel data model with fixed effects, and the other one is through a dual-factor model for HDFTS. Different from the functional panel data model with fixed effects, where the homogeneous part and the heterogenous part are additive, the dual-factor model puts the collection of functional time series in a lower dimension (homogeneous) and a group of population-specific basis functions (heterogeneous) as a product, which avoids clustering. An empirical study shows that the proposed model produces relatively accurate point and interval forecasts for age-specific mortality rates in 32 countries. The financial benefits associated with the improved mortality forecasts are translated into a life annuity pricing scheme.

EO025 Room 210 DESIGN AND ANALYSIS FOR ORDER-OF-ADDITION EXPERIMENTS Chair: Fasheng Sun

E0407: **Optimal designs for order-of-addition two-level factorial experiments**
Presenter: **Fasheng Sun**, Northeast Normal University, China
Co-authors: Qiang Zhao, Qian Xiao, Abhyuday Mandal
A new type of experiment called the order-of-addition factorial experiment, has recently received considerable attention in medical science and bioengineering. These experiments aim to simultaneously optimize the order of addition and dose levels of drug components. In the experimental design literature, dual-orthogonal arrays (DOAs) were recently introduced for such experiments. However, constructing flexible DOAs is a challenging task. A novel theory-guided search method is proposed that efficiently identifies DOAs of any size (if present). An algebraic construction is also provided that instantly leads to certain DOAs. Moreover, to address the potential issue of DOA ignoring interaction effects, a new type of optimal design is constructed under the expanded compound model, named the strong DOA (SDOA). Two algebraic constructions of the SDOA are provided. Theoretical results are established on the optimality of both DOAs and SDOAs. Simulation studies are performed to demonstrate the superiority of the proposed designs.

E0436: **Symmetrical analysis for the order-of-addition experiments**
Presenter: **Xueping Chen**, Jiangsu University of Technoology, China
The order-of-addition experiment has received significant attention in recent literature. When non-negligible block effects are present, the design and inference pose a novel and demanding challenge. A new inference approach is presented that is model-free but puissant in the interpretability of the experiment. This new approach uses variance-based decomposition to capture the symmetry of the output on the order permutations, then forms an intelligent data collection strategy to search the optimal settings of orders. It provides multiple options for additional experiments, facilitating adaptability to different block capacities. Theoretical supports are provided to illustrate the effectiveness of the proposed method, and numerical experiments demonstrate its effectiveness.

E0736: **Design for order-of-addition experiments with two level components**
Presenter: **Hengzhen Huang**, Guangxi Normal University, China
The statistical design for order-of-addition (OofA) experiments has received much recent interest due to its potential to determine the optimal sequence of multiple components, such as the optimal sequence of drug administration for disease treatment. The traditional OofA experiments focus mainly on the sequence effects of components, i.e., the experimenters fix the factor level of each component and observe how the response is affected by varying the sequences of components. However, the components may also have factorial effects in that changing their factor levels in a given sequence can affect the response. In view of this, the design problem for OofA experiments is considered, where each component experiments at two levels. A systematic method is given to construct OofA designs that jointly consider the sequence design and factorial design for all components. By appropriately choosing the sequence design and defining relations for the factorial design, it is shown that the combination of the two parts results in a balanced design with an economical run size. Moreover, the constructed designs enjoy a number of optimality properties such as D-, A- and E-optimalities under some empirical models. The design method proposed can be extended to some other practical situations like the number of process variables is different from the number of components and OofA experiments with multi-level components.

E0572: **Minimum aberration designs for of experiments**
Presenter: **Bing Guo**, Sichuan University, China
In fractional factorial experiments, the minimum aberration designs play an important role and are widely used because they can minimize the confounding among low-order effects. For the order of addition designs, concepts such as word length, resolution, and word length pattern are introduced. By minimizing the word length pattern in sequence, the order of addition minimum aberration designs are defined. These designs have a simple structure and are easy to interpret, making them a generalization of the component projection balanced designs. Meanwhile, such designs can minimize the confounding of low-order effects in data analysis. Some theoretical properties of the order of addition minimum aberration designs are also explored.

EO165 Room 307 RECENT DEVELOPMENT ON DEPENDENT FUNCTIONAL DATA Chair: Han Lin Shang

E0467: **Nonstationary functional time series forecasting**
Presenter: **Yang Yang**, University of Newcastle, Australia
Co-authors: Han Lin Shang
A nonstationary functional time series forecasting method is proposed with an application to age-specific mortality rates observed over the years. The method begins by taking the first-order differencing and estimates its long-run covariance function. Through eigen-decomposition, a set of estimated functional principal components and their associated scores are obtained for the differenced series. These components allow the reconstruction of the original functional data and compute the residuals. To model the temporal patterns in the residuals, dynamic functional principal component analysis is again performed, and its estimated principal components and the associated scores for the residuals are extracted. As a byproduct, a geometrically decaying weighted approach is introduced to assign higher weights to the most recent data than those from the distant past. Using the Swedish age-specific mortality rates from 1751 to 2022, the weighted dynamic functional factor model is demonstrated to produce more accurate point and interval forecasts, particularly for male series exhibiting higher volatility.

E0480: **Forecasting density-valued functional panel data**
Presenter: **Han Lin Shang**, Macquarie University, Australia
Co-authors: Cristian Felipe Jimenez Varon, Ying Sun
A statistical method is introduced for modeling and forecasting functional panel data, where each element is a density. Density functions are nonnegative and have a constrained integral and thus do not constitute a linear vector space. A center log-ratio transformation is implemented to transform densities into unconstrained functions. These functions exhibit cross-sectionally correlation and temporal dependence. Via a functional analysis of variance decomposition, the unconstrained functional panel data is decomposed into a deterministic trend component and a time-varying residual component. A functional time series forecasting method based on the estimation of the long-range covariance is implemented to produce forecasts for the time-varying component. By combining the forecasts of the time-varying residual component with the deterministic trend component, h-step-ahead forecast curves are obtained for multiple populations. Illustrated by age- and sex-specific life-table death counts in the United States, the proposed method is applied to generate forecasts of the life-table death counts for 51 states.

E0635: **Eigen-analysis for functional time series**
Presenter: **Yanrong Yang**, The Australian National University, Australia
Co-authors: Yuan Gao, Han Lin Shang, Yang Yang
Spectral analysis is important in dimension reduction for functional data. Under a general data structure for functional time series, the influence of temporal dependence is studied on empirical eigenvalues and eigenvectors from the sample covariance operator. Asymptotic properties of empirical eigenvalues are established to quantify such influence. Based on the developed theory, a new algorithm is proposed to recover the non-stationary subspace and the principal component subspace. Various simulations are constructed and empirical analysis includes eigen-analysis on global mean temperature data and Australian mortality data.

E0738: **Enhanced functional data alignment with exogenous variables**
Presenter: **Wenlin Dai**, Renmin University of China, China
The alignment of functional data has been a topic of significant interest in research. The alignment of functional data is addressed while considering the temporal dependence between sample curves. Meanwhile, the warping functions are modeled as the function of exogenous variables, e.g., time of record. Numerical simulations showcase the superiority of the approach, and the validation of actual LOFAR graph data further confirms the effectiveness of the proposed method. The results highlight the importance of incorporating exogenous variables in functional data alignment and the potential applications of this method in various fields.

EO074 Room 313 LARGE-SCALE TIME SERIES MODELS Chair: Yubo Tao

E0279: **Estimation and inference for three-dimensional panel data models**
Presenter: **Bin Peng**, Monash University, Australia
Estimation and inferential methods are developed for three-dimensional (3D) panel data models with homogeneous/heterogeneous coefficients. The 3D panel data models specify the nature of common shocks through the use of a hierarchical factor structure (i.e., global factors and sector factors). Accordingly, an approach to estimating the hierarchy is developed, thus enabling a better understanding of the relative importance of the two types of unobservable shocks. Second, bias-corrected estimators are proposed, and bootstrap procedures are given to construct the confidence intervals for the parameters of interest while allowing for correlation along three dimensions of idiosyncratic errors. The theoretical findings are justified using extensive simulations. In an empirical study, the twin hypotheses of conditional and unconditional convergence are examined for manufacturing industries across countries.

E0506: **Transfer learning in conditional factor models**
Presenter: **Yubo Tao**, University of Macau, China
The focus is on considering the estimation and prediction of a conditional latent factor model in the setting of transfer learning where, in addition to observations from the target model, auxiliary datasets are available. To effectively utilize the auxiliary datasets, a transfer learning algorithm is employed in conjunction with the instrumented principle component analysis (IPCA) to estimate the conditional latent factor models. Given the informativeness of the auxiliary datasets, a trans-IPCA algorithm is proposed, and its ℓ_1/ℓ_2-estimation error bounds are derived. It is proven that when the target and sources are sufficiently close to each other, these bounds could be improved over those of the classical IPCA estimator and its penalized variants using only target data under mild conditions. When the set of informative auxiliary data is unknown, a data-driven and algorithm-free procedure is introduced to detect transferable samples. Monte Carlo simulations confirm the superior performance of the proposed estimator compared to classical and penalized IPCA models, both in-sample and out-of-sample.

E0554: **Estimating time-varying networks for high-dimensional time series**
Presenter: **Yuning Li**, University of York, United Kingdom
Time-varying networks are explored for high-dimensional locally stationary time series, using the large VAR model framework with transition and (error) precision matrices evolving smoothly over time. Two types of time-varying graphs are investigated: one containing directed edges of Granger causality linkages and the other containing undirected edges of partial correlation linkages. Under the sparse structural assumption, a penalized local linear method is proposed with time-varying weighted group LASSO to jointly estimate the transition matrices and identify their significant entries and a time-varying CLIME method to estimate the precision matrices. The estimated transition and precision matrices are then used to determine the time-varying network structures. Under some mild conditions, the theoretical properties of the proposed estimates are derived, including the consistency and oracle properties. In addition, the methodology and theory are extended to cover highly correlated large-scale time series, for which the sparsity assumption becomes invalid, and it is allowed for common factors before estimating the factor-adjusted

time-varying networks. Extensive simulation studies and an empirical application are provided to a large U.S. macroeconomic dataset to illustrate the finite-sample performance of the methods.

E0551: **Shrinkage estimation of multiple structural breaks in spatial panel data models with multifactor error structure**
Presenter: **Chaowen Zheng**, University of Southampton, United Kingdom
Co-authors: Siqi Dai
The aim is to consider a spatial panel data model where multiple structural breaks occur in both the coefficients for the spatial lagged dependent variable and regressors. The model can accommodate cross-sectional dependence arising from spatial dependence and unknown common factors. To tackle the challenging issues of endogeneity and time heterogeneity, a novel penalized generalized method is proposed for moments estimation with common correlated effects (PGMM-CCEX). Specifically, the PGMM-CCEX method uses cross-sectional averages of regressors as factor proxies when constructing the instrumental variables and employs adaptive group fused lasso to detect multiple structural breaks. It is shown that the PGMM-CCEX method can consistently estimate the number of breaks and their dates, and the resulting regime-specific coefficients are also consistent and asymptotically normally distributed. Monte Carlo simulations show that the PGMM-CCEX method has the superior finite-sample performance of the proposed estimators, which is quite satisfactory. An empirical application to cross-country economic growth of 103 countries from 1970-2019 reveals a more complete and time-varying picture of the driving forces behind economic growth.

EO091 Room 405 RECENT ADVANCES IN STATISTICAL METHODS FOR STOCHASTIC PROCESSES Chair: Masayuki Uchida

E0328: **Estimation for a discretely observed linear parabolic SPDE in two space dimensions based on triple increments**
Presenter: **Masayuki Uchida**, Osaka University, Japan
Co-authors: Yozo Tonaki, Yusuke Kaino
Parameter estimation for a linear parabolic second-order stochastic partial differential equation (SPDE) is addressed in two space dimensions using high-frequency spatio-temporal data. A driving process of the SPDE is assumed to be a Q-Wiener process. A prior study investigated parameter estimation of a linear parabolic second-order SPDE in one space dimension driven by a cylindrical Wiener process based on temporal and spatial increments (double increments) and proposed minimum contrast estimators (MCEs) with asymptotic normality. MCEs of the coefficient parameters of the SPDE are first introduced in two space dimensions based on temporal and two-dimensional spatial increments (triple increments) by applying the prior study's method to the SPDE in two space dimensions. Next, by using the MCEs, an approximate coordinate process of the SPDE is derived. Finally, parametric adaptive estimators of the coefficient parameters of the SPDE are proposed using the approximate coordinate process. Under certain regularity conditions, asymptotic normality of the adaptive estimators is shown. In addition, numerical simulations of the proposed estimators are presented.

E0347: **Approximation and estimation of scale functions for spectrally negative Levy processes**
Presenter: **Yasutaka Shimizu**, Waseda University, Japan
The scale function holds significant importance within the fluctuation theory of Levy processes, particularly in addressing exit problems. However, its definition is established through the Laplace transform, thereby lacking explicit representations in general. A novel series representation is introduced for this scale function, employing Laguerre polynomials to construct a uniformly convergent approximate sequence. Additionally, statistical inference is derived based on specific discrete observations, presenting estimators of scale functions that are asymptotically normal.

E0474: **Estimation of the number of relevant factors from high-frequency data**
Presenter: **Yuta Koike**, University of Tokyo, Japan
Factor models play an important role in modeling financial asset prices, both theoretically and practically. Traditionally, only "strong" factors that are correlated with almost all the assets under analysis have been considered, but in recent years, "weak" factors that are correlated with only some assets have attracted attention. It is discussed how to estimate the number of factors that drive the model, including "some" weak factors, from high-frequency data. In particular, a general setting is considered in which the log price process is modeled as a semimartingale, possibly with jumps. Theoretically, the growth rate of the spectral norm of the realized covariance matrix plays a key role, and a new result is given from this perspective.

E0479: **Locally differentially private drift parameter estimation for iid paths of diffusion processes**
Presenter: **Arnaud Gloter**, Universite d Evry Val d Essonne, France
Co-authors: Chiara Amorino, Helene Halconruy
The problem of parametric drift estimation is addressed for N discretely observed iid SDEs, considering the additional constraints that only privatized data can be published and used for inference. The concept of local differential privacy is formally introduced for a system of stochastic differential equations. The aim is to estimate the drift parameter by proposing a contrast function based on a pseudo-likelihood approach. A suitably scaled Laplace noise is incorporated to satisfy the privacy requirement. One main result consists of deriving explicit conditions on the privacy level for which the associated estimator is proven to be consistent. The asymptotic behavior of the estimator is also derived, and how the rate of convergence is linked to the privacy level is determined. This holds true as the discretization step approaches zero and the number of processes N tends to infinity.

EO081 Room 406 MODELING MULTIVARIATE EXTREMES: THEORY AND APPLICATIONS Chair: Pavel Krupskiy

E0368: **Estimation and inference for extreme continuous treatment effects**
Presenter: **Liuhua Peng**, The University of Melbourne, Australia
Co-authors: Wei Huang, Shuo Li
Estimation and inference for the treatment effect are studied on deep tails of the potential outcome distributions corresponding to a continuously valued treatment, namely the extreme continuous treatment effect. Two measures are considered for the tail characteristics: the quantile function and the tail mean function, which is defined as the conditional mean beyond a quantile level. Then, for a quantile level close to 1, we define the extreme quantile treatment effect (EQTE) and extreme average treatment effect (EATE), which are, respectively, the differences of the quantile and tail mean at different treatment statuses. Estimators are proposed for the EQTE and EATE based on tail approximations from the extreme value theory. The limiting theory is for the EQTE and EATE processes indexed by a set of quantile levels and hence facilitates uniform inference for the EQTE and EATE over multiple tail levels. Simulations suggest that the method works well in finite samples, and an empirical study of its practical merits is illustrated.

E0482: **A deep geometric approach to modelling multivariate extremes**
Presenter: **Jordan Richards**, King Abdullah University of Science and Technology, Saudi Arabia
Co-authors: Callum Murphy-Barltrop, Reetam Majumder
The geometric representation for multivariate extremes, where data is split into radial and angular components and the radial component is modelled conditionally on the angle, provides an exciting new approach to modelling the extremes of multivariate data. Through a consideration of scaled

sample clouds and limit sets, it provides a flexible, semi-parametric model for extremes that connects multiple classical extremal dependence measures; these include the coefficients of tail dependence and asymptotic independence and parameters of the conditional extremes framework. Although the geometric approach is becoming an increasingly popular modelling tool for multivariate extremes, inference with this framework is limited to a low dimensional setting ($d < 4$). The first deep representation is proposed for geometric extremes. By leveraging the predictive power and computational scalability of neural networks, asymptotically-justified yet flexible semi-parametric models are constructed for extremal dependence of high-dimensional data. The efficacy of the deep approach is showcased by modelling the complex extremal dependence between metocean variables sampled from the North Sea.

E0524: **Flexible max stable processes for fast and efficient inference**
Presenter: **Peng Zhong**, University of New South Wales, Australia
Co-authors: Boris Beranger, Scott Sisson
Max-stable processes serve as the fundamental distribution family in extreme value theory. However, likelihood-based inference methods for max-stable processes still heavily rely on composite likelihood, rendering them intractable in high dimensions due to their intractable densities. A fast and efficient inference method is introduced for max-stable processes based on their angular densities for a group of max-stable processes whose angular densities do not put mass on the boundary space of the simplex. The efficiency of the proposed method is demonstrated through two new max-stable processes, the truncated extremal-t process and the skewed Brown-Resnick process. The proposed method is shown to be computationally efficient and can be applied to large datasets. Furthermore, the skewed Brown-Resnick process contains the popular Brown-Resnick model as a special case and possesses nonstationary extremal dependence structures. Finally, the method is showcased with the skewed Brown-Resnick model on a real dataset.

E0559: **Estimation of max-stable random fields using the periodogram for extreme events**
Presenter: **Laleh Tafakori**, RMIT University, Australia
In recent years, there has been significant interest in the analysis of extreme events within stochastic processes. This has led to the exploration of various methodologies and techniques aimed at understanding and predicting extreme occurrences in both spatial and temporal domains. One approach involves the study of strictly stationary random fields with regularly varying marginal and finite-dimensional distributions. By exploiting the regular variation, the concept of the spatial extremogram is developed, which focuses specifically on the largest values present within the random field. Additionally, methods for estimating the corresponding extremal spectral density and its associated estimator, the extremal periodogram, have been devised. These existing procedures help in facilitating the analysis of extreme events and providing insights into their underlying characteristics. Furthermore, with advancements in computational capabilities, max-stable processes have become a cornerstone in the analysis of spatiotemporal extreme events. Simulation techniques play a crucial role in the inference of certain characteristics, particularly in spatial risk assessment for future events. An overview of existing procedures for simulating extreme events within stochastic processes is provided. By contextualizing these methodologies and comparing their theoretical properties, the aim is to enhance understanding and facilitate informed decision-making in the analysis of extreme events.

EO254 Room 408 RECENT ADVANCES IN EXPERIMENTAL DESIGN AND ANALYSIS Chair: Qian Xiao

E0366: **A construction method for Maximin L1-distance Latin hypercube designs**
Presenter: **Ru Yuan**, Zhongnan University of Economics and Law, China
Co-authors: Yuhao Yin, Hongquan Xu, Min-Qian Liu
Maximin distance designs are a kind of space-filling design and are widely used in computer experiments. However, although much work has been done on constructing such designs, doing so for a large number of rows and columns remains challenging. A theoretical construction method is proposed that generates a maximin L_1-distance Latin hypercube design with a run size that is close to the number of columns or half the number of columns. The theoretical results show that some of the constructed designs are both maximin L_1-distance and equidistant designs, which means that their pairwise L_1-distances are all equal and that they are uniform projection designs. Other designs are asymptotically optimal under the maximin L_1-distance criterion. Moreover, the proposed method is efficient for constructing high-dimensional Latin hypercube designs that perform well under the maximin L_1-distance criterion.

E0367: **Penalized additive Gaussian process for auto-tuning of quantitative and qualitative factors in Black-Box systems**
Presenter: **Yongxiang Li**, Shanghai Jiao Tong University, China
Optimizing black-box systems with both quantitative and qualitative (QQ) factors is critical in various applications where resource-intensive or time-consuming evaluations make factor-level selection critical. Traditional sensitivity analysis lacks a unified framework for simultaneously screening important QQ factors and struggles to select important qualitative levels. To address this, a penalized additive Gaussian process (PAGP) model is introduced, featuring an interpretable additive (IA) covariance function for QQ factors. This allows sparsity penalties that enable the identification of critical qualitative levels. The three-step model fitting approach utilizes derivatives for acceleration, and a tailored ADMM is proposed to optimize the L1 regularized likelihood. Then, qualitative level screening is proposed utilizing sparse regularization and quantitative factor selection leveraging Shapley value. Finally, Bayesian optimization is introduced to PAGP for the optimization of black-box systems with QQ factors, and the sparse covariance function will guide Bayesian optimization in efficiently searching the important qualitative levels. PAGP distinguishes itself by enabling sparse regularization and efficient screening of qualitative levels. Simulation studies validate the outperformance of PAGP, and it is also applied to the design of paper pilots and neural networks.

E0579: **Uniform designs for experiments with branching and nested factors**
Presenter: **Feng Yang**, Sichuan Normal University, China
The factors that only exist at certain levels of other factors are called nested factors. The factors that lead to such nested factors are called the branching factors. Experiments with branching and nested factors occur frequently in practical applications. Designing for such experiments is challenging because of the special relevancy between the branching and nested factors. Uniform designs are proposed for experiments involving branching and nested factors. A novel criterion is introduced to measure the uniformity of such designs, and the corresponding lower bound is also given. The construction methods of uniform designs for experiments with branching and nested factors are provided, and their effectiveness is verified by simulation comparisons and a practical manufacturing experiment. The proposed method allows each of the branching, nested and shared factors to be either qualitative or quantitative. Moreover, the run size and the levels of quantitative factors are very flexible, such that the method works well for physical and computer experiments.

E0854: **A distance metric based space filling subsampling method for nonparametric models**
Presenter: **Dianpeng Wang**, Beijing Institute of Technology, China
Taking subset samples from the original data set is an efficient and popular strategy for handling massive data too large to be directly modelled. Employing a subsampling scheme to collect observations intelligently to optimize inference and prediction accuracy is crucial. A proportionate sampling method is proposed that uses distance metric-based strata to select subsamples from high-volume data sets. To minimize the maximal distance from pairs of samples that are located in the same stratum, Voronoi cells of the thinnest covering lattices are used to partition the space.

With the help of an algorithm to quickly identify the cell an observation is located in, the computational cost of the subsampling method is proportional to the number of observations and irrelevant to the number of cells, which makes the method applicable to extremely large data sets. Results from simulated studies and real data analysis show that the new method is remarkably better than existing approaches when used in conjunction with a k-nearest neighbor or Gaussian process models.

EO016 Room 411 (Virtual sessions) EXTREME VALUE MODELLING, PREDICTION AND RISK ASSESSMENT Chair: Stefano Rizzelli

E0552: **Inference for extremal regression with dependent heavy-tailed data**
Presenter: **Gilles Stupfler**, University of Angers, France
Co-authors: Abdelaati Daouia, Antoine Usseglio-Carleve
Nonparametric inference on tail conditional quantiles and their least squares analogs, expectiles, remains limited to i.i.d. data. A fully operational inferential theory is developed for extreme conditional quantiles and expectiles in the challenging framework of strong mixing, conditional heavy-tailed data whose tail index may vary with covariate values. It requires a dedicated treatment to deal with data sparsity in the far tail of the response, in addition to handling difficulties inherent to mixing, smoothing, and sparsity associated with covariate localization. The pointwise asymptotic normality of the estimators is proven, and optimal rates of convergence reminiscent of those found in the i.i.d. regression setting are obtained but have not been established in the conditional extreme value literature. The assumptions hold in a wide range of models. Full bias and variance reduction procedures are proposed, and simple but effective data-based rules for selecting tuning hyperparameters are used. The inference strategy is shown to perform well in finite samples and is showcased in applications to stock returns and tornado loss data.

E0570: **Asymptotic properties of the maximum likelihood estimator within the block maxima framework**
Presenter: **David Carl**, Bocconi University, Italy
Co-authors: Simone Padoan, Stefano Rizzelli
The asymptotic properties of the maximum likelihood estimator for the extreme value index within the block maxima setting are still not fully understood. So far, it has been shown that likelihood maximizers over compact sets that contain the truth are consistent, but no convergence rates for such estimators were derived. On the other hand, for suitably fast shrinking sets around the truth, there exist local maximizers of the likelihood that are asymptotically Gaussian distributed with the usual parametric convergence rate and that are eventually unique. In this work, we show that we can extend the results concerning uniqueness, convergence rate, and asymptotic Gaussianity to likelihood maximizers over compact sets if the extreme value index is positive.

E0883: **Extremal random forests**
Presenter: **Nicola Gnecco**, University of Geneva, Switzerland
Co-authors: Edossa Merga Terefe, Sebastian Engelke
Classical methods for quantile regression fail in cases where the quantile of interest is extreme, and only a few or no training data points exceed it. Asymptotic results from extreme value theory can be used to extrapolate beyond the range of the data, and several approaches exist that use linear regression, kernel methods or generalized additive models. Most of these methods break down if the predictor space has more than a few dimensions or if the regression function of extreme quantiles is complex. A method for extreme quantile regression that combines the flexibility of random forests with the theory of extrapolation is proposed. The extremal random forest (ERF) estimates the parameters of a generalized Pareto distribution, conditional on the predictor vector, by maximizing a local likelihood with weights extracted from a quantile random forest. The shape parameter is penalized in this likelihood to regularize its variability in the predictor space. Under the general domain of attraction conditions, we show the consistency of the estimated parameters in both the unpenalized and penalized cases. Simulation studies show that the ERF outperforms both classical quantile regression methods and existing regression approaches from extreme value theory. The methodology is applied to extreme quantile prediction for U.S. wage data.

E0932: **Likelihood-based inference for the peaks-over-threshold method in time series**
Presenter: **Simone Padoan**, Bocconi University, Italy
Co-authors: Stefano Rizzelli
The focus is on the popular peaks-over-threshold method and stationary time series that satisfy the beta-mixing condition. Within the framework, it is described how accurate inference can still be achieved using likelihood-based procedures that rely on the generalized Pareto likelihood function derived from the independence case.

EC287 Room 108 COMPUTATIONAL AND METHODOLOGICAL STATISTICS Chair: Shan Yu

E0952: **Fast computation of the bootstrap method for incomplete data**
Presenter: **Masahiro Kuroda**, Okayama University of Science, Japan
The bootstrap method is a powerful tool for making statistical inference about a parameter of a statistical model. The bootstrap generates a sample by randomly sampling with replacement from the observed data. The maximum likelihood estimate (MLE) of the parameter is computed from this sample. By repeating the bootstrap sampling procedure, the parameter distribution can be obtained. When dealing with incomplete observed data, an iterative computation step is required to find the MLE in the bootstrap procedure. The EM algorithm is used in this step and applied to each of the bootstrap samples. However, the bootstrap can be time-consuming due to the slow convergence of the EM algorithm. To address this issue, a fast bootstrap method is proposed that includes an acceleration step to speed up the convergence of the EM algorithm. Numerical experiments apply the proposed bootstrap method to contingency table analysis and examine its performance in terms of the number of iteration and CPU time.

E0977: **Wasserstein k-centers clustering for distributional data**
Presenter: **Ryo Okano**, The University of Tokyo, Japan
Co-authors: Masaaki Imaizumi
Distributional data arise when each data point can be regarded as a probability distribution, and its analysis is gaining increasing attention in statistics. Because the space of probability distributions does not have a vector space structure, distributional data cannot be analyzed using existing methods devised for Euclidean functional data. In particular, cluster analysis of distributional data is still under development. Adopting the Wasserstein metric, a novel clustering method for distributional data on the real line is proposed. The clustering method follows the k-centers clustering approach for functional data that accounts for the mean and the modes of variation differentials between clusters. The notions of Frechet mean, and geodesic principal component analysis are employed in the Wasserstein space to define the mean and the modes of variation structures of clusters of distributional data. These structures are used in a reclassification step to predict cluster membership of each distribution based on a non-parametric random-effect model. Through a simulation study and real data application, the proposed distributional clustering method is demonstrated to improve cluster quality compared to conventional clustering algorithms.

E0290: **A multiplier approach for nonparametric estimation of the extreme quantiles of compound frequency distributions**
Presenter: **Helgard Raubenheimer**, North-West University, South Africa
Co-authors: Riaan de Jongh, Charl Pretorius, Tertius de Wet
Estimating operational risk reserves is still widely done using the loss distribution approach. The accuracy of the estimation depends heavily on the accuracy with which the extreme quantiles of the aggregate loss distributions are estimated. Many approaches have been proposed to estimate the extreme quantiles of this compound distribution, amongst other approximations based on the underlying severity distribution, such as the single loss and perturbative approximations. Both the approximation approaches suggest using an even more extreme quantile of the underlying severity distribution. The estimation of these extreme quantiles lacks accuracy, so the obvious question is, why not estimate a less extreme or lower quantile of the severity distribution, hopefully with better accuracy, and then use a multiplier to approximate the extreme quantile? Using extreme value theory, first- and second-order multipliers are derived to estimate the extreme quantile of the severity distribution, which is approximated by the single loss and perturbative approximations. Nonparametric estimators for the multipliers are suggested and evaluated using a simulation study. The simulation study results suggest that the second-order multiplier, based on the second-order perturbative approximation, should be a good choice for practical applications.

E1119: **Refining soccer rankings: Balancing scored points, goals for and against with kendall correlations and radar charts**
Presenter: **Valerio Ficcadenti**, London South Bank University, United Kingdom
Co-authors: Raffaele Mattera, Roy Cerqueti
Computational challenges of ranking when more than a feature is used in the ranking exercise are addressed. We introduce a novel ranking method showcasing an application in soccer championships, utilizing computational statistics derived from a combination of Kendall correlations and radar charts. The proposed method considers goals scored and conceded by teams in individual matches as additional scoring factors beyond the traditional win-draw-lose system in soccer. This approach aims to reduce biases in existing scoring rules, such as the mathematical certainty of some teams winning before the end of the competitions. The methodology involves calculating the areas of radar charts, which represent the normalized Kendall correlation values between different ranking metrics. By transforming these areas into target Kendall tau values, we identify alternative team rankings that align with the target correlations. This process provides a different view of team performance and reduces the impact of biases present in the current ranking system. We apply our methodology to the Italian Serie A championships, showcasing its capability to generate different rankings through a robust computational framework. This innovative approach demonstrates the potential for improving ranking systems in sports competitions by incorporating detailed performance metrics.

Thursday 18.07.2024 08:15 - 09:55 Parallel Session F – EcoSta2024

EO048 Room 102 MACROECONOMIC POLICIES (VIRTUAL) Chair: Etsuro Shioji

E0788: **Automation and monetary policy: An empirical investigation**
Presenter: **Takuji Fueki**, Hitotsubashi University, Japan
Co-authors: Jouchi Nakajima
The evolving impact of monetary policy is examined amid advancements in autonomous machines or robots. To this end, a nonlinear VAR model is developed, distinguishing between a low automation regime and a highly developed automation regime. This model allows capturing potential nonlinearities in macroeconomic dynamics due to advances in automation in the U.S. and Japan. The findings reveal significant asymmetries in the responses of key macroeconomic variables to identified monetary policy shocks across these two regimes. Notably, the effectiveness of monetary policy diminishes as automation advances, aligning with theoretical predictions. This finding holds crucial implications for post-automation monetary policies.

E0796: **How do people tweet against inflation in Japan**
Presenter: **Toshitaka Sekine**, Hitotsubashi University, Japan
During the chronic deflation era starting in the 1990s, Japanese inflation expectations were said to be firmly anchored at a very low level, around zero, and households were quite against any price hikes. This resulted in the zero-inflation norm, implying prices are not only expected to rise but also should not be raised. A natural language processing technique is applied to tweets to uncover whether there has been any change in households' sentiments against price hikes in recent years.

E0663: **Forecasting GDP growth using stock returns in Japan: A factor-augmented MIDAS approach**
Presenter: **Hiroshi Morita**, Tokyo Institute of Technology, Japan
Using the rich time-series and cross-sectional information of the stock market, the purpose is to examine which dimensions of information contribute to the accuracy of GDP growth rate forecasts. Methodologically, MIDAS (mixed data sampling) regression analysis is combined with factor analysis and applied to the Japanese economy. The results reveal that the use of factors significantly improves forecast accuracy and that extracting factors from a broader set of stock prices further improves accuracy, suggesting the important role of cross-sectional stock market information in forecasting macroeconomic activity.

E0432: **Responses of households' expected inflation to oil prices and the exchange rate: Evidence from daily data**
Presenter: **Etsuro Shioji**, Chuo University, Japan
The purpose is to examine if daily data can help predict monthly changes in the inflation expectations of households. The mixed-data sampling (MIDAS) method is utilized to study determinants of a monthly survey-based measure of household inflation expectations in Japan. Two types of higher-than-monthly frequency variables are incorporated into the analysis. The first is a group of "traditional" daily indicators, such as the exchange rate and crude oil prices. The second consists of more unconventional measures. They include a daily price index called CPINOW, which is constructed from scanner data recorded at supermarkets from all over the country and (weekly) retail gasoline prices. It is found that, although both groups of variables help explain actual CPI inflation, only the latter turns out to be significant when the dependent variable is expected inflation. This finding suggests that people's perceptions are affected predominantly by prices that they actually observe in their everyday lives at supermarkets and gasoline stations.

EO080 Room 103 BOOTSTRAP METHODS IN MODERN SETTINGS Chair: Miles Lopes

E0298: **When does massive data bootstrap work**
Presenter: **Nan Zou**, Macquarie University, Australia
Co-authors: Patrice Bertail, Liuhua Peng, Dimitris Politis, Han Lin Shang, Stanislav Volgushev
In classic statistical inference, the bootstrap stands out as a simple, powerful, and data-driven technique. However, when coping with massive data sets, which are increasingly prevalent these days, the bootstrap can be computationally infeasible. To speed up the bootstrap for massive data sets, the bag of little bootstraps was invented in 2014. Despite its considerable popularity, little is known about the theoretical properties of the bag of little bootstrap, including reliability. Indeed, the preliminary results have already raised questions on the applicability of the bag of little bootstraps under a simple but important setting. The procedure for the bag of little bootstraps is first introduced, and then its theoretical applicability is investigated. Specifically, for this applicability, a counterexample for the claimed sufficient condition is presented in the literature and, as a remedy, a hopefully correct, generic sufficient condition is provided.

E0314: **New Gaussian and bootstrap approximations for suprema of empirical processes**
Presenter: **Alexander Giessing**, National University of Singapore, Singapore
New non-asymptotic Gaussian approximation results are developed for the sampling distribution of suprema of empirical processes when the indexing function class F varies with the sample size n and may not be Donsker. Prior approximations of this type required upper bounds on the metric entropy of F and uniform lower bounds on the variance of f in F, which both limited their applicability to high-dimensional inference problems. In contrast, the new results are based on simpler conditions of boundedness, continuity, and the strong variance of the approximating Gaussian process. The results are broadly applicable and yield a novel procedure for bootstrapping the distribution of empirical process suprema based on the truncated Karhunen-Loeve decomposition of the approximating Gaussian process. The flexibility of this new bootstrap procedure is demonstrated by applying it to three fundamental problems in high-dimensional statistics: simultaneous inference on parameter vectors, inference on the spectral norm of covariance matrices, and construction of simultaneous confidence bands for functions in reproducing kernel Hilbert spaces.

E0320: **Yurinskii's coupling for martingales**
Presenter: **Ricardo Masini**, UC Davis, United States
Yurinskii's coupling is a popular theoretical tool for non-asymptotic distributional analysis in mathematical statistics and applied probability, offering a strong Gaussian approximation with an explicit error bound under easily verified conditions. Originally started in ℓ^2-norm for sums of independent random vectors, it has recently been extended both to the ℓ^p-norm, for $1 \leq p \leq \infty$, and to vector-valued martingales in ℓ^2-norm, under some strong conditions. As the main result is presented, a Yurinskii coupling for approximate martingales in ℓ^p-norm, under substantially weaker conditions than those previously imposed. The formulation further allows for the coupling variable to follow a more general Gaussian mixture distribution, and a novel third-order coupling method is provided that gives tighter approximations in certain settings. The main result specializes in mixingales, martingales, and independent data, and uniform Gaussian mixture strong approximations are derived for martingale empirical processes. Applications to non-parametric partitioning-based and local polynomial regression procedures are provided.

E0336: **A resized parametric bootstrap method for inference of a high-dimensional generalized linear model**
Presenter: **Qian Zhao**, University of Massachusetts, Amherst, United States
Co-authors: Emmanuel Candes
Accurate statistical inference can be challenging when the ratio between the number of parameters and the sample size is not negligible. One example is logistic regression: when the number of parameters increases with the sample size, approximations based on either classical asymptotic theory or bootstrap are grossly off the mark. A resized bootstrap method is introduced to infer model parameters from a logistic regression in arbitrary dimensions. As in the parametric bootstrap, observations from a distribution are resampled, which depends on an estimated regression coefficient sequence. The novelty is that this estimate is actually far from the maximum likelihood estimate (MLE). The estimate is obtained by appropriately shrinking the MLE towards the origin. The amount of shrinkage is motivated by recent theories of high-dimensional MLEs. It is demonstrated that the resized bootstrap method yields valid confidence intervals in both simulated and real data examples. It is further shown that the resized bootstrap method extends to other high-dimensional generalized linear models.

EO250 Room 104 RECENT ADVANCES IN FACTOR MODELS (VIRTUAL) Chair: Sung Hoon Choi

E0518: **An assessment of the marginal predictive content of economic uncertainty indexes and business conditions predictors**
Presenter: **Yang Liu**, Rutgers University, United States
The marginal predictive content of a variety of new business conditions (BC) predictors, as well as nine economic uncertainty indexes (EUIs) constructed using these predictors, are evaluated. The predictors are defined as selected observable variables and latent factors extracted from a high dimensional macroeconomic dataset, and the EUIs are functions of predictive errors from models that incorporate these predictors. The estimation of the predictors is based on a number of extant and novel machine-learning methods that combine dimension reduction and shrinkage. When predicting 14 monthly U.S. economic series selected from 8 different groups of economic variables, the new indexes and predictors are shown to result in significant improvements in forecast accuracy relative to predictions made using benchmark models. Moreover, while the inclusion of either BC predictors or EUIs often yields forecast accuracy improvements, greater predictive gains accrue when using BC predictors with real economic activity type variables. Also, adding both BC predictors and EUIs together is particularly useful when forecasting housing market variables at short horizons.

E0652: **Identification and estimation of parameter instability in high dimensional approximate factor models**
Presenter: **Yiru Wang**, University of Pittsburgh, United States
Co-authors: Ruiqi Liu
A novel approach is introduced for estimating structural break ratios in the factor loadings of high-dimensional approximate factor models, where the breaks occur at unknown common dates and the number of factors is unknown. The method is based on the observation that the sum of the numbers of pseudo factors in the pre- and post-split subsamples is minimized when the sample is split at the structural break. By appropriately transforming these criteria using the eigenvalue ratios of the covariance matrices of the pre- and post-split subsamples, consistent estimators are derived for the structural break ratios. Notably, the framework exhibits remarkable flexibility in accommodating weak factors and can be easily extended to handle multiple breaks. A data-driven process is also introduced to determine the number of breaks. Monte Carlo simulations demonstrate the good performance of the proposed estimators. Furthermore, in an empirical analysis of the FRED-MD dataset, two structural breaks are identified around January 1983 and March 2009.

E0829: **Matrix-based prediction approach for intraday instantaneous volatility vector**
Presenter: **Sung Hoon Choi**, University of Connecticut, United States
Co-authors: Donggyu Kim
A novel method is introduced for predicting intraday instantaneous volatility based on Ito semimartingale models using high-frequency financial data. Several studies have highlighted stylized volatility time series features, such as interday auto-regressive dynamics and the intraday U-shaped pattern. To accommodate these volatility features, an interday-by-intraday instantaneous volatility matrix process is proposed that can be decomposed into low-rank conditional expected instantaneous volatility and noise matrices. To predict the low-rank conditional expected instantaneous volatility matrix, the Two-sIde Projected-PCA (TIP-PCA) procedure is proposed. Asymptotic properties of the proposed estimators are established, and a simulation study is conducted to assess the finite sample performance of the proposed prediction method. Finally, the TIP-PCA method is applied to an out-of-sample instantaneous volatility vector prediction study using high-frequency data from the S&P 500 index and 11 sector index funds.

E0669: **Window selection in FAR models with structural instabilities**
Presenter: **Antoine Djogbenou**, York University, Canada
A theory for rolling window selection is developed for generating out-of-sample forecasts using factor-augmented regression (FAR) models in the presence of structural instabilities. It shows how a rolling window can be selected by minimizing the conditional mean square forecast error (MSFE) while accounting for factor estimation uncertainty. Because the conditional MSFE is unobserved and the factors are latent, a feasible version of the criterion is proposed, and conditions under which the new method is valid are derived. A simulation experiment and an empirical application are used to document the performance of the procedure.

EO207 Room 105 APPLYING DOUBLY ROBUST METHODS TO IMPROVE FINITE POPULATION INFERENCES Chair: Lingxiao Wang

E0176: **Doubly robust inference for measuring (un)explained health disparities**
Presenter: **Yan Li**, University of Maryland, United States
A general framework for statistical inferences is established in measuring (un)explained health disparities between privileged (AG) and marginalized (DG) groups. While the Peters and Belson (PB) method is commonly employed in literature, its reliance on parametric modelling of the outcome can yield misleading results under model misspecification. To address this, an alternative method based on propensity scores (PS) adjusts the empirical distribution of the outcome variable by constructing pseudo-weights that equalize the empirical distributions of the explanatory variables across groups. However, the PS method hinges on the assumption of a valid propensity model to construct pseudo-weights. A rigorous procedure for constructing doubly robust (DR) estimators is developed to measure health disparities. DR estimators use the estimated propensity scores as well as an outcome regression model and remain consistent as long as one of the two models is correctly specified. Variance estimation is discussed under the proposed framework. Results from simulation studies show the robustness and efficiency of the proposed DR estimators as compared to existing PS and PB methods. The proposed method measures the disparity in body mass index in the national health and nutrition examination survey from 1999 to 2004.

E0312: **Model-assisted weighting methods for improving robustness and efficiency of relative risk estimation in the population**
Presenter: **Lingxiao Wang**, University of Virginia, United States
Model-assisted calibration weighting methods are widely used to improve the robustness and efficiency of the weighted estimator in survey statistics. Calibration on auxiliary variables has been shown not only to gain efficiency in estimating finite population total/mean when the auxiliary

variables are highly correlated with the outcome of interest but also to reduce bias if the calibration totals are error-free. However, for regression coefficients, such as linear association, relative risks, and hazard ratios, calibration on auxiliary variables only does not improve efficiency. Model-assisted calibration weighting methods are presented using novel auxiliary variables generated from the estimating equations for parameters of interest to obtain robust and efficient estimators. The proposed method can be applied under various data structures, such as two-phase samples and data integration. The proposed methods are applied to estimating hazard ratios of lung cancer incidence from a large-scale volunteer-based epidemiologic cohort by using the National Health Interview Survey (NHIS) as the reference.

E0610: **Doubly robust estimation for non-probability samples with modified intertwined probabilistic factors decoupling**
Presenter: **Zhan Liu**, School of Mathematics and Statistics, Hubei University, China

In recent years, non-probability samples, such as web survey samples, have become increasingly popular in many fields, but they may be subject to selection biases, which results in the difficulty of inference from them. Doubly robust (DR) estimation is one of the approaches to making inferences from non-probability samples. When many covariates are available, variable selection becomes important in DR estimation. A new DR estimator for the finite population mean is constructed, where the intertwined probabilistic factors decoupling (IPAD) and modified IPAD are used to select important variables in the propensity score model and the outcome superpopulation model, respectively. Unlike the traditional variable selection approaches, such as adaptive least absolute shrinkage and selection operator and smoothly clipped absolute deviations, IPAD and the modified IPAD not only can select important variables and estimate parameters but also can control the false discovery rate, which can produce more accurate population estimators. Asymptotic theories and variance estimation of the DR estimator with a modified IPAD are established. Results from simulation studies indicate that the proposed estimator performs well. The proposed method is applied to the analysis of the Pew Research Center data and the Behavioral Risk Factor Surveillance System data.

E0628: **Quantile regression-based data integration for combining probability and nonprobability samples**
Presenter: **Sixia Chen**, University of Oklahoma, United States
Co-authors: Emily Berg, Cindy Yu

Researchers often encounter nonprobability samples in practice, including biomedical research, business study, educational research, and other fields. Statistical analysis by using nonprobability samples without further adjustment may lead to biased results due to selection bias. Date integration has been regarded as one of the effective ways to handle nonprobability samples. It combines the information from both nonprobability samples and probability samples to reduce selection bias. Commonly used data integration methods include mass imputation, Propensity score weighting, Calibration, and Hybrid methods. The validity of those methods depends on the underlying model assumptions. To improve the robustness of model misspecification and protect the outliers, a novel quantile regression-based mass imputation method is proposed as a doubly robust method with a nonparametric estimation of the propensity score model. The proposed methods are more robust compared to some existing methods in terms of model misspecification and outliers. Asymptotic theory, including consistency, asymptotic normality, and variance estimation procedures, has been developed. The methods are further evaluated by using a Monte Carlo simulation study and one real data application.

EO069 Room 106 HIGH-DIMENSIONAL INFERENCE AND NETWORK ANALYSIS Chair: Daoji Li

E0232: **Robust knockoffs inference via coupling**
Presenter: **Lan Gao**, University of Tennessee Knoxville, United States
Co-authors: Jinchi Lv, Yingying Fan

The robustness of the model-X knockoffs framework is investigated with respect to the misspecified or estimated feature distribution. Such a goal is achieved by theoretically studying the feature selection performance of a practically implemented knockoffs algorithm, which is named the approximate knockoffs (ARK) procedure, under the measures of the false discovery rate (FDR) and family-wise error rate (FWER). The approximate knockoffs procedure only differs from the model-X knockoffs procedure in that the former uses the misspecified or estimated feature distribution. A key technique in the theoretical analyses is to couple the approximate knockoffs procedure with the model-X knockoffs procedure so that random variables in these two procedures can be close in realizations. It is proven that if such a coupled model-X knockoffs procedure exists, the approximate knockoffs procedure can achieve the asymptotic FDR or FWER control at the target level. Three specific constructions of such coupled model-X knockoff variables are showcased, verifying their existence and justifying the robustness of the model-X knockoffs framework.

E0242: **Integrative conformal p-values for out-of-distribution testing with labeled outliers**
Presenter: **Ziyi Liang**, University of Southern California, United States

The focus is on presenting a conformal inference method for out-of-distribution testing that leverages side information from labelled outliers, which are commonly underutilized or even discarded by conventional conformal p-values. The solution is practical and blends inductive and transductive inference strategies to adaptively weight conformal p-values while also automatically leveraging the most powerful model from a collection of one-class and binary classifiers. Further, the approach leads to rigorous false discovery rate control in multiple tests when combined with a conditional calibration strategy. Extensive numerical simulations show the proposed method outperforms existing approaches.

E1099: **On partial envelope approach for modeling spatial-temporally dependent data**
Presenter: **Wenbo Wu**, University of Texas at San Antonio, United States

In the new era of big data, modeling multivariate spatial-temporally dependent data is a challenging task due to not only the high dimensionality of the features but also complex spatial-temporal associations among the observations across different locations and time points. To improve the estimation efficiency, a spatial-temporal partial envelope model is proposed, which is parsimonious and effective in modeling high-dimensional spatial-temporal data. The partial envelope model is proposed under a linear coregionalization model framework, which allows for a heterogeneous spatial-temporal covariance structure for different components of the response vector. The maximum likelihood estimator for the proposed model can be obtained through a Grassmann manifold optimization. An asymptotic result is obtained for the estimator, and thorough simulation studies are conducted to demonstrate the soundness and effectiveness of the proposed method. The proposed model is also applied to analyze the crowdsourcing weather data collected from personal weather stations in the city of San Antonio, TX, of the United States.

E1117: **A comparison of methods for externally validating the Kidney Donor Risk Index in the UK kidney transplant population**
Presenter: **Yinghui Wei**, Plymouth University, United Kingdom
Co-authors: Stephanie Riley, Andrew Connor, Wai-yee Tse

Transplantation represents the optimal treatment for many patients with end-stage kidney disease. The Kidney Donor Risk Index (KDRI) was developed to predict graft failure following kidney transplantation. The survival process following transplantation consists of semi-competing events, where recipient death precludes graft failure but not vice-versa. We sought to externally validate the KDRI in the UK kidney transplant population, and assess whether validation under a competing risks framework had an impact on predictive performance. Additionally, we updated the KDRI using data from the United Kingdom to explore whether this improved the predictive performance. Using data from recipients of deceased donor single kidney-only transplants, held by NHS Blood and Transplant, we externally validated the KDRI. Our outcomes of interest were one- and five-year graft failure. Considering the semi-competing events, we modelled the outcome in two ways: censoring the recipient at the time of death, and modelling death as a competing event. Cox proportional hazard models were used to validate the KDRI when censoring for death, and

cause-specific Cox models were used to account for death as a competing event. KDRI performance was assessed by discrimination, calibration, and overall accuracy of predictions.

EO090 Room 108 STATISTICAL LEARNING FROM COMPLEX DATA Chair: Fengrong Wei

E0240: **Holding-linked network in mutual fund and the predictability of performance**
Presenter: **Fengrong Wei**, University of West Georgia, United States
Co-authors: Weizhong Tian

A holding-linked network of mutual funds is used, measured by the similarity between funds portfolios, to examine the network predictability of fund performance and flows. Using the new network method, evidence of significant predictability between funds with similar holdings is found. The predictability persists for three to six months for alternative performance measures and at least twelve months for fund flows. In addition, a long-short strategy based on these holding links yields a significant annual alpha of about 4.5 per cent. These findings reflect the similar underlying drivers of funds portfolio holdings and show the persistent prediction of fund performance and flows by the holding linked network.

E0515: **Bayesian single and multiple index models with additive regression trees**
Presenter: **Ruijin Lu**, Washington University in St. Louis School of Medicine, United States

In analyzing data of environmental mixtures, single (SIM) and multiple index models (MIM) are powerful tools given their nonparametric links and interpretable index coefficients. A Bayesian additive regression tree (BART) perspective is taken for these models, and variable selections are considered, particularly the selection of exposures in the indices. The challenge of applying BART to SIM/MIM is tackled by using a sigmoid gating function in place of the binary routine at each splitting node and that of variable selection by placing a sparsity-inducing Dirichlet hyperprior. The performance of the proposed approach is examined by conducting extensive simulations and applying them to commonly used benchmark data sets. For real data application, the link between birth weight and exposures to environmental pollutants, dietary intakes, and physical activities during pregnancy is investigated using data from the NICHD Fetal Growth Study.

E0594: **Learning from heterogeneous data with stick-breaking variational autoencoder**
Presenter: **Jaeyoung Lee**, Virginia Tech, United States
Co-authors: Hongxiao Zhu

Modern data often exhibit intricate heterogeneous structures caused by diverse data sources, subpopulations, nested experimental designs, or other unknown factors. Conventional statistical models are inadequate for such data as they frequently overlook the inherent heterogeneous structure. An innovative statistical learning framework is introduced to capture latent heterogeneous structures among samples while facilitating prediction and association analysis. Specifically, the framework employs a stick-breaking variational autoencoder to characterize the heterogeneous data structure and link the latent stick-breaking process with a response variable. The advantages of modeling latent heterogeneous structures are illustrated through simulations and a real data application involving brain tumor images.

E0680: **Discovering dependence structure of transcription factors based on a nonhomogeneous Poisson process model**
Presenter: **Xiaowei Wu**, Virginia Tech, United States
Co-authors: Hongxiao Zhu

Transcription factors (TFs) are proteins that bind to DNA sequences, playing the central role of regulating gene expressions. Discovering the dependence structure among TFs is a critical problem in statistical genetics, and it advances the understanding of the underlying epigenetic regulatory mechanism. Based on a nonhomogeneous Poisson process (NHPP) model, a two-step approach is developed for detecting pairwise TF relations, whether they bind to DNA sequences cooperatively or independently. The key of the proposed method lies in the goodness-of-fit test of the NHPP model to recurrent TF binding events, which is an important but often neglected issue in modeling complex data arising from real-world problems. Simulation studies show that the proposed method provides a powerful test for interactions between TFs. The method is applied to analyze ChIP-seq data of 14 TFs collected in recent mouse embryonic stem cell research. Findings provide new insights into the orchestration of TFs in gene regulation processes.

EO084 Room 109 INNOVATIVE STATISTICAL LEARNING METHODS FOR COMPLEX DATA Chair: Yeying Zhu

E0187: **Jackknife empirical likelihood for infinite-order U-statistics with applications to ensemble predictions**
Presenter: **Qing Wang**, Wellesley College, United States
Co-authors: Yichuan Zhao, Ting Zhang

Infinite-order U-statistics have abundant practical applications, such as subsampling-based ensemble methods. However, due to the dependence of the degree kn on the sample size n, theories and results under the traditional fixed-degree U-statistic framework cannot be applied directly. In particular, there has yet to be a promising method to estimate the variance of an infinite-order U-statistic, especially when kn is not of a much lower order of n. The jackknife empirical likelihood methodology is extended to infinite-order U-statistics. It is proven that the fundamental framework of jackknife empirical likelihood still holds under some regularity conditions. In the context of subsampling-based ensemble methods, the performance of the proposed methodology is evaluated to construct confidence intervals for ensemble predictions through simulation studies. The proposal yields superior results compared to existing methods across various settings. In addition, it gives a coverage probability that is approaching the nominal level as the number of trees used to build the ensemble increases.

E0627: **Estimation of multiple large covariance matrices and its application to high-dimensional quadratic discriminant analysis**
Presenter: **Yingli Qin**, University of Waterloo, Canada
Co-authors: Mu Zhu, Liyuan Zheng, Yilei Wu, Weiming Li

When estimating covariance matrices for data from multiple related categories, such as different subtypes of a particular disease, it is possible that these covariance matrices may exhibit shared structural components. The precision matrix (the inverse of the covariance matrix) of each category is assumed to be decomposed into a common diagonal component, a common low-rank component and a category-specific low-rank component. This decomposition can be motivated by a factor model, in which the effects of some latent factors are common across all categories while others are specific to individual categories. A method to jointly estimate these precision matrices is proposed, thereby inferring the covariance matrices as well, starting with an estimation of the number of factors. The approach incorporates a complexity penalty to promote the imposition of low-rank structures. Under moderate conditions, The consistency of the estimators is established. Furthermore, these estimators are applied to formulate a high-dimensional quadratic discriminant analysis rule. Its convergence rate is established for the classification error. Finally, the method is illustrated through numerical examples.

E0657: **Mediation analysis with latent factors using simultaneous group-wise and parameter-wise penalization**
Presenter: **Xizhen Cai**, Williams College, United States
Co-authors: Qing Wang, Yeying Zhu

Mediation analysis aims to uncover the underlying mechanism of how an exposure variable affects the outcome of interest through one or more

mediating variables. In the event that the number of candidate mediators is large, variable selection or dimension reduction techniques are often utilized to reduce the dimension of the initial set of mediators. The proposed latent variable approach is discussed using sparse factor analysis with both group-wise and parameter-wise penalization to remove irrelevant candidate mediators and estimate the latent factors simultaneously. After the low-dimensional latent mediating factors are obtained, the direct and indirect effects can be estimated and tested from a multivariate mediation model. To demonstrate the practical applications of the proposed methodology, real-world applications are discussed with a weight behavior dataset and an environmental dataset.

E0873: **Polya trees for survival data**
Presenter: **Liqun Diao**, University of Waterloo, Canada
Polya trees are commonly used as priors in nonparametric Bayesian analysis. The purpose is to discuss approaches for utilizing Polya trees to characterize the distribution of time-to-event data, which may be subjected to different forms of censoring, such as right-censoring or interval-censoring. The discussion will cover different aspects of Polya trees, including partitions, prior strength, and choices of prior distributions. Comparisons of the proposed methods to existing approaches for estimating survival probabilities are provided in both simulated settings and through applications to real datasets. It is shown that the proposed methods either improve upon or remain competitive with existing nonparametric estimation methods.

EO075 Room 110 SPACE FILLING DESIGNS AND FACTORIAL DESIGNS Chair: Wei Zheng

E0358: **Construction of maximin distance Latin hypercube designs via good lattice point sets**
Presenter: **Xueru Zhang**, Purdue University, United States
Space-filling Latin hypercube designs have found widespread applications in computer experiments, yet their design construction poses significant challenges. The design constructed through algebraic methods is limited to very restricted numbers of runs and factors, whereas those designs generated by algorithmic searches are limited to small numbers of runs and factors. To address these limitations, an approach is proposed for producing space-filling Latin hypercube designs that can accommodate flexible numbers of runs and factors. The proposed approach is hybrid in nature, incorporating an algebraic method and its corresponding algorithm. The algebraic method, built on good lattice point sets and level permutation techniques, applies to any run size and flexible numbers of factors. The proposed algorithmic search can further handle any number of factors, especially those not covered by the algebraic method. A theoretical analysis of optimality is provided for the algebraic component. Numerical studies demonstrate the superior L_p-distance properties of the proposed designs. Furthermore, it is demonstrated that the proposed designs exhibit good column-orthogonality and projection uniformity as well.

E0410: **On construction of nonregular two-level factorial designs with maximum generalized resolutions**
Presenter: **Chenlu Shi**, Colorado State University, United States
Co-authors: Boxin Tang
The generalized resolution was introduced and justified as a criterion for selecting nonregular factorial designs. Although extensive research has been conducted on other aspects of nonregular designs, few works have investigated the construction of nonregular designs with maximum generalized resolutions. To date, the knowledge of nonregular designs with maximum generalized resolutions is predominantly computational, except for very few theoretical results. Lower bounds on relevant J-characteristics are derived, and the construction results are presented. With the assistance of the lower bounds, many of the constructed designs are shown to have maximum generalized resolutions.

E0497: **Fast approximation of Shapley values via design methods**
Presenter: **Zheng Zhou**, Nankai University, China
Shapley value is a well-known concept in cooperative game theory that provides a fair way to distribute revenues or costs to each player. Recently, it has been widely applied in various fields, such as data science, marketing, and genetics. However, the computation of the Shapley value is an NP-hard problem. For a cooperative game with n players, calculating Shapley values for all players requires calculating the value function for 2^n different coalitions, which makes it infeasible for a large n. A remarkable connection is first identified between cooperative games and two-level factorial designs, which is a familiar concept from the field of design of experiments (DOE). This inspires the proposal of fast approximation approaches for Shapley values based on design methods. Multiple simulations and real case examples demonstrate that, with equivalent computational cost, the method provides significantly more accurate approximations compared with several popular methods.

E0540: **Space filling designs on Riemannian manifolds**
Presenter: **Xiangshun Kong**, Beijing Institute of Technology, China
The aim is to propose a new approach to generating space-filling designs over Riemannian manifolds by using a Hilbert curve. Different from ordinary Euclidean spaces, a novel transformation is constructed to link the uniform distribution over a Riemannian manifold and that over its parameter space. With the help of this transformation, the uniformity of the design points in the sense of the Riemannian volume measure can be guaranteed by the intrinsic measure preserving the property of the Hilbert curve. It has been proved that these generated designs are not only asymptotically optimal under minimax and maximin distance criteria but also perform well in minimizing the Wasserstein distance from the target distribution and controlling estimation errors in numerical integration. Furthermore, an efficient algorithm is developed for the numerical generation of these space-filling designs. Compared with the existing methods, the advantages of the new approach are verified through numerical simulations.

EO208 Room 111 CONFORMAL PREDICTION AND INFERENCE Chair: Yuan Zhang

E0173: **Model-free selective inference: Selecting trusted decisions from black boxes**
Presenter: **Ying Jin**, Stanford University, United States
Co-authors: Emmanuel Candes
Many decision-making and scientific discovery processes aim to identify candidates whose unknown outcomes satisfy a desired property, e.g., drugs with high binding affinities to a disease target in drug discovery. While predictive AI excels in accelerating such processes, ensuring the reliability of AI remains a challenge in critical situations. In the example of drug discovery, any false lead in shortlisted drug candidates may incur substantial costs in later stages, which harms its initial promise. A model-free selective inference framework is introduced to identify candidates whose unobserved outcomes exceed user-specified values with the assistance of any prediction model. The framework controls the average proportion of false positives (FDR) among the selected set of units without any modelling assumptions on the data distribution. In addition, new ideas on dealing with distribution shifts are discussed between training and new samples, a scenario often encountered in applications. It is shown via several empirical studies in drug discovery that the methods help scientists narrow down the drug candidates to a manageable size of promising ones with finite-sample error control.

E0202: **Conformalized matrix completion**
Presenter: **Yu Gui**, University of Chicago, United States
Co-authors: Rina Foygel Barber, Cong Ma
Matrix completion aims to estimate missing entries in a data matrix using the assumption of a low-complexity structure (e.g., low rank) to make imputation possible. While many effective estimation algorithms exist in the literature, uncertainty quantification for this problem has proved challenging, and existing methods are extremely sensitive to model misspecification. A distribution-free method is proposed for predictive inference in the matrix completion problem. The method adapts the framework of conformal prediction, which provides confidence intervals with guaranteed distribution-free validity in the regression setting, to the matrix completion problem. The resulting method, conformalized matrix completion (CMC), offers provable predictive coverage regardless of the accuracy of the low-rank model. Empirical results on simulated and real data demonstrate that CMC is robust to model misspecification while matching the performance of existing model-based methods when the model is correct.

E0489: **A conformal test of linear models via permutation-augmented regressions**
Presenter: **Leying Guan**, Yale University, United States
Permutation tests are widely recognized as robust alternatives to tests based on normal theory. Random permutation tests have been frequently employed to assess the significance of variables in linear models. Despite their widespread use, existing random permutation tests lack finite-sample and assumption-free guarantees for controlling type I errors in partial correlation tests. To address this ongoing challenge, a conformal test is developed through permutation-augmented regressions, referred to as PALMRT. PALMRT not only achieves power competitive with conventional methods but also provides reliable control of type I errors at no more than two alpha, given any targeted level alpha, for arbitrarily fixed designs and error distributions. This, through extensive simulations, is confirmed. Compared to the cyclic permutation test (CPT) and residual permutation test (RPT), which also offer theoretical guarantees, PALMRT does not compromise as much on power or set stringent requirements on the sample size, making it suitable for diverse biomedical applications. The differences in a long-Covid study are further illustrated when PALMRT validated key findings previously identified using the t-test after multiple corrections, while both CPT and RPT suffered from a drastic loss of power and failed to identify any discoveries. PALMRT is endorsed as a robust and practical hypothesis test in scientific research for its superior error control, power preservation, and simplicity.

E0674: **Conformal prediction for fragmented functional data**
Presenter: **Fangyi Wang**, The Ohio State University, United States
Co-authors: Sebastian Kurtek, Yuan Zhang
Predicting missing segments in incomplete curves is a significant challenge in functional data analysis due to unknown nuisance transformations among different curves. Existing methods typically rely on correct (parametric) model specifications and often involve computationally intensive estimation procedures. A very different approach, using conformal prediction, is proposed to tackle this problem. Applying conformal prediction to fragmented functional data is highly non-trivial due to the lack of naturally defined predictor and response variables. These variables are constructed from given complete functions, and the downstream analysis is carefully designed such that exchangeability is preserved, even in the presence of unknown nuisance transformations. Based on a neighborhood smoothing algorithm, various types of pointwise prediction bands can be produced. The method is simple, easy to implement, and supported by finite sample theoretical guarantees under rather weak assumptions. It also computes much faster than existing methods and allows straightforward parallelization. Extensive numerical studies and real-world examples clearly demonstrate the effectiveness and practical utility of the approach.

EO053 Room 212 STATISTICAL AND/OR PHARMACOMETRIC CONSIDERATIONS IN DRUG DEVELOPMENT Chair: Yisheng Li

E0648: **DEMO: Dose exploration, monitoring, and optimization using a biological mediator for clinical outcomes**
Presenter: **Ruitao Lin**, The University of Texas MD Anderson Cancer Center, United States
Phase 1-2 designs provide a methodological advance over phase 1 designs for dose finding by using both clinical response and toxicity. A phase 1-2 trial still may fail to select a truly optimal dose. Early response is not a perfect surrogate for long-term therapeutic success. To address this problem, a generalized phase 1-2 design first uses a phase 1-2 design component to identify a set of candidate doses, adaptively randomizes patients among the candidates, and, after a longer follow-up, selects a dose to maximize long-term success rate. This paradigm is extended by proposing a design that exploits an early treatment-related, real-valued biological outcome, such as pharmacodynamic activity or an immunological effect, that may act as a mediator between dose and clinical outcomes, including tumor response, toxicity, and survival time. Multivariate dose-outcome models are assumed to include effects appearing in causal pathways from dose to the clinical outcomes. Bayesian model selection is used to identify and eliminate biologically inactive doses. At the end of the trial, a therapeutically optimal dose is chosen from the set of doses that are acceptably safe, clinically effective, and biologically active to maximize restricted mean survival time. Results of a simulation study show that the proposed design may provide substantial improvements over designs that ignore the biological variable.

E0673: **SUDO: A Bayesian subgroup-specific utility-based dose testing-optimization design for multi-dose randomized trials**
Presenter: **Fangrong Yan**, China Pharmaceutical University, China
Phase II basket trials are increasingly adopted, as they enable the concurrent evaluation of multiple tumor types, thereby expediting the drug development process, especially for oncology treatments. With growing evidence supporting the promising efficacy of lower doses for many novel agents, and in line with the FDA's project optimus, a trending practice in oncology phase IIA studies is to combine routine preliminary testing of treatment effects and dose optimization with investigations of multiple doses in a single trial. A novel Bayesian adaptive design is proposed for testing and optimizing subgroup-specific doses in multi-dose randomized basket trials. To address potential heterogeneity between subgroups, the Bayesian model averaging approach is utilized to adaptively cluster predefined patient subgroups. A Bayesian hierarchical dynamic linear model is developed to facilitate efficient information sharing across multiple doses and within specific subgroup clusters. Under the Bayesian inference framework, proof of concept for the treatment effect in each subgroup can be established, and the subgroup-specific optimal dose can be identified based on a utility function that quantifies the trade-off between toxicity and efficacy. Extensive simulation studies are conducted to evaluate the operating characteristics of the proposed design. The results demonstrate its favorable performance across various scenarios.

E0802: **Using modeling and simulation to evaluate response variation and optimal dose in clinical development**
Presenter: **Yanguang Cao**, University of North Carolina at Chapel Hill, United States
Project Optimus is an initiative by the US FDA aimed at revolutionizing the approach to dose selection and optimization in the development of oncology drugs. The challenges and opportunities associated with selecting the optimal doses are explored in the early phases of drug development, especially when faced with limited data and a small patient cohort. Traditional exposure-response (E-R) analyses in drug development are primarily conducted to identify therapeutic doses or subpopulations of patients with distinct response/safety profiles, usually for labeling purposes rather than to pinpoint the optimal doses. However, the scarcity of data and information complicates robust E-R analyses, introducing significant uncertainty in our selection of optimal doses. Therefore, integrating Bayesian methods into conventional E-R analyses promises considerable benefits. The application of these Bayesian approaches and using modeling and simulation techniques to realize the potential of Project Optimus are briefly discussed.

E0816: **Graphormer supervised de novo protein design method and function validation**
Presenter: **Ting Wei**, Shanghai Jiao Tong Univeristy, China

Protein design is central to nearly all protein engineering problems, as it can enable the creation of proteins with new biological functions, such as improving the catalytic efficiency of enzymes. One key facet of protein design, fixed-backbone protein sequence design, seeks to design new sequences that will conform to a prescribed protein backbone structure. Nonetheless, existing sequence design methods present limitations, such as low sequence diversity and shortcomings in experimental validation of the designed functional proteins. To improve these limitations, the Graphormer-based protein design (GPD) model is initially developed. This model utilizes the transformer on a graph-based representation of 3D protein structures and incorporates Gaussian noise and a sequence random mask to node features, thereby enhancing sequence recovery and diversity. The performance of the GPD model was significantly better than that of the state-of-the-art ProteinMPNN model on multiple independent tests, especially for sequence diversity. GPD is employed to design CalB hydrolase, and nine artificially designed CalB proteins are generated. The results show significant improvement in the catalytic activity, which is 1.7 times higher than the CalB wild type. Thus, the GPD method could be used for the de novo design of industrial enzymes and protein drugs with specific functions.

EO112 Room 202 RECENT ADVANCES IN STATISTICAL METHODS AND THEORY Chair: Mengyu Xu

E0253: **Asymptotics of sample tail autocorrelations for tail dependent time series: Phase transition and visualization**
Presenter: **Ting Zhang**, University of Georgia, United States

An asymptotic theory on sample tail autocorrelations of time series data is developed that can exhibit serial dependence in both tail and non-tail regions. Unlike the traditional autocorrelation function, the study of tail autocorrelations requires a double asymptotic scheme to capture the tail phenomena, and the results do not impose any restriction on the dependence structure in non-tail regions and allow processes that are not necessarily strong mixing. The asymptotic theory indicates a phase transition phenomenon for sample tail autocorrelations, whose asymptotic behavior, including the convergence rate, can transit from one phase to the other when the lag index moves past the point beyond which serial tail dependence vanishes. The phase transition fills the gap of existing research on tail autocorrelations and can be used to construct the lines of significance, in analogy to the traditional autocorrelation plot, when visualizing sample tail autocorrelations to assess the existence of serial tail dependence or to identify the maximal lag of tail dependence.

E0661: **Accounting for network noise in graph-guided Bayesian modeling of structured high-dimensional data**
Presenter: **Wenrui Li**, University of Pennsylvania, United States
Co-authors: Changgee Chang, Suprateek Kundu, Qi Long

There is a growing body of literature on knowledge-guided statistical learning methods for analysis of structured high-dimensional data (such as genomic and transcriptomic data) that can incorporate knowledge of underlying networks derived from functional genomics and functional proteomics. These methods have been shown to improve variable selection and prediction accuracy and yield more interpretable results. However, these methods typically use graphs extracted from existing databases or rely on subject matter expertise, which are known to be incomplete and may contain false edges. To address this gap, a graph-guided Bayesian modeling framework is proposed to account for network noise in regression models involving structured high-dimensional predictors. Specifically, two sources of network information are used, including the noisy graph extracted from existing databases and the estimated graph from observed predictors in the dataset at hand, to inform the model for the true underlying network via a latent scale modeling framework. This model is coupled with the Bayesian regression model with structured high-dimensional predictors involving an adaptive structured shrinkage prior. An efficient Markov chain Monte Carlo algorithm is developed for posterior sampling. The advantages of the method are demonstrated over existing methods in simulations and through analyses of a genomics dataset and another proteomics dataset for Alzheimer's disease.

E0684: **Whittle estimation based on the extremal spectral density of a heavy-tailed random field**
Presenter: **Yuwei Zhao**, Xián Jiaotong-Liverpool University, China

A strictly stationary random field is considered on the two-dimensional integer lattice with regularly varying marginal and finite-dimensional distributions. Exploiting the regular variation, the spatial extremogram is defined, which takes into account only the largest values in the random field. This extremogram is a spatial autocovariance function. The corresponding extremal spectral density and its estimator are defined as the extremal periodogram. Based on the extremal periodogram, the Whittle estimator is considered for suitable classes of parametric random fields, including the Brown-Resnick random field and regularly varying max-moving averages.

E0742: **Joint graphical lasso with regularized aggregation for high-dimensional time series with long-memory**
Presenter: **Jongik Chung**, University of Central Florida, United States
Co-authors: Qihu Zhang, Cheolwoo Park

The purpose is to outline methods for estimating multiple precision matrices for high-dimensional long-memory time series within the framework of Gaussian graphical models, particularly focusing on analyzing functional magnetic resonance imaging (fMRI) data collected from multiple subjects. The aim is to estimate individual brain networks and a collective structure representing a group of subjects. A method is proposed that utilizes regularized aggregation to simultaneously estimate individual and group precision matrices, assigning varying weights to each individual based on their outlier status within the group. The convergence rates of the precision matrix estimators are examined across different norms and expectations, evaluating their performance under sub-Gaussian and heavy-tailed assumptions. The efficacy of the methods is demonstrated through simulations and real fMRI data.

EO188 Room 204 ADVANCES IN MARKOV CHAIN MONTE CARLO Chair: Qian Qin

E0682: **Multivariate strong invariance principle and uncertainty assessment for time in-homogeneous cyclic MCMC samplers**
Presenter: **Haoxiang Li**, University of Minnesota, Twin Cities, United States
Co-authors: Qian Qin

Time in-homogeneous cyclic Markov chain Monte Carlo (MCMC) samplers, including deterministic scan Gibbs samplers and Metropolis within Gibbs samplers, are extensively used for sampling from multi-dimensional distributions. A multivariate strong invariance principle (SIP) is established for Markov chains associated with these samplers. The rate of this SIP essentially aligns with the tightest rate available for time-homogeneous Markov chains. The SIP implies the strong law of large numbers (SLLN) and the central limit theorem (CLT) and plays an essential role in uncertainty assessments. Using the SIP, conditions are given under which the multivariate batch means estimator for estimating the covariance matrix in the multivariate CLT is strongly consistent. Additionally, conditions are provided for a multivariate fixed volume sequential termination rule, which is associated with the concept of effective sample size (ESS), to be asymptotically valid. The uncertainty assessment tools are demonstrated through various numerical experiments.

E0364: **Importance tempering of Markov chain Monte Carlo methods**
Presenter: **Quan Zhou**, Texas A&M University, United States
Co-authors: Aaron Smith, Guanxun Li
Informed importance tempering (IIT) is an easy-to-implement MCMC algorithm that can be seen as an extension of the familiar Metropolis-Hastings algorithm with the special feature that informed proposals are always accepted and which was shown to converge much more quickly in some common circumstances. A new, comprehensive guide is developed for the use of IIT in many situations. First, two IIT schemes are proposed that run faster than existing informed MCMC methods on discrete spaces by not requiring the posterior evaluation of all neighbouring states. Second, IIT is integrated with other MCMC techniques, including simulated tempering, pseudo-marginal and multiple-try methods (on general state spaces), which have been conventionally implemented as Metropolis-Hastings schemes and can suffer from low acceptance rates. The use of IIT allows to always accept proposals and brings about new opportunities for optimizing the sampler, which is not possible under the Metropolis-Hastings framework. Numerical examples illustrating the findings are provided for each proposed algorithm, and a general theory on the complexity of IIT methods is developed.

E0389: **Spectral gap bounds for reversible hybrid Gibbs chains**
Presenter: **Nianqiao Ju**, Purdue University, United States
Co-authors: Qian Qin, Guanyang Wang
Hybrid Gibbs samplers represent a prominent class of approximated Gibbs algorithms that utilize Markov chains to approximate conditional distributions, with the Metropolis-within-Gibbs algorithm standing out as a well-known example. Despite their widespread use in both statistical and non-statistical applications, very little is known about their convergence properties. Novel methods are introduced to establish bounds on the convergence rates of hybrid Gibbs samplers. In particular, the convergence characteristics of hybrid random-scan Gibbs and data augmentation algorithms are examined. The analysis reveals that the absolute spectral gap of a reversible hybrid chain can be bounded based on the absolute spectral gap of the exact Gibbs chain and the absolute spectral gaps of the Markov chains employed for conditional distribution approximations. The new techniques are applied to four algorithms: a random-scan Metropolis-within-Gibbs sampler, a hybrid proximal sampler, random-scan Gibbs samplers with block updates, and a hybrid slice sampler.

E0361: **MCMC when you do not want to evaluate the target distribution**
Presenter: **Guanyang Wang**, Rutgers University, United States
Co-authors: Wei Yuan
In sampling tasks, it is common for target distributions to be known up to a normalizing constant. However, in numerous situations, evaluating even the unnormalized distribution proves to be costly or infeasible. This issue arises in scenarios such as sampling from the Bayesian posterior for large datasets and the 'doubly intractable' distributions. The aim is to introduce a unified framework that includes various MCMC algorithms, including several minibatch MCMC algorithms and the exchange algorithm. This framework not only simplifies the theoretical analysis of existing algorithms but also leads to the development of new, more efficient algorithms.

EO244 Room 207 DESIGN AND ANALYSIS FOR EVALUATING CAUSAL, MODERATION, AND MEDIATION EFFECTS Chair: Xu Qin

E0831: **Optimal sample size planning for longitudinal multisite experiments to investigate the main and moderator effects**
Presenter: **Wei Li**, University of Florida, United States
Co-authors: Spyros Konstantopoulos, Zuchao Shen
Longitudinal multisite experimental designs are commonly employed in educational interventions, where, for example, students from the same schools are randomly assigned to either a treatment or control group and subsequently followed and measured over time. One objective of longitudinal studies is to examine how treatment effects evolve over time. Additionally, educational researchers are interested in assessing whether changes in treatment effects vary among subgroups of students or schools. These student and school characteristics, often referred to as moderators, can be investigated through interaction analyses between the treatment and specific student or school characteristics. A crucial consideration in designing longitudinal experiments is determining the sample size allocation across levels and treatment conditions to ensure sufficient power to detect the effect of interest. Researchers typically plan their longitudinal studies with budget constraints in mind, as different sampling plans under the same budget can yield varying levels of statistical power. The contribution to the literature is that it provides optimal sample size computation methods for three-level longitudinal multisite experiments to explore main and moderator effects and implements these methods into an R package and a Shiny App to assist applied researchers in planning longitudinal experiments.

E0833: **Stochastic noncompliance and endogenous confounding in evaluating a multi-phase treatment: Multi-site IV as a solution**
Presenter: **Guanglei Hong**, University of Chicago, United States
The average cumulative effect of a multi-phase treatment sequence is hard to identify due to stochastic noncompliance and endogenous confounding. Despite the initial randomization of treatment assignment, individual responses to the phase-1 treatment may predict non-compliant behaviors in the subsequent phase, thereby confounding the effect of the phase-2 treatment on the outcome. Principal stratification resorts to a deterministic framework, often a mismatch with reality. In contrast, non-compliant behaviors are allowed to be influenced by stochastic random events. Extending the instrumental variable (IV) method to a multi-site randomized trial for evaluating a multi-phase treatment sequence, the approach requires neither sequential ignorability nor exclusion restriction. The key is to obtain the conditional distribution of the potential intermediate outcome under the counterfactual phase-1 treatment condition as a function of not just baseline covariates but also the observed intermediate outcome under the actual phase-1 treatment condition for each individual. The cumulative treatment effect is further allowed to depend on these potential/counterfactual intermediate outcome values. Reanalyzing the well-known Project STAR data, a multi-site randomized trial for studying class size reduction, the average impact of receiving two years of instruction is evaluated in a small-size class as opposed to a regular-size class on student achievement.

E0837: **A causal investigation of heterogeneity in mediation mechanisms in multisite randomized trials**
Presenter: **Xu Qin**, University of Pittsburgh, United States
Multisite randomized trials have been pervasive in the past three decades. The importance of investigating the variation in the total impact of an intervention has become increasingly valued. An intervention may generate heterogeneous impacts due to natural variations in participant characteristics, context, and local implementation. Important research questions include whether the intervention impact is generalizable across individuals and contexts, for whom and under what contexts the intervention is effective, and why. To advance this line of research, a method is developed to assess the mediation mechanism underlying the total impact of the intervention in multisite randomized trials and how it varies by individual and contextual factors. The findings may help practitioners improve and tailor intervention designs and implementations for different individuals and contexts. The method is evaluated through comprehensive Monte Carlo simulations. It is also applied to the National Study of Learning Mindsets to evaluate the mediation mechanism underlying the impact of a growth mindset intervention on math performance and its heterogeneity.

E0856: **Heterogeneous causal mediation analysis with Bayesian additive regression trees**
Presenter: **Chen Liu**, University of Pittsburgh, United States
Co-authors: Xu Qin, Jiebiao Wang
Causal mediation analysis can help explain how an exposure affects an outcome. The mediation effects are often heterogeneous based on individual characteristics, but most existing methods ignore this heterogeneity and estimate the population average effects. To address this gap, a heterogeneous causal mediation analysis method is developed using Bayesian regression tree ensembles. Distinct from traditional methods, the approach captures complex non-linear interactions and heterogeneous effects in mediation processes more flexibly, offering a refined understanding of the heterogeneity of causal mechanisms. By sampling from the posterior trees of mediator and outcome models, rigorous credible intervals are obtained for causal mediation effects. Partial dependent plots are also used to illustrate which moderators play more important roles and how each effect changes with a moderator. Utilizing simulated datasets, the superiority of the approach is demonstrated in the accurate estimation and inference of heterogeneous mediation effects, especially in scenarios characterized by non-linear relationships and interaction effects. The proposed method is applied to estimate heterogeneous mediation effects in genetic mechanisms of Alzheimer's disease.

EO221 Room 209 STATISTICAL INNOVATIONS FOR COMPLEX DATA ANALYSIS IN BIOMEDICAL RESEARCH Chair: Kaiqiong Zhao

E0261: **Minor issues escalated to critical levels in large samples: A permutation-based fix**
Presenter: **Xuekui Zhang**, University of Victoria, Canada
In the big data era, the need to reevaluate traditional statistical methods is paramount due to the challenges posed by vast datasets. While larger samples theoretically enhance accuracy and hypothesis testing power without increasing false positives, practical concerns about inflated Type-I errors persist. The prevalent belief is that larger samples can uncover subtle effects, necessitating dual consideration of p-value and effect size. Yet, the reliability of p-values from large samples remains debated. DE analysis of single-cell genomic data often identifies thousands of DE genes using adjusted p-values, and subjective log-fold change thresholds must be used to filter them. Since larger fold changes always have smaller p-values, p-values are nearly obsolete in decision-making. It is warned that larger samples can exacerbate minor issues into significant errors, leading to false conclusions. Through the simulation study, growing sample sizes are demonstrated to amplify issues arising from two commonly encountered violations of model assumptions in real-world data and lead to incorrect decisions. This underscores the need for vigilant analytical approaches in the era of big data. In response, a permutation-based test is suggested to counterbalance the effects of sample size and assumption discrepancies by neutralizing them between actual and permuted data.

E0969: **Changepoint detection in the variability of multivariate and functional data**
Presenter: **Kelly Ramsay**, York University, Canada
Co-authors: Shojaeddin Chenouri
The problem of robustly detecting changepoints is considered in the variability of a sequence of independent multivariate functions and vectors. Novel changepoint procedures, called the functional and multivariate Kruskal-Wallis for covariance (FKWC and MKWC) changepoint procedures, are presented based on rank statistics and data depth. The MKWC and FKWC changepoint procedures allow the user to test for at most one changepoint or an epidemic period or to estimate the number and locations of an unknown amount of changepoints in the data. It is shown that when the "signal-to-noise" ratio is bounded below, the changepoint estimates produced by the MKWC and FKWC procedures attain the minimax localization rate for detecting general changes in distribution in the univariate setting. The behavior of the proposed test statistics is also provided for the AMOC and epidemic setting under the null hypothesis, and, as a simple consequence of the main result, these tests are consistent.

E0654: **Addressing dispersion in mis-measured multivariate binomial outcomes: Analyzing bisulfite sequencing data**
Presenter: **Kaiqiong Zhao**, York University, Canada
Motivated by a DNA methylation application, the purpose is tackling fitting and inferring a multivariate binomial regression model for outcomes that are contaminated by errors and exhibit extra-parametric variations, also known as dispersion. While dispersion in univariate binomial regression has been extensively studied, addressing dispersion in the context of multivariate outcomes remains a complex and relatively unexplored task. The complexity arises from a noteworthy data characteristic observed in our motivating dataset: non-constant yet correlated dispersion across outcomes. To address this challenge and account for possible measurement error, a novel hierarchical quasi-binomial varying coefficient mixed model is proposed, which enables flexible dispersion patterns through a combination of additive and multiplicative dispersion components. To maximize the Laplace-approximated quasi-likelihood of the model, a specialized two-stage EM algorithm is further developed. Simulations demonstrated that the approach yields accurate inference for smooth covariate effects and exhibits excellent power in detecting non-zero effects. The proposed method is also applied to investigate the association between genome-wide whole blood DNA methylation and levels of ACPA, a preclinical marker for rheumatoid arthritis (RA). The analysis highlights important insights into RA risk factors. The method is implemented in the R Bioconductor package called "SOMNiBUS"

E0769: **Improving understanding of complex diseases genetics with Bayesian sparse models and variational inference**
Presenter: **Wenmin Zhang**, Montreal Heart Institute, Canada
Genome-wide association studies (GWAS) have discovered many associations between genetic variants and complex diseases. Yet, the interpretation of GWAS results, including identifying causal variants, understanding the interplay between traits, and characterizing disease heterogeneity, is complicated by linkage disequilibrium, as univariate regression models cannot account for correlation between variants. Additionally, the large number of genetic variants poses computational challenges and incurs a high burden of multiple testing. Two novel Bayesian sparse models and efficient variational inference algorithms are presented to address these challenges and facilitate the interpretation of GWAS results. The first method, SparsePro, integrates GWAS associations and functional annotations for prioritizing causal variants, demonstrating improved performance in simulations and identifying biologically relevant causal variants. The second method, SharePro, assesses whether two or more traits share the same genetic signals identified in GWAS. SharePro achieved improved power with a well-controlled false positive rate and identified biologically plausible colocalizations missed by other methods. SharePro could be further adapted for gene-environment interaction analysis by accounting for genetic effect heterogeneity and could effectively reduce multiple testing burdens. These new methods serve as valuable tools for improving the understanding of complex disease genetics.

EO078 Room 210 STATISTICAL INFERENCE ON COMPLEX DATA Chair: Zhao Ren

E0811: **Early indicators of degradation of materials with applications to batteries**
Presenter: **Satish Iyengar**, University of Pittsburgh, United States
The degradation of materials is a common phenomenon that can lead to failure that can require considerable expense to repair. Rechargeable battery-powered devices have become a very common, hence the interest in better understanding degradation. These studies are complicated by nonlinear features of degradation patterns. The use of diffusion processes is studied to approximate recent proposals involving compound Poisson process inputs to model transient phenomena.

E0555: **COVID-19 surveillance via adaptive Fisher's method using weakly geometric grid for combining p-values**
Presenter: **Yusi Fang**, University of Pittsburgh, United States
In COVID-19 surveillance, detecting significant case increases within regions over specific periods is crucial. Classical methods, typically relying on strict parametric assumptions, struggle with the rare events characteristic of COVID-19's early spread. An alternative strategy is employing nonparametric approaches based on p-value combination methods. However, initial COVID-19 outbreaks across regions exhibit varying signal sparsity levels, while existing p-value combination methods demonstrate power in detecting either ultra-sparse or moderately sparse signals in practice, but not both. A modified Fisher's method is presented, utilizing a weakly geometric system-based search strategy to adapt across the entire spectrum of signal sparsity. The method is theoretically and numerically powerful across the whole spectrum of sparsity. Under mild conditions, the method's robustness is examined by combining approximated p-values, demonstrating its powerful performance even when the number of p-values far surpasses the sample sizes for their derivation, offering a novel nonparametric strategy for COVID-19 surveillance. An efficient algorithm is developed to calculate the p-value of our method. Focusing on the early COVID-19 surveillance in the United States, the method consistently detects outbreaks across regions with varying signal sparsity, uncovering diverse patterns of COVID-19's spread, while competing methods struggle with either ultra-sparse or moderately sparse signals.

E0667: **A unifying dependent combination framework with applications to association tests**
Presenter: **Xiufan Yu**, University of Notre Dame, United States
A novel meta-analysis framework is introduced to combine dependent tests under a general setting and utilize it to synthesize various association tests that are calculated from the same dataset. The development builds upon the classical meta-analysis methods of aggregating p-values and also a more recent general method of combining confidence distributions but makes generalizations to handle dependent tests. The proposed framework ensures rigorous statistical guarantees, and a comprehensive study is provided and compared with various existing dependent combination methods. Notably, it is demonstrated that the widely used Cauchy combination method for dependent tests, referred to as the vanilla Cauchy combination in this article, can be viewed as a special case within the framework. Moreover, the proposed framework provides a way to address the problem when the distributional assumptions underlying the vanilla Cauchy combination are violated. The numerical results demonstrate that ignoring the dependence among the to-be-combined components may lead to a severe size distortion phenomenon. Compared to the existing p-value combination methods, including the vanilla Cauchy combination method, the proposed combination framework can handle the dependence accurately and utilizes the information efficiently to construct tests with accurate size and enhanced power.

E0167: **Innovative unsupervised approach for simultaneous subgroup recovery and group-specific feature identification**
Presenter: **Wen Zhou**, New York University, United States
Co-authors: Xiwei Tang, Lyuou Zhang, Lulu Wang
Simultaneously identifying heterogeneous subgroups and the informative features defining them, especially in the absence of responses and with a plethora of features, has long been a challenge in various domains, including omics studies, clinical research, etc. Existing methods have either focused narrowly on global informative features or performed feature selection and group recovery as separate tasks, overlooking their interactions. Such methods might miss scientifically relevant information and lead to suboptimal feature identification and subgroup recovery solutions. To overcome these limitations, a novel unsupervised learning approach is introduced, PAirwise REciprocal fuSE (PARSE), which concurrently pinpoints cluster-specific informative features and conducts high-dimensional clustering. The method employs a new regularization that heavily penalizes features with minor differences across clusters, thus avoiding selecting less informative features that define clusters. The oracle property of PARSE is obtained, and lower bounds for clustering and cluster-specific feature identification are established, affirming the method's optimality in both aspects. For implementations, a computationally efficient enhanced expectation-maximization algorithm is devised. Extensive numerical studies and analysis on identifying gene signatures in human pancreatic cell subtypes using scRNAseq data showcase PARSE's superiority over existing methods.

EO049 Room 307 TOPICS IN FUNCTIONAL AND OBJECT DATA ANALYSIS Chair: Kuang-Yao Lee

E1082: **Nonlinear global Frechet regression for random objects via weak conditional expectation**
Presenter: **Bing Li**, The Pennsylvania State University, United States
The notion of a weak conditional Frechet mean is introduced based on Carleman operators, and then a global nonlinear Frechet regression model is proposed by reproducing kernel Hilbert space (RKHS) embedding. Furthermore, the relationships between the conditional Frechet mean and the weak conditional Frechet mean are established for both Euclidean and object-valued data. The state-of-the-art global Frechet regression is shown to emerge as a special case of the method by choosing a linear kernel. The metric space is required for the predictor to admit a reproducing kernel, while the intrinsic geometry of the metric space for the response is utilized to study the asymptotic properties of the proposed estimates. Numerical studies, including extensive simulations and a real application, are conducted to investigate the performance of the estimator in a finite sample.

E0676: **Binary regression and classification with covariates in metric spaces**
Presenter: **Zhenhua Lin**, University of California, Davis, United States
Co-authors: Yinan Lin
Inspired by logistic regression, a regression model is introduced for data tuples consisting of a binary response and a set of covariates residing in a metric space without vector structures. Based on the proposed model, a binary classifier is also developed for metric-space-valued data. A maximum likelihood estimator is proposed for the metric-space valued regression coefficient in the model, and upper bounds are provided on the estimation error under various metric entropy conditions that quantify the complexity of the underlying metric space. Matching lower bounds are derived for the important metric spaces commonly seen in statistics, establishing the optimality of the proposed estimator in such spaces. Similarly, an upper bound on the excess risk of the developed classifier is provided for general metric spaces. A finer upper bound and a matching lower bound, and thus optimality of the proposed classifier, are established for Riemannian manifolds. To the best of knowledge, the proposed regression model and the above minimax bounds are the first of their kind for analyzing a binary response with covariates residing in general metric spaces. The numerical performance of the proposed estimator and classifier is also investigated via simulation studies, and their practical merits are illustrated via an application to task-related fMRI data.

E1030: **Geodesic optimal transport regression**
Presenter: **Changbo Zhu**, University of Notre Dame, United States
Co-authors: Hans-Georg Mueller
Classical regression models do not cover non-Euclidean data that reside in a general metric space, while the current literature on non-Euclidean regression, by and large, has focused on scenarios where either predictors or responses are random objects, i.e., non-Euclidean, but not both. Geodesic optimal transport regression models are proposed for the case where both predictors and responses lie in a common geodesic metric space, and predictors may include not only one but also several random objects. This provides an extension of classical multiple regression to the case where both predictors and responses reside in non-Euclidean metric spaces, a scenario that has not been considered before. It is based on the concept of optimal geodesic transports, which is defined as an extension of the notion of optimal transports in distribution spaces to more general geodesic metric spaces, where optimal transports are characterized as transports along geodesics. The proposed regression models cover many

spaces of practical statistical interest, including one-dimensional distributions viewed as elements of the 2-Wasserstein space and multidimensional distributions with the Fisher-Rao metric represented as data on the Hilbert sphere. Also included are data on finite-dimensional Riemannian manifolds, with an emphasis on spheres, covering directional and compositional data, as well as data that consist of symmetric positive definite matrices.

E1079: **Functional structural equation models**
Presenter: **Kuang-Yao Lee**, Temple University, United States
Co-authors: Lexin Li
A functional structural equation model is introduced for estimating directional relations from multivariate functional data. The estimation is decoupled into two major steps: directional order determination and selection through sparse functional regression. A score function is first proposed at the linear operator level. It shows that its minimization can recover the true directional order when the relation between each function and its parental functions is nonlinear. A sparse functional additive regression is then developed, where both the response and the multivariate predictors are functions, and the regression relation is additive and nonlinear. Strategies are also proposed to speed up the computation and scale the method. In theory, the consistencies of order determination are established, sparse functional additive regression, and directed acyclic graph estimation while allowing both the dimension of the Karhunen-Loeve expansion coefficients and the number of random functions to diverge with the sample size. The efficacy of the method is illustrated through simulations and an application to brain-effective connectivity analysis.

EO246 Room 313 NEW STATISTICAL METHODS FOR SPATIAL TRANSCRIPTOMICS Chair: Yunshan Duan

E0357: **Statistical identification of cell type-specific spatially variable genes in spatial transcriptomics**
Presenter: **Lulu Shang**, MD Anderson, United States
An essential task in spatial transcriptomics involves identifying genes with spatial expression patterns, known as spatially variable genes (SVGs). Importantly, a subset of SVGs displays diverse spatial expression patterns within a given cell type, thus representing key transcriptomic signatures underlying cellular heterogeneity. Celina, a statistical method, is presented for systematically detecting this subset of cell type-specific SVGs (ct-SVGs). Celina utilizes a spatially varying coefficient model to accurately capture each gene's spatial expression pattern in relation to the distribution of cell types across tissue locations, ensuring effective type I error control and high statistical power. The performance of Celina is evaluated through comprehensive simulations and applications to five real datasets, where existing methods are also adapted and examined, originating from other analytic settings to detect ct-SVGs. Celina proves powerful compared to these ad hoc method adaptations in single-cell resolution spatial transcriptomics and stands as the only effective solution for spot-resolution spatial transcriptomics. The ct-SVGs detected by Celina also enable novel biologically informed downstream analyses, unveiling functional cellular heterogeneity at an unprecedented scale.

E0442: **Spotiphy enables single-cell spatial whole transcriptomics via generative modeling**
Presenter: **Jiyang Yu**, St. Jude Childrenś Research Hospital, United States
Spatial transcriptomics (ST) has revolutionized the understanding of tissue regionalization by making it possible to visualize gene expression across an intact tissue section, but the approach remains dogged by the challenge of achieving single-cell resolution without sacrificing whole genome coverage. Spotiphy (Spot imager with pseudo-single-cell resolution histology) is presented, a novel computational toolkit that transforms sequencing-based ST data into single-cell-resolved whole-transcriptome images. It achieves this by (i) leveraging generative modeling with single-cell RNA sequencing (scRNA-seq) data and high-resolution histological images to convert spot-level ST data into single-cell whole transcriptomic profiles, (ii) utilizing Gaussian processes to impute the cellular composition and expression profiles of non-capture areas, and then (iii) merging the results of the first two steps into a whole-slide image. In evaluations that used matched scRNA-seq, Visium, Xenium, CosMx, and immunohistochemistry datasets for Alzheimer's Disease and normal mouse brains, Spotiphy delivers the most precise cell-type proportions and accurately depicts the distribution of rare cell populations such as immune cells. By making it possible to visualize the cellular localization and expression profiles of an intact tissue section, Spotiphy will be an important tool for gaining insight into the cellular organization, heterogeneity, and function of complex biological systems.

E0556: **Integrating transcriptomic and pathomic features to reconstruct 3D tissue maps with super-resolution**
Presenter: **Mingyao Li**, University of Pennsylvania, United States
Solid tissues form complex 3D structures, and examining the tissue microenvironment in a 3D context allows researchers to gain a comprehensive understanding of how cells interact within the original tissue context. This 3D information also reveals spatial relationships between different cell types and signaling pathways that are not observable in 2D tissue sections. The purpose is to present the recently developed tool that is aimed at generating single-cell resolution 3D ST tissue maps while significantly reducing experimental costs. By integrating information from spatial transcriptomics and pathology imaging data, our method gradually increases gene expression resolution down to the single-cell level. Additionally, an algorithm is developed to register tissue sections obtained from serial tissue cuts and impute missing gene expression data between tissue gaps, enabling the construction of accurate 3D tissue volumes. The resulting analysis will not only generate a single-cell resolution spatial transcriptomics tissue map but also facilitate detailed characterization and quantification of tissue structures of interest in 3D.

E0584: **Immune profiling among colorectal cancer subtypes using dependent mixture models**
Presenter: **Yunshan Duan**, University of Texas at Austin, United States
Co-authors: Shuai Guo, Wenyi Wang, Peter Mueller
The comparison of transcriptomic data across different conditions is of interest to many biomedical studies. Comparative immune cell profiling is considered for early-onset (EO) versus late-onset (LO) colorectal cancer (CRC). EOCRC, diagnosed between ages 18-45, is a rising public health concern that needs to be urgently addressed. However, its etiology remains poorly understood. The purpose is to work towards filling this gap by identifying homogeneous T cell sub-populations that show significantly distinct characteristics across the two tumor types and identifying others that are shared between EOCRC and LOCRC. Dependent finite mixture models are developed where immune subtypes enriched under a specific condition are characterized by terms in the mixture model with common atoms but distinct weights across conditions, whereas common subtypes are characterized by sharing both atoms and relative weights. The proposed model facilitates the desired comparison across conditions by introducing highly structured multi-layer Dirichlet priors. Inferences from simulation studies and data examples are illustrated. Results identify EO- and LO-enriched T cell subtypes whose biomarkers are found to be linked to mechanisms of tumor progression. The findings reveal distinct characteristics of the immune profiles in EOCRC and LOCRC and potentially motivate insights into the treatment of CRC.

EO110 Room 405 STATISTICS ADVANCES IN CHANGE POINTS, BAYESIAN MODELING AND PREDICTION Chair: Xin Wang

E0285: **Borrow dose-response information from an historical experiment in aquatic toxicity assessment**
Presenter: **Jing Zhang**, Miami University, United States
Co-authors: Shixuan Wang, Bin Zhang

Aquatic toxicity tests assess the negative impact of toxicants on the survival, reproduction, and growth of organisms. Since such tests are often repeated periodically in the same lab, borrowing historical experimental outcomes that are congruent would increase the precision in potency estimation. Historical borrowing has been a popular research topic, especially in clinical trials. An existing Bayesian historical borrowing technique, calibrated power prior (CPP), is extended to borrow dose-response information from a historical experiment with a modified calibration algorithm in aquatic toxicity assessment. The effectiveness and flexibility of the proposed method are demonstrated via simulation studies. An application to the Ceriodaphnia dubia reproduction test is presented.

E0286: **Simultaneously detecting spatiotemporal changes with penalized Poisson regression models**
Presenter: **Xin Wang**, San Diego State University, United States

In the realm of large-scale spatiotemporal data, abrupt changes are commonly occurring across both spatial and temporal domains. The aim is to address the concurrent challenges of detecting change points and identifying spatial clusters within spatiotemporal count data. An innovative method based on the Poisson regression model is introduced, employing doubly fused penalization to unveil the underlying spatiotemporal change patterns. To efficiently estimate the model, an iterative shrinkage and threshold-based algorithm to minimize the doubly penalized likelihood function is presented. The statistical consistency properties of the proposed estimator confirm its reliability and accuracy. Furthermore, extensive numerical experiments are conducted to validate the theoretical findings, thereby highlighting the superior performance of the method when compared to existing competitive approaches.

E0351: **Identification and estimation of change points in factor models for high-dimensional time series**
Presenter: **Xialu Liu**, San Diego State University, United States

A new method is proposed to estimate and identify a factor model for high-dimensional time series that contains structural breaks in the factor loading space at unknown time points. The case when there is one change point in factor loadings is first studied, and a consistent estimator for the structural break location is proposed. The proposed estimators are shown for change-point location, and loading spaces are consistent when the number of factors is correctly estimated or overestimated. An algorithm for multiple change-point detection has also been developed. A distinguishing feature of the proposed method is that it is specifically designed for the changes in the factor loading space, and the stationarity assumption is not imposed on either the factor or noise process, while most existing methods for change-point detection of high-dimensional time series with/without a factor structure require the data to be stationary or close to a stationary process between two change points, which is rather restrictive. Numerical experiments, including a Monte Carlo simulation and a real data application, are presented to illustrate how the proposed estimators perform well.

E0557: **Two-part predictive modeling for COVID-19 deaths in the U.S.**
Presenter: **Xiyue Liao**, SDSU, United States

COVID-19 prediction has been essential in the prevention and control of the disease. The motivation of this case study is to develop predictive models for COVID-19 deaths based on a cross-sectional data set with a total of 28,955 observations and 18 variables, which is compiled from 5 data sources from Kaggle. A two-part modeling framework, in which the first part is a binary logistic classifier and the second part includes machine learning or statistical smoothing methods, is introduced to model the highly skewed distribution of COVID-19 deaths. The aim is to understand what factors are most relevant to COVID-19's occurrence and fatality. Evaluation criteria such as root mean squared error (RMSE) and mean absolute error (MAE) are used. It is found that the two-part XGBoost model performs best in predicting the entire distribution of COVID-19 deaths. The most important factors relevant to COVID-19 deaths include population, the rate of primary care physicians, etc.

EO139 Room 406 ADVANCED TOPICS IN STATISTICS AND DATA SCIENCE Chair: Shufei Ge

E0246: **Nonlinear prediction of functional time series**
Presenter: **Haixu Wang**, University of Calgary, Canada

A nonlinear prediction (NOP) method is proposed for functional time series. Conventional methods for functional time series are mainly based on functional principal component analysis or functional regression models. These approaches rely on the stationary or linear assumption of the functional time series. However, real data sets are often non-stationary, and the temporal dependence between trajectories cannot be captured by linear models. Conventional methods also make it hard to analyze multivariate functional time series. To tackle these challenges, the NOP method employs a nonlinear mapping for functional data that can be directly applied to multivariate functions without any preprocessing step. The NOP method constructs feature space with forecast information; hence, it provides a better ground for predicting future trajectories. The NOP method avoids calculating covariance functions and enables online estimation and prediction. The finite sample performance of the NOP method is examined using simulation studies that consider linear, nonlinear, and non-stationary functional time series. The NOP method shows superior prediction performances in comparison with the conventional methods. Three real applications demonstrate the advantages of the NOP method model in predicting air quality, electricity price and mortality rate.

E0423: **Federated event predictions with divergence-guided global aggregation**
Presenter: **Xuhui Fan**, Macquarie University, Australia

Clients-specific events, such as hospital visits, stock market events, and app-based riding-hailing, are typically modelled and predicted in a centralized manner, which may aggregate all the data for training purposes. Such events involve client privacy and are often sparse and uncertain; their existing prediction methods may raise concerns about privacy leakage and require more effort in modelling event sparsity and uncertainty. To enable client-specific privacy-preserving event prediction, the first attempt is made by proposing federated event prediction models (FedEvent), where deep sigmoidal Gaussian Cox processes are developed to generate flexible intensity functions to characterize client-specific event dynamics and capture client event uncertainties simultaneously. Further, a novel framework is proposed using divergence to guide global aggregation over all clients' modeling information, which shares and converges client information and event uncertainty. Divergence measures, including the KL divergence and the Wasserstein distance, are presented to elaborate the approach for uncertain client event prediction. Extensive experimental results verify the advantages of our approach.

E0499: **A parsimonious joint model of survival outcomes and time-varying biomarkers**
Presenter: **Zhiyang Zhou**, University of Wisconsin-Milwaukee, United States
Co-authors: Lihui Zhao

Dynamic risk prediction dynamically updates an individual's risk assessment for a particular outcome by integrating new information over time. The core challenge of this approach involves estimating the intricate interplay between time-varying risk factors and survival outcomes. The shared-random-effects joint model, a key strategy for dynamic risk prediction, simultaneously fits submodels for longitudinal/survival outcomes. However,

as the number of time-varying biomarkers increases, so does the size of unknown parameters, making the model computationally demanding. Additionally, this inflation may potentially compromise the predictive accuracy due to approximation errors in handling the complex likelihood function. To mitigate these issues, a parsimonious joint model is introduced to enhance computational efficiency. The method demonstrates competitive predictive accuracy, verified by numerical studies.

E0567: **A flexible distribution-guided tool in topology data analysis**
Presenter: **Shufei Ge**, ShanghaiTech University, China
A distribution-guided tool is developed that utilizes the property of the probability model and data intrinsic characteristics to achieve topology data analysis. In addition, a metric is introduced to evaluate the performance of similar tools in both non-topological and topological aspects. The numerical experiments showed that the method outperformed the traditional approach in various scenarios.

EO024 Room 408 NEW DEVELOPMENTS IN MICROBIOME RESEARCH Chair: Gen Li

E0387: **QuanT: Identifying unmeasured heterogeneity in microbiome data via quantile thresholding**
Presenter: **Ni Zhao**, Johns Hopkins University, United States
Unmeasured technical and biomedical heterogeneity in microbiome data can arise from differential processing and design. Uncorrected for, they can lead to spurious results. The quantile thresholding (QuanT) approach is proposed, a comprehensive non-parametric hidden variable inference method that accommodates the complex distributions of microbial read counts. QuanT is applied to synthetic and real data sets and demonstrates its ability to identify unmeasured heterogeneity and improve downstream analysis.

E0398: **A general testing method for inference of microbial networks with compositional data**
Presenter: **Yijuan Hu**, Emory University, United States
Inference of microbial networks reveals inter-dependencies or interactions among microbial taxa within communities. The compositional, sparse, high-dimensional, highly overdispersed, and sometimes clustered sequencing data pose significant challenges to this task. There is a lack of testing methods that control the false discovery rate (FDR) and thus calibrate the discoveries. A novel testing method called TestNet has been introduced. It is based on the empirical covariance and distance covariance of the centred-log-ratio data for capturing linear and nonlinear dependencies, respectively. A permutation procedure is developed for generating null replicates that account for the compositional effects and the extensive zero counts in microbiome data, assuming sparse dependencies in a microbial community. The permutation procedure readily allows schemes that preserve clustering structures in the samples, e.g., longitudinal samples. Therefore, the method applies to general scenarios involving the inference of microbial networks. The extensive simulation studies indicate that TestNet controls the FDR well while achieving high efficiency in a wide range of scenarios; the results from existing methods are not calibrated for any error rate.

E0905: **Locally sparse varying coefficient mixed model with application to longitudinal microbiome differential abundance**
Presenter: **Simon Fontaine**, University of Michigan, Canada
Co-authors: Gen Li, Ji Zhu
Differential abundance (DA) analysis in microbiome studies has recently been used to uncover a plethora of associations between microbial composition and various health conditions. While current approaches to DA typically apply only to cross-sectional data, many studies feature a longitudinal design to understand the underlying microbial dynamics better. A novel varying coefficient mixed-effects model with local sparsity is introduced to study DA on longitudinal microbial studies. The proposed method can identify time intervals of significant group differences while accounting for temporal dependence. Specifically, a penalized kernel smoothing approach is exploited for parameter estimation, and local regression is extended to include a random effect without any requirements for the sampling design. In particular, it operates effectively regardless of whether sampling times are shared across subjects, accommodating irregular sampling or potentially missing observations. Simulation studies demonstrate the necessity of modelling dependence for precise estimation and support recovery. The method's application to a longitudinal study of mice's oral microbiome during cancer development revealed significant scientific insights that were otherwise not discernible through cross-sectional analyses.

E0989: **Microbial interactions and community stability from longitudinal microbiome study**
Presenter: **Huilin Li**, New York University, United States
Dynamic changes in microbiome communities may play important roles in human health and diseases. The recent rise in longitudinal microbiome studies calls for statistical methods to model temporal dynamic patterns and quantify microbial interactions and community stability simultaneously. The aim is to propose a novel autoregressive zero-inflated mixed-effects model (ARZIMM) to capture the sparse microbial interactions and estimate the community stability. ARZIMM employs a zero-inflated Poisson autoregressive model to model the excessive zero abundances and the non-zero abundances separately, a random effect to investigate the underlying dynamic pattern shared within the group, and a Lasso-type penalty to capture and estimate the sparse microbial interactions. Based on the estimated microbial interaction matrix, the estimate of community stability is further derived, and the core dynamic patterns are identified through network inference. ARZIMM is evaluated in comparison with the other methods through extensive simulation studies and real data analyses.

EO142 Room 411 (Virtual sessions) STATISTICAL METHODS FOR ANALYZING HIGH-THROUGHPUT DATA Chair: Elif Acar

E0622: **Mediation analysis to infer direct genetic effects on disease risks**
Presenter: **Yildiz Yilmaz**, Memorial University of Newfoundland, Canada
Co-authors: Brady Ryan
Many genetic associations have been identified with disease risks. However, the associations do not infer the causal genetic effects. To distinguish direct genetic effects from indirect genetic effects, a directed acyclic graph is considered to have a direct genetic effect on the primary disease occurrence phenotype and indirect effects through intermediate phenotypes, which are potentially confounded by measured and unmeasured factors. A mediation analysis method is discussed using the odds ratio scale to infer controlled direct genetic effects on disease risks while removing indirect effects through mediators and adjusting the model for measured and unmeasured confounders. The proposed method uses the estimating function methodology with robust sandwich standard errors. The method allows the inclusion of genetic mediator interaction. It provides consistent controlled direct genetic effect estimates and valid tests for testing the absence of the direct effect in both cohort and case-control studies. The proposed method is applied to genome sequence data to estimate and test controlled direct genetic effects on hypertension.

E0731: **Leveraging genetic correlation for multi-trait polygenic scores construction via L1 penalized regression**
Presenter: **Osvaldo Espin-Garcia**, University of Western Ontario, Canada
Polygenic risk scores (PRS) quantify the genetic contribution of an individual's genotype to a trait, e.g. disease or phenotype. PRS can be used to group subjects into different risk strata and can thus be treated as predictors in clinical and epidemiological studies. However, PRS methods typically focus on a single trait at a time, ignoring the potential simultaneous influence of genes on multiple traits. To address this limitation, a recently published model that uses an L1 penalty is extended by incorporating a genetic correlation matrix among traits into the cost function of the penalized regression framework. The main objective is to improve the predictive ability of multi-trait PRS models when multiple traits are

of equal interest. The proposed method is evaluated against alternatives via comprehensive numerical studies with a focus on marginal and joint performance metrics. The penalized approach is applied to two studies: one analyzing smoking and drinking behaviors in a cohort of head and neck participants and another one examining cardiovascular traits from a birth cohort from Finland. Lastly, a memory-efficient re-implementation of the penalized regression framework will be discussed.

E0997: **Automated statistical methods for high-throughput phenotyping experiments**

Presenter: **Elif Acar**, University of Manitoba, Canada

Many health applications produce ever-increasing quantities of biological data. As such applications often rely on automated pipelines for data analysis, an important statistical challenge is to evaluate and refine these pipelines as more and more data are acquired. This challenge is exemplified by the high-throughput phenotyping experiments conducted by the International Mouse Phenotyping Consortium (IMPC), where multiple phenotype measurements are obtained for a small set of gene-edited mice and a large set of controls acquired continually over time. Model selection is a fundamental component of the automated pipeline, increasing the power of detecting the gene effect. However, the effect of post-selection inference in this setting is not well understood. Moreover, due to the size and complexity of the data, gene function is assessed by combining the results of univariate phenotype analyses. However, analyzing multiple phenotypes simultaneously at the individual level greatly improves the power of detection. The focus is on evaluating and improving the IMPC statistical pipeline along these lines of inquiry.

E1001: **Bayesian dimension reduction in microbiome platforms**

Presenter: **Kevin McGregor**, University of Manitoba, Canada

Dimension reduction techniques are among the most essential analytical tools in analyzing high-dimensional data. Generalized principal component analysis is an extension to standard principal component analysis (PCA) for various types of non-Gaussian data and has been widely used to identify low-dimensional features in high-dimensional data. For microbiome count data, the multinomial PCA is a natural counterpart to standard PCA. However, this technique fails to account for the excessive number of zero values frequently observed in microbiome count data. To allow for sparsity, zero-inflated multivariate distributions can be used. A Bayesian zero-inflated probabilistic PCA model is proposed for extracting information in compositional count data. A classification variational approximation algorithm is developed to fit the model. A simulation study and an application in a pediatric-onset multiple sclerosis metagenomic dataset will be further featured.

Thursday 18.07.2024 10:25 - 12:30 Parallel Session G – EcoSta2024

EI007 Room 106 NEW CHALLENGES IN HIGH-DIMENSIONAL DATA ANALYSIS Chair: Binyan Jiang

E0208: **Modeling emotional expressions for multiple cancers via a linguistic analysis of an online health community**
Presenter: **Steven Ma**, Yale University, United States

The diagnosis and treatment of cancer can evoke a variety of adverse emotions. Online health communities (OHCs) provide a safe platform for cancer patients and those closely related to express emotions without fear of judgment or stigma. In the literature, linguistic analysis of OHCs is usually limited to a single disease and based on methods with various technical limitations. Posts from September 2010 to September 2022 are analyzed on nine publicly available cancers at the American Cancer Society's Cancer Survivors Network (CSN). A novel network analysis technique is proposed based on a latent space model. The proposed approach decomposes the emotional expression semantic networks into an across-cancer time-independent component (which describes the baseline that is shared by multiple cancers), a cancer-specific time-independent component (which describes cancer-specific properties), and an across-cancer time-dependent component (which accommodates temporal effects on multiple cancer communities). A novel clustering structure and a change point structure are considered for the second and third components, respectively. A penalization approach is proposed, and its theoretical and computational properties are carefully examined. The analysis of the CSN data leads to sensible networks and deeper insights into emotions for cancer overall and specific cancer types.

E0934: **Enveloped Huber regression**
Presenter: **Le Zhou**, Hong Kong Baptist University, Hong Kong
Co-authors: Dennis Cook, Hui Zou

Huber regression (HR) is a popular flexible alternative to the least squares regression when the error follows a heavy-tailed distribution. A new method called the enveloped Huber regression (EHR) is proposed by considering the envelope assumption that some subspace of the predictors exist that have no association with the response, which is referred to as the immaterial part. More efficient estimation is achieved via the removal of the immaterial part. Different from the envelope least squares (ENV) model, whose estimation is based on maximum normal likelihood, the estimation of the EHR model is through the generalized method of moments. The asymptotic normality of the EHR estimator is established, and it is shown that EHR is more efficient than HR. Moreover, EHR is more efficient than ENV when the error distribution is heavy-tailed while maintaining a small efficiency loss when the error distribution is normal. Moreover, the theory also covers the heteroscedastic case in which the error may depend on the covariates. The envelope dimension in EHR is a tuning parameter that is determined by the data in practice. A novel generalized information criterion (GIC) is further proposed for dimension selection and its consistency is established. Extensive numerical studies confirm the messages of the theory.

E1010: **Residual importance weighted transfer learning For high-dimensional linear regression**
Presenter: **Junlong Zhao**, Beijing Normal University, China

Transfer learning is an emerging paradigm for leveraging multiple sources to improve the statistical inference on a single target. A novel approach named residual importance weighted transfer learning (RIW-TL) is proposed for high-dimensional linear models built on penalized likelihood. Compared to existing methods, such as trans-Lasso, which selects sources in an all-in-all-out manner, RIW-TL includes samples via importance weighting and thus may permit more effective sample use. To determine the weights, remarkably, RIW-TL only requires the knowledge of one-dimensional densities dependent on residuals, thus overcoming the curse of dimensionality of having to estimate high-dimensional densities in naive importance weighting. It is shown that the oracle RIW-TL provides a faster rate than its competitors and develops a cross-fitting procedure to estimate this oracle. Variants of RIW-TL by adopting different choices for residual weighting are discussed. The theoretical properties of RIW-TL and variants are established and compared with those of Lasso and trans-Lasso. Extensive simulation and real data analysis confirm its advantages.

EO068 Room 102 STATISTICS AND DATA SCIENCE FOR DIGITAL FINANCE AND TOKENOMICS Chair: Stephen Chan

E0698: **Network transitions in the cryptocurrency market: The impact of regional conflicts**
Presenter: **Jeffrey Chu**, Renmin University of China, China

Over the past 15 years, cryptocurrencies have been exposed to a wide variety of significant global events, such as financial crises, rising inflation, booms and recessions, and, most recently, the coronavirus (COVID-19) pandemic. Before 2022, cryptocurrencies had never witnessed a military conflict and simultaneously played a significant role. This changed in February 2022 with the Russia-Ukraine conflict and, most recently, in October 2023 with the Israel-Hamas conflict. The aim is to study how the cryptocurrency markets have evolved throughout these conflict periods through a network graph approach. Some of the key questions investigated are: How have military conflicts impacted cryptocurrency markets? Do different classes of cryptocurrency assets react differently to conflicts? What factors are driving the cryptocurrency networks during these times?

E0732: **Detecting illicit activity in digital cryptocurrency networks**
Presenter: **Mingkun Yuan**, Renmin University of China, China

Anti-money laundering (AML) regulations exist to protect financial systems, but they often hinder financial inclusion and result in higher participation costs for those who are the poorest and on the edges of society. Digital cryptocurrencies offer one possible solution for financial inclusion, but like traditional financial systems, they are still susceptible to illicit activity. The literature on anomaly and fraud detection in cryptocurrency networks is reviewed, and attempts to understand the temporal evolution of cryptocurrency networks are made to determine the most significant and relevant factors for determining fraud in cryptocurrency networks are made.

E0845: **Empirical analysis of the metaverse non-fungible tokens**
Presenter: **Stephen Chan**, American University of Sharjah, United Arab Emirates
Co-authors: Jeffrey Chu, Yuanyuan Zhang

A thorough investigation of non-fungible tokens (NFTs) in the metaverse is presented, initiated by exploring the metaverse's development, the rise of NFTs, and their transformative impacts. Furthermore, it conducts a detailed empirical analysis of five specific metaverse NFTs. Adopting a comprehensive analytical approach, this research utilizes descriptive statistics, Hill's estimator, detrended fluctuation analysis, volatility and asymmetric volatility clustering, and quantile-on-quantile regression. This robust methodology reveals distinctive market behaviors, pricing dynamics, and investor trends in the metaverse NFT domain. The results shed light on the complex intricacies of the NFT market, offering essential insights for investors, creators, and regulators. It underscores the importance of innovative strategies and thoughtful regulatory frameworks to effectively navigate the metaverse's evolving landscape. To ensure the reliability of the conclusions, a comparative analysis is conducted using metaverse indices from Bloomberg (BBMI) and data from Yield Guild Games (YGG), providing a solid foundation for our findings.

E0847: **Empirical analysis of illicit transactions in blockchain networks**
Presenter: **Yuanyuan Zhang**, University of Manchester, United Kingdom
Co-authors: Stephen Chan, Jeffrey Chu
In the last ten years, over 19 billion dollars have been stolen through breaches and fraud, with Ethereum being one of the most targeted cryptocurrencies, accounting for 33.3 per cent of hacking and fraud cases. The aim is to provide the first empirical analysis that explores the network of Ethereum transactions that are associated with real entities belonging to illicit categories. To examine the Ethereum network topology, a range of network measurements are implemented for characteristics, local network properties, and global network properties. The network analysis will reveal the nature of Ethereum users that use it for illicit transactions.

E1015: **Effective multidimensional persistence for Ethereum network representation learning**
Presenter: **Yuzhou Chen**, Temple University, United States
Topological data analysis (TDA) is gaining prominence across a wide spectrum of machine learning tasks spanning manifold learning to graph classification. A pivotal technique within TDA is persistent homology (PH), which furnishes an exclusive topological imprint of data by tracing the evolution of latent structures as a scale parameter changes. Present PH tools are confined to analyzing data through a single filter parameter. However, many scenarios necessitate the consideration of multiple relevant parameters to attain finer insights into the data. The issue is addressed by introducing the effective multidimensional persistence (EMP) framework, empowering data exploration by simultaneously varying multiple scale parameters. The framework integrates descriptor functions into the analysis process, yielding a highly expressive data summary. It seamlessly integrates established single PH summaries into multidimensional counterparts like EMP landscapes, silhouettes, images, and surfaces. In addition, EMP's utility is demonstrated in Ethereum network prediction tasks, showing its effectiveness. Results reveal EMP enhances various single PH descriptors, outperforming state-of-the-art baselines.

EO039 Room 103 SPATIAL PANEL DATA MODELS Chair: Zhenlin Yang

E0329: **Threshold spatial panel data models with fixed effects**
Presenter: **Xiaoyu Meng**, Nankai University, China
General estimation and inference methods are introduced for threshold spatial panel data models with two-way fixed effects in a diminishing-threshold-effects framework. A valid objective function is obtained through a simple adjustment on the concentrated quasi-loglikelihood with fixed effects being concentrated out, which leads to a consistent estimation of all common parameters. It is shown that the estimation of the threshold parameter has a negligible effect on the asymptotic distribution of the main parameter estimators, and thereby, regular inference methods apply, though a bias correction may be necessary. The limiting distribution of the threshold parameter estimator is shown to be non-regular and infeasible, and for inference, a likelihood ratio test procedure is proposed. Test for the non-existence of threshold effects faces an identification issue at the null, and a sup-Wald test is proposed with critical values being bootstrapped. Monte Carlo results show that the proposed methods perform well in finite samples. An empirical application is presented on age-of-leader effects on political competitions across Chinese cities.

E0338: **Dynamic spatial panel data models with interactive fixed effects**
Presenter: **Liyao Li**, East China Normal University, China
An M-estimation method is proposed for estimating dynamic spatial panel data models with interactive fixed effects based on (relatively) short panels. Unbiased estimating functions (EF) are obtained by adjusting the concentrated conditional quasi scores, given initial values and with factor loadings being concentrated out, to account for the effects of conditioning and concentration. Solving the estimating equations gives M-estimators of common parameters and factor parameters. Under fixed T, $\sqrt{n}$-consistency and joint asymptotic normality of both sets of M-estimators are established; under $T = o(n)$, the M-estimators of common parameters are shown to be $\sqrt{nT}$-consistent and asymptotically normal. For inference, EF is decomposed into a sum of n nearly uncorrelated terms. Outer products of these n terms, together with a covariance adjustment, lead to a consistent estimator of the VC matrix under both fixed T and $T = o(n)$. Important extensions of the methods, allowing for unknown heteroskedasticity, time-varying spatial weight matrices, and high-order dynamic and spatial effects, are critically discussed. Monte Carlo results show that the proposed methods perform well in finite samples and outperform the existing methods when T is not large.

E0339: **Efficient and sequential estimation of high-order dynamic spatial panels with time-varying strongly dominant units**
Presenter: **Chen Yahui**, Xiamen University, China
Co-authors: Han Xiaoyi, Zhang Jiajun, Jin Fei
The estimation of a high-order spatial dynamic panel data model is considered with time-varying strongly dominant units and heteroskedasticity. The dominant units vary over time, and the numbers of dominant units are finite or infinite. To accommodate the model specification, a central limit theorem (CLT) is developed where the column sum magnitude in the quadratic form can be equal to one and the existence of heteroskedasticity is allowed. The generalized method of moments estimator (GMME) and root estimator (RE) is proposed, as well as the consistency and asymptotic normality of these estimators when both n and T are large. The advantage of RE is that it has a closed-form solution and is asymptotically as efficient as the best GMME. Monte Carlo simulations demonstrate that the estimators have satisfactory finite sample performances. Finally, an empirical application is presented to illustrate the usefulness of the model on the peer effects of firm finance decisions across Chinese listed firms.

E0342: **Learning from neighbors: Peer effects in Chinese household financial investments**
Presenter: **Juncong Guo**, Shanghai Jiao Tong University, China
Co-authors: Xi Qu
Nationally representative survey data from China is employed to examine peer effects on the investment behaviors of Chinese households with respect to wealth management products. The empirical findings indicate that neighbors' behaviors have a statistically significant impact on investments in these financial products. These peer effects exist at both the extensive and intensive margins, even under incomplete information. Heterogeneity analyses suggest that the underlying mechanism driving these effects is the spread and learning of information. Additionally, it is observed that the rise in the participation rate in wealth management product investments is linked to a reduction in inequality. Accordingly, the findings propose that policymakers could leverage peer effects from influencers to promote household investments, thereby contributing to the mitigation of inequality.

E0783: **Firm to firm supply network, urban agglomeration and firms' structure change evidence from structural model**
Presenter: **ZhiQiang Zhang**, NanKai University of China, China
Inter-firm network connections and urban agglomeration economies exert a significant influence on firms' total factor productivity. A structural model is constructed in which inter-firm network linkages, agglomeration economies, and their interactive effects endogenously impact firm productivity. Utilizing panel data on publicly listed companies in China spanning the period 2008-2020, it empirically investigates the relationships between inter-firm network ties, urban agglomeration externalities, and firm-level total factor productivity in the Chinese context. The findings reveal that the intensity of inter-firm connections significantly shapes the non-linear relationship between agglomeration economies and total factor productivity. Regarding the underlying mechanisms, price markups and inter-firm synergies in innovation appear to play a significant role, and inter-firm network linkages coupled with urban agglomeration economies further facilitate firms' structural transformation. Grounded in these findings,

it puts forward policy recommendations aimed at enhancing firm competitiveness by expanding firms' supply chain networks and bolstering the agglomeration economies of urban clusters.

EO209 Room 104 FRONTIERS OF BAYESIAN METHODS FOR COMPLEX DATA Chair: Fan Bu

E0484: **Regression analysis for single-cell RNA-seq data**
Presenter: **Fangda Song**, The Chinese University of Hong Kong, Shenzhen, China
Co-authors: Kevin Y Yip, Yingying Wei
scRNA-seq studies that assay a large cohort of donors are emerging, which provides opportunities for us to understand how gene expression profiles of a cell type are affected by donors' characteristics, such as age, gender and disease status. However, statistical methods developed to study the association between bulk gene expression data and a set of covariates are not applicable to scRNA-seq data because the cell-type label of each cell is unknown. In addition, batch effects, variations in cell-type abundance between donors, and missing data due to dropout events all add challenges to detecting the association between scRNA-seq data and covariates. Regress-seq, a Bayesian hierarchical model, is developed that can simultaneously cluster cell types, correct batch effects and the effects of the covariates inferred on cell-types-specific gene expression profiles. Moreover, the conditions are derived from the experimental designs under which the integrative analysis of multiple scRNA-seq studies is valid so that the cell type effects, batch effects and covariate effects can be separated. Regress-seq is envisioned to greatly facilitate the development of personalized medicine.

E0517: **Phylogenetic latent position model for populations of networks**
Presenter: **Federico Pavone**, Universita Paris Dauphine-PSL, France
In many applications, networks are characterized by a hierarchical or multiresolution organization of the nodes responsible for the connectivity. A phylogenetic latent position model is proposed that effectively learns the multiresolution structure via modelling the latent positions as realizations of a branching process on a phylogenetic tree. The model is applied to the problem of learning the underlying structure responsible for the connectivity patterns in the human brain. A population of networks is analyzed to represent the brain's structural connectivity for a set of subjects. The model reveals a tree organization of the brain regions coherent with known hemisphere and lobe partitions. Such a result uncovers interesting new possible clustering of the brain regions at different levels of resolution.

E0573: **A Bayesian non-parametric approach: Integrating VAEs and GANs using Wasserstein and MMD**
Presenter: **Forough Fazeliasl**, University of Hong Kong, Hong Kong
Co-authors: Michael Minyi Zhang
Generative models like GANs and VAEs have shown promise in generating realistic images. While GANs produce sharp images, they often miss out on the full diversity of the target distribution. On the other hand, VAEs generate diverse images but tend to be blurry. To overcome these limitations, a novel approach is proposed that combines GANs and VAEs using a Bayesian non-parametric (BNP) framework. The method incorporates Wasserstein and maximum mean discrepancy (MMD) measures in the loss function to effectively learn the latent space and generate diverse, high-quality samples. By merging the discriminative power of GANs and the reconstruction capabilities of VAEs, our model outperforms existing approaches in tasks like anomaly detection and data augmentation. Additionally, an extra generator is introduced in the code space to explore areas overlooked by the VAE. The BNP perspective allows for the modeling of data distribution using an infinite-dimensional space, providing flexibility and reducing overfitting risks. This framework enhances the performance of GANs and VAEs, resulting in a more robust generative model suitable for various applications. By combining the strengths of these models and mitigating their weaknesses, the approach opens new possibilities for generating high-quality images that closely resemble real ones.

E0649: **Inferring HIV transmission patterns from viral deep-sequence data via latent spatial Poisson processes**
Presenter: **Fan Bu**, UCLA, United States
Viral deep-sequencing technologies play a crucial role in understanding disease transmission patterns because the higher resolution of these data provides evidence on transmission direction. To better utilize these data and account for uncertainty in phylogenetic analysis, a spatial Poisson process model is proposed to uncover HIV transmission flow patterns at the population level. Pairings of two individuals are represented with viral sequence data as typed points, with coordinates representing covariates such as sex and age and the point type representing the unobserved transmission statuses (linkage and direction). Points are associated with deep-sequence phylogenetic analysis summary scores that reflect the strength of evidence for each transmission status. The method jointly infers the latent transmission status for all pairings and the transmission flow surface on the source-recipient covariate space. In contrast to existing methods, the framework does not require pre-classification of the transmission statuses of data points; instead, it learns them probabilistically through full Bayesian inference. By directly modeling continuous spatial processes with smooth densities, the method enjoys significant computational advantages over previous methods that discretize the covariate space. In an HIV transmission study from Rakai, Uganda, the framework is demonstrated to capture age structures in HIV transmission at high resolution and bring valuable insights.

EO189 Room 105 RELIABILITY AND PRECISION IN MODERN BIOMEDICAL RESEARCH Chair: Andrew Chen

E0857: **Reliability in functional brain measurement**
Presenter: **Brian Caffo**, Johns Hopkins University, United States
Repeatability in functional brain measurement is discussed. The use of fingerprinting in functional connectomics is discussed, which involves matching an individual's brain scans from two different sessions in a group of subjects. The process involves statistical tests based on the assumption of exchangeability. The soundness of these tests, their power, and the factors that influence them are discussed. Theoretical investigations and numerical studies are presented using fMRI datasets from the human connectome project (HCP) and the Baltimore longitudinal study of ageing (BLSA). The sensitivity of the tests to various factors is further examined, such as familial status or demographics, and a detailed analysis of single regional connections in the HCP data is performed.

E0248: **Challenges in measuring individual differences and reliability of brain function**
Presenter: **Ting Xu**, Child Mind Institute, United States
With a growing interest in personalized medicine, functional neuroimaging research has recently shifted focus from the evaluation of group-level summaries to associating individual differences in brain function with behaviors. However, this new focus brings forth challenges related to accurately measuring the sources of individual variation in functional signals. The impact of within-individual variations is highlighted, and the concept of measurement reliability is discussed as a critical tool for accounting for within- and between-individual variations when measuring individual differences in brain function. A tool, reliability explorer (ReX), is also presented, which facilitates the examination of individual variations and reliability and the effective direction for optimizing individual differences in biomarker discovery.

E0797: **Distance-based reliability**
Presenter: **Philip Reiss**, University of Haifa, Israel
Co-authors: Meng Xu, Ivor Cribben
The intraclass correlation coefficient (ICC) is a classical index of measurement reliability. With the advent of new and complex types of data for which the ICC is not defined, there is a need for new ways to assess reliability. To meet this need, a distance-based ICC (dbICC) defined in terms of arbitrary distances among observations is proposed. It is shown that naive bootstrap confidence intervals for the dbICC suffer from undercoverage, and a bias correction is introduced to remedy this. The Spearman-Brown (SB) formula, which shows how more intensive measurement increases reliability, is extended to encompass the dbICC. The generalized SB formula depends on a notion of measurement intensity that generalizes simple averaging over multiple measurements. The dbICC is illustrated by analyzing test-retest reliability in several settings, including brain connectivity matrices derived from functional magnetic resonance imaging, as well as complex phenotypes derived from experience sampling and psychotherapy research.

E0181: **Transfer learning methods to get more reliable estimates of effects and predictions**
Presenter: **Haotian Zheng**, Vertex Pharmaceuticals, United States
Co-authors: Sai Li, Hongzhe Li
One common problem in modern genomics and multiomics studies is that the model often has weak classification or prediction performance due to a small set of training samples compared to the number of genomic features. On the other hand, there are often auxiliary data sets that are related to the target learning problem, which can potentially be transferred to improve parameter estimation or prediction. Estimation and prediction methods are introduced for high-dimensional transfer learning for two problems. The first problem is transfer learning for high dimensional linear discriminant analysis (LDA), one of the most commonly used methods for building a classification rule when the data are approximately Gaussian. The second problem is transfer learning for high-dimensional linear regression in settings where summary statistics is only observed in the auxiliary studies. It is shown that such summary statistics, together with external data for estimating the feature covariance matrix (e.g., linkage disequilibrium (LD) matrix), can be effectively used in transfer learning. It is shown that under some assumptions, transfer learning methods have lower error rates in estimating the effect sizes and in classification/prediction. The methods are demonstrated using numerical studies and applications to several data sets, including the polygenic risk score (PRS) prediction of blood-related phenotypes using Penn Medicine Biobank genotype data and UK Biobank summary statistics.

E0583: **Subject-level segmentation accuracy weights for volumetric studies involving label fusion**
Presenter: **Christina Chen**, University of Pennsylvania, United States
Co-authors: Sandhitsu Das, Matthew Tisdall, Fengling Hu, Andrew Chen, Paul Yushkevich, David Wolk, Russell Shinohara
In neuroimaging research, volumetric data contribute valuable information for understanding brain changes during healthy ageing and pathological processes. Extracting these measures from images requires segmenting the regions of interest (ROIs), and many popular methods accomplish this by fusing labels from multiple expert-segmented images called atlases. However, post-segmentation, current practices typically treat each subject's measurement equally without incorporating any information about variation in their segmentation precision. This naive approach hinders comparing ROI volumes between different samples to identify associations between tissue volume and disease or phenotype. A novel method is proposed that estimates the variance of the measured ROI volume for each subject due to the multi-atlas segmentation procedure. It is demonstrated in real data that weighting by these estimates markedly improves the power to detect a mean difference in hippocampal volume between controls and subjects with mild cognitive impairment or Alzheimer's disease.

EO124 Room 108 STATISTICAL LEARNING AND INFERENCE FOR LARGE-SCALE COMPLEX DATA Chair: Xiufan Yu

E0379: **Data-driven robust change detection using Wasserstein ambiguity sets**
Presenter: **Liyan Xie**, The Chinese University of Hong Kong - Shenzhen, China
Co-authors: Yiran Yang
The problem of quickest detection of a change in the distribution of a sequence of independent observations is considered. It is assumed that the pre-change distribution is known (accurately estimated), while the only information about the post-change distribution is through a (small) set of labeled data. This post-change data is used in a data-driven minimax robust framework, where an uncertainty set for the post-change distribution is constructed using the Wasserstein distance from the empirical distribution of the data. The robust change detection problem is studied in an asymptotic setting where the meantime to the false alarm goes to infinity. A cumulative sum (CuSum) test based on the least favorable distribution, which is referred to as the distributionally robust (DR) CuSum test, is then shown to be asymptotically robust. The results are further extended to the case where the uncertainty set is constructed adaptively. The proposed method is applied to a real-world human activity detection scenario, and validation results are presented.

E0753: **High-dimensional statistical inference for linkage disequilibrium score regression and its cross-ancestry extensions**
Presenter: **Fei Xue**, Purdue University, United States
Co-authors: Bingxin Zhao
Linkage disequilibrium score regression (LDSC) has emerged as an essential tool for genetic and genomic analyses of complex traits, utilizing high-dimensional data derived from genome-wide association studies (GWAS). LDSC is investigated within a fixed-effect data integration framework, underscoring its ability to merge multi-source GWAS data and reference panels. In particular, genome-wide dependence is considered among the high-dimensional GWAS summary statistics, along with the block-diagonal dependence pattern in estimated LD scores. The analysis uncovers several key factors of both the original GWAS and reference panel datasets that determine the performance of LDSC. It is shown that it is relatively feasible for LDSC-based estimators to achieve asymptotic normality when applied to genome-wide genetic variants, whereas it becomes considerably challenging when the focus is on a much smaller subset of genetic variants (e.g., in partitioned heritability analysis). Moreover, by modelling the disparities in LD patterns across different populations, it is unveiled that LDSC can be expanded to conduct cross-ancestry analyses using data from distinct global populations.

E0809: **ByMI Byzantine machine identification with false discovery rate control**
Presenter: **Haojie Ren**, Shanghai Jiao Tong University, China
Various robust estimation methods or algorithms have been proposed to hedge against Byzantine failures in distributed learning. However, there is a lack of systematic approaches to provide theoretical guarantees of significance in detecting those Byzantine machines. A general detection procedure, ByMI, is developed via error rate control to address this issue, which applies to many robust learning problems. The key idea is to apply the sample-splitting strategy on each worker machine to construct a score statistic integrated with a general robust estimation and then to utilize the symmetry property of those scores to derive a data-driven threshold. The proposed method is dimension insensitive and p-value-free with the help of the symmetry property and can achieve false discovery rate control under mild conditions. Numerical experiments on both synthetic and real data validate the theoretical results and demonstrate the effectiveness of the proposed method on Byzantine machine identification.

E0843: **Network regression and supervised centrality estimation**
Presenter: **Dan Yang**, University of Hong Kong, Hong Kong
Co-authors: Junhui Jeffrey Cai, Haipeng Shen, Wu Zhu, Linda Zhao, Ran Chen
The centrality in a network is often used to measure nodes' importance and model network effects on a certain outcome. Empirical studies widely adopt a two-stage procedure, which first estimates the centrality from the observed noisy network and then infers the network effect from the estimated centrality, even though it lacks theoretical understanding. A unified modeling framework is proposed, under which the shortcomings of the two-stage procedure are first proven, including the inconsistency of the centrality estimation and the invalidity of the network effect inference. Furthermore, a supervised centrality estimation methodology is proposed, which aims to simultaneously estimate both centrality and network effect. The advantages in both regards are proved theoretically and demonstrated numerically via extensive simulations and a case study in predicting currency risk premiums from the global trade network.

EO147 Room 109 HIGH-DIMENSIONAL STATISTICS AND RANDOM MATRIX THEORY Chair: Zhaoyuan Li

E0363: **Tracy-Widom, Gaussian, and Bootstrap: Approximations for leading eigenvalues in high-dimensional PCA**
Presenter: **Nina Doernemann**, Aarhus University, Denmark
Co-authors: Miles Lopes
The leading eigenvalues of sample covariance matrices play a fundamental role in many aspects of high-dimensional statistics. Under certain conditions, when the data dimension and sample size diverge proportionally, these eigenvalues undergo a well-known phase transition: In the sub-critical regime, the eigenvalues have Tracy-Widom fluctuations of order $n^{-2/3}$, while in the supercritical regime, they have Gaussian fluctuations of order $n^{-1/2}$. However, the statistical problem of determining which regime underlies a given dataset has remained largely unresolved. A new testing framework and procedure to address this problem is developed. In particular, the procedure is demonstrated to be at an asymptotically controlled level, and it is power-consistent for certain spiked alternatives. Also, this testing procedure enables the design of a new bootstrap method for approximating the distributions of functionals of the leading eigenvalues within the sub-critical regime, which is the first such method supported by theoretical guarantees.

E0504: **On eigenvalue distributions of large auto-covariance matrices**
Presenter: **Wangjun Yuan**, Department of Mathematics, University of Luxembourg, Luxembourg
Co-authors: Jianfeng Yao
A limiting distribution for eigenvalues of a class of auto-covariance matrices was established. The same distribution has been found in the literature for a regularized version of these auto-covariance matrices. The original non-regularized auto-covariance matrices are non-invertible, thus introducing supplementary difficulties for studying their eigenvalues through Girko's Hermitization scheme. The key result is a new polynomial lower bound for a specific family of least singular values associated with a rank-defective quadratic function of a random matrix with independent and identically distributed entries. Another innovation is that the lag of the auto-covariance matrices can grow to infinity with the matrix dimension.

E0805: **Concentration of the measure as an hypothesis for theoretical machine learning**
Presenter: **Cosme Jean Leon Louart**, Chinese university of Hong Kong Shenzhen, China
Michel Talagrand won this year's Abel Prize after devoting a large part of his research life to concentrating on measure theory, a very conceptual theory that is actually very powerful when applied to machine learning problems. Following this year's great news, a general picture of this theory is given, and some tools that will be applied to a toy example are provided.

E0541: **High-dimensional consistent independence testing with maxima of rank correlations**
Presenter: **Hongjian Shi**, Technical University of Munich, Germany
Co-authors: Mathias Drton, Fang Han
Testing mutual independence for high-dimensional observations is a fundamental statistical challenge. Popular tests based on linear and simple rank correlations are known to be incapable of detecting nonlinear, non-monotone relationships, calling for methods that can account for such dependencies. A family of tests is proposed to address this challenge, which is constructed using maxima of pairwise rank correlations that permit consistent assessment of pairwise independence. Built upon a newly developed Cramer-type moderate deviation theorem for degenerate U-statistics, our results cover a variety of rank correlations, including Hoeffdings D, Blum-Kiefer-Rosenblatts R and Bergsma-Dassios-Yanagimotos τ^*. The proposed tests are distribution-free in the class of multivariate distributions with continuous margins, implementable without the need for permutation, and are shown to be rate-optimal against sparse alternatives under the Gaussian copula model. As a by-product of the study, an identity between the aforementioned three rank correlation statistics is revealed, hence making a step towards proving a conjecture of Bergsma and Dassios.

E0838: **Statistical inference for principal components of spiked covariance matrices**
Presenter: **Jingming Wang**, Harvard University, United States
Co-authors: Zhigang Bao, Xiucai Ding, Ke Wang
In random matrix theory, one of the central topics is the limiting behavior of eigenvalues and eigenvectors of random matrices under fixed-rank perturbations. A famous model raised by Johnstone is the so-called spiked covariance matrix model. It is a sample covariance matrix whose population has all its eigenvalues equal to one except for a few top eigenvalues (spikes). From the principal component analysis (PCA) point of view, the main task is to study the limiting behavior of the top eigenvalues and eigenvectors of the spiked sample covariance matrix. The high dimensional setting is considered; namely, both the sample size n and the dimension p are large. The limiting distribution of the eigenvectors associated with the largest eigenvalues is first identified for the sample covariance matrix in the supercritical regime. Second, the joint distribution is derived from the extreme eigenvalues and the associated eigenvectors. Third, based on these results, accurate and powerful statistics are proposed, and their asymptotic distributions are derived in order to conduct hypothesis testing on the principal components. Numerical simulations also confirm the accuracy and power of the proposed statistics and illustrate significantly better performance compared to the existing methods in the literature.

EO186 Room 110 BIOSTATISTICS, BIOINFORMATICS, AND CAUSAL INFERENCE (VIRTUAL) Chair: Li-Pang Chen

E0697: **Length-biased and partly interval-censored survival data analysis with measurement error in covariates**
Presenter: **Li-Pang Chen**, National Chengchi University, Taiwan
Length-biased and partly interval-censored data are considered, whose challenges primarily come from biased sampling and interfere with interval censoring. Unlike existing methods that focus on low-dimensional data and assume the covariates to be precisely measured, researchers are able to encounter high-dimensional data subject to measurement error, which are ubiquitous in applications and make estimation unreliable. To address those challenges, a valid inference method is explored for handling high-dimensional length-biased and interval-censored survival data with measurement error in covariates under the accelerated failure time model. The SIMEX method is primarily employed to correct the measurement error effects, and the boosting procedure for variable selection and estimation is proposed. The proposed method is able to handle the case that the dimension of covariates is larger than the sample size and enjoys appealing features that the distributions of the covariates are left unspecified.

E0749: **Bayesian model for disease-specific gene detection in high-dimensional spatially resolved transcriptomics**
Presenter: **Qihuang Zhang**, McGill University, Canada
Identifying disease-indicative genes is critical for deciphering disease mechanisms and continues to attract significant interest. Spatial transcriptomics offers unprecedented insights for the detection of disease-specific genes by enabling within-tissue contrasts. However, this new technology poses challenges for conventional statistical models developed for RNA-seq, as these models often neglect the spatial organization of tissue spots. A new Bayesian shrinkage model is discussed to characterize the relationship between high-dimensional gene expressions and the disease status of tissue spots, incorporating spatial correlation among these spots through autoregressive terms. The model adopts a hierarchical structure to accommodate for the missing data within tissues and is further extended to facilitate the analysis of multiple correlated samples. To ensure the model's applicability to datasets of varying sizes, two computational frameworks are carried out for Bayesian parameter estimation, tailored to both small and large sample scenarios. Simulation studies are conducted to evaluate the performance of the proposed model, and the model is also applied to analyze the data arising from a HER2-positive breast cancer study.

E0750: **Robust and flexible high-dimensional causal mediation model for DNA methylation studies**
Presenter: **An-Shun Tai**, National Cheng Kung University, Taiwan
In the pathogenesis of diseases, DNA methylation (DNAm) markers play a pivotal role in influencing gene expression and engaging in diverse biological processes. Given the extensive number of DNAm markers, exceeding half a million, implementing a high-dimensional mediation model is necessary to identify the activated DNAm markers within the mediation pathway and assess their mediation effects. Most existing high-dimensional mediation models necessitate stringent assumptions, including correctly prespecifying the mediation relationship and determining all outcomes, mediators, and exposure models. However, fulfilling these assumptions is challenging in the context of high-dimensional mediators. A novel Bayesian estimation procedure is studied for interventional mediation effects, offering robustness against model misspecification and flexibility in prespecifying the mediation structure. Spike-and-slab priors are employed to integrate Bayesian variable selection into the modeling process. The proposed method is demonstrated using publicly available genome-wide array-based cancer studies to estimate the causal effects mediated through DNAm.

E0771: **Ultrahigh-dimensional discriminant analysis and its application to gene expression data**
Presenter: **Jou Chin Wu**, National Chengchi University, Taiwan
Co-authors: Li-Pang Chen
Discriminant analysis has been a commonly used strategy to handle classification with binary or multi-label responses. Under multivariate normal distributions of covariates, a linear or quadratic discriminant function can be derived, which is used as a boundary to classify subjects. In the discriminant function, the estimation of the precision matrix, which is defined as the inverse of the covariance matrix, is a crucial issue. While a large body of estimation methods is available to estimate the precision matrix, most methods not only fail to handle ultrahigh-dimensionality in the sense that the dimension of variables is extremely larger than the sample size but also require longer computational time. In addition, in the presence of nonlinear dependence among variables, existing methods may falsely miss the detection of dependence. To tackle those challenges, the model-free feature screening method is extended to reduce the dimension of variables and detect possibly nonlinear pairwise dependence structures among variables. After that, the graphical lasso and joint graphical lasso methods are adopted to estimate the precision matrix and then implement the estimator to the discriminant function. Numerical studies are conducted to assess the prediction performance of the proposed method.

E0774: **Weighted least-squares estimation for semiparametric multivariate accelerated failure time model with regularization**
Presenter: **Sy Han Chiou**, Southern Methodist University, United States
Co-authors: Ying Chen, Chuan-Fa Tang, Min Chen
In large-scale epidemiology and medical studies, informative sampling and high-dimensional covariates pose significant challenges in statistical inference. These challenges warrant an accurate inference procedure that addresses the sampling bias arising from informative sampling and incorporates an efficient variable selection process within it. A weighted least-squares estimation is considered in the semiparametric multivariate accelerated failure time model framework for right-censored clustered data emerging from informative sampling. By embedding the generalized estimating equation (GEE) techniques, the proposed estimating procedure accommodates situations when observations within a cluster are correlated. The regularization techniques are further incorporated to perform variable selection by penalizing the proposed GEE procedure. The consistency and asymptotic normality of proposed estimators is established, and it is shown that the proposed variable selection method has an oracle property. Extensive simulation results indicate superior performance of the proposed estimators over the existing method that does not account for sampling bias or within-cluster dependence for multivariate responses. The proposed regularization procedure also achieves favorable variable selection performance under moderate sample sizes. Two dental studies are used to illustrate the practical applications of the proposed methods.

EO220 Room 111 MEASUREMENT ERROR AND SURVIVAL DATA MODELS Chair: Liqun Wang

E1087: **New procedure for controlling false discovery rate in Cox model**
Presenter: **Hengjian Cui**, Capital Normal University,Beijing,China, China
A novel feature selection method for the Cox model in high-dimensional data analysis is developed. The method is constructed under the framework of the FDR control for multiple testing, and the multiple data-splitting strategy is adopted. For each splitting, the data is divided into two disjoint parts. The first part of the data is used for feature selection, and multiple tests are conducted for the set of selected features in the two parts. Then, the z-values of the statistics are aggregated to control the FDR, and the set of important features is chosen by rejecting the null hypotheses. The asymptotic theory of FDR control for the proposed method is established under mild conditions. The finite sample performance of the feature selection procedure is evaluated by Monte Carlo simulations. It is shown that the proposed new procedure effectively controls the empirical FDR. The new approach is illustrated through a real dataset from the diffuse large-B-cell lymphoma study.

E0951: **Variable selection for the generalized odds rate non-mixture cure model with current status data**
Presenter: **Xuewen Lu**, University of Calgary, Canada
Co-authors: Saba Saghatchi , Jingjing Wu
Current status or case I interval-censored data are common in labour economics, clinical trials, and hospital visits. In some cases, there exists a cured sub-population where individuals never experience the event of interest. Meanwhile, when the current status data is modeled, one may also encounter an extraordinarily large number of risk factors; variable selection is desired in the model building. The purpose is to study variable selection methods for the generalized odds rate non-mixture cure model with current status data using penalized semiparametric likelihood function when the dimension of the covariates is diverging. The proposed model encompasses the proportional hazards (PH) and proportional odds (PO) non-mixture cure models as special cases. The broken adaptive bridge (BAR) penalty is utilized for regularization, and the asymptotic properties of the resultant estimators of regression parameters, including the oracle and group properties, are studied. The sieve method based on Bernstein polynomials is employed to estimate the unknown cumulative distribution function. To facilitate the computation, a novel penalized expectation maximization (EM) algorithm is implemented. Furthermore, a simulation study is conducted to assess the finite sample performance of the proposed method and compare it with other existing penalized methods such as Lasso, ALasso, and SCAD. Finally, the method is applied to a real data set for illustration.

E0770: **Extrapolation estimation for nonparametric regression with measurement error**
Presenter: **Weixing Song**, Kansas State University, United States
The purpose is to introduce an extrapolation algorithm designed for estimating regression functions in nonparametric regression models when covariates are affected by normal measurement errors. The approach involves applying conditional expectation directly to the kernel-weighted least squares of the discrepancies between the local linear approximation and the observed responses. This innovative algorithm eliminates the need for the simulation step, a characteristic of classical simulation extrapolation methods, thereby significantly reducing computational time. It is worth noting that the method provides an exact form of the extrapolation function. However, obtaining the extrapolation estimate is not as straightforward as merely substituting a negative one for the extrapolation variable in the fitted extrapolation function when the bandwidth is less than the standard deviation of the measurement error. Furthermore, the large sample properties of the proposed estimation procedure are delved into, and simulation studies are conducted. Additionally, a real data example is presented to showcase its practical applications.

E0658: **A robust minimum distance estimation of Cox PH models**
Presenter: **Jingjing Wu**, University of Calgary, Canada
Cox proportional hazard (PH) models are simple but frequently used in survival analysis. The unknown coefficient parameters in a Cox PH model are usually estimated using the partial maximum likelihood estimation (PMLE) introduced in a past study. Nevertheless, PMLE is generally non-robust against model misspecification and outlying observations. When data is contaminated, PMLE produces inaccurate estimates with a large bias. A robust distance-based method is proposed instead, specifically, a minimum Hellinger distance estimation (MHDE). Both discrete and continuous covariates are considered, and different types of covariates are accommodated by introducing different versions of MHDE. Through an extensive simulation study, the finite-sample performance of the proposed MHDEs is examined and compared with the PMLE. Numerical results show that the proposed MHDEs are competitive with the PMLE under the true model, while they outperform the PMLE when outliers are present, which testifies to the robustness property of the MHDEs. The applications of the proposed MHDE are also demonstrated in real data analysis.

E0672: **Instrumental variable method in regularized regression with mismeasured predictors**
Presenter: **Liqun Wang**, University of Manitoba, Canada
Regularization methods are widely used in high-dimensional regression models, and most methods are developed for situations where all variables are correctly and precisely measured. However, in real data analysis, measurement error is common. The variable selection and estimation problems are studied in linear and generalized linear models where some of the predictors are not directly observable. How measurement error impacts the selection results is demonstrated, and regularized instrumental variable methods are proposed to correct the effects of measurement error. The proposed methods are consistent in selection and estimation, and their asymptotic distributions are derived under general conditions. The performance of the methods is also investigated through Monte Carlo simulations, and they are compared with the naive method, which ignores measurement errors. Finally, the proposed method is applied to a real dataset.

EO095 Room 212 RANDOM FIELDS AND THEIR STATISTICAL APPLICATIONS Chair: Dan Cheng

E0586: **Excursion sets and critical points of Gaussian random fields over high thresholds**
Presenter: **Yi Shen**, University of Waterloo, Canada
Co-authors: Weinan Qi, Paul Marriott
The excursion sets and the location and type of the critical points of isotropic Gaussian random fields are discussed, satisfying certain conditions over high thresholds. After quickly introducing the Poisson limit result for the critical points and the excursion sets as the threshold tends to infinity, a discussion of the local behavior of the critical points is proceeded with, and it is shown that a pair of close critical points in R^n, both above a high threshold, predominantly consists of one local maxima and one saddle point with index n-1. The possibility of approximating these locations when the threshold is high but not extremely high is also possible and is further discussed.

E0427: **Classes of multivariate and space-time power-law covariance functions**
Presenter: **Pulong Ma**, Iowa State University, United States
Understanding marginal covariance and cross-covariance structures is essential for modeling continuously indexed multivariate and space-time processes. The Matern covariance function, which only allows short-range dependence, has enabled several notable developments in multivariate and space-time covariance models in the past few decades. However, many geophysical processes possess long-range dependence in space and space-time domains, for which the Matern-based covariance models often fail to capture. To address this issue, new classes of multivariate and space-time covariance functions are introduced with power-law decay in the tail. Several validity conditions are derived to ensure the positive definiteness of the proposed multivariate covariance models. The interplay among long-range dependence, Markov property, and screening effect is examined theoretically to provide foundations for their practical usefulness. Extensive simulation examples and real datasets are used to illustrate the superior performance of the proposed covariance models over the state-of-the-art models.

E0728: **Confidence regions for geometric features via extreme value distributions of Gaussian random fields on manifolds**
Presenter: **Wanli Qiao**, George Mason University, United States
Geometric features, frequently expressed as level sets of functions in Euclidean spaces, often manifest as manifolds. Employing kernel estimators, establishing confidence regions for these features is commonly framed as discerning extreme value distributions of locally stationary Gaussian random fields along the level sets. The aim is to delve into the asymptotic formulations of these distributions and explore bootstrap techniques for their approximation. Furthermore, the application of these methodologies is demonstrated in constructing confidence regions for ridges.

E0620: **Estimation of expected Euler characteristic curves of nonstationary smooth random fields**
Presenter: **Dan Cheng**, Arizona State University, United States

The expected Euler characteristic curve (EEC) summarizes the topology of the excursion sets of a random field above the excursion threshold in terms of its expected Euler characteristic. For large thresholds, the EEC is an excellent approximation for the tail distribution of the supremum of a smooth Gaussian field and has applications in the control of familywise error rate (FWER) and construction of simultaneous confidence bands. Therefore, it is important and valuable to estimate the EEC. Viewed as a function of the excursion threshold, the EEC of a Gaussian-related field is expressed by the Gaussian kinematic formula as a finite sum of known functions multiplied by the Lipschitz Killing curvatures (LKC) of the generating Gaussian field. This transforms the estimation of EEC into estimating LKC. A new method is presented to estimate the LKC as linear projections of pinned Euler characteristic curves obtained from realizations of Gaussian fields. This provides an efficient and accurate tool to estimate the EEC and, hence, high excursion probabilities of Gaussian fields.

EO120 Room 202 HIGH-DIMENSIONAL GENETIC AND GENOMIC DATA Chair: Linxi Liu

E0454: **Identification of differentially expressed genes via knockoff statistics in single-cell RNA sequencing data analysis**
Presenter: **Linxi Liu**, University of Pittsburgh, United States
Co-authors: Lixia Yi

Single-cell RNA sequencing (scRNA-seq) is a technology that provides high-resolution gene expression data. With scRNA-seq data, an important statistical task is to identify differentially expressed genes (DEGs) in case-control studies, as results from DEG analysis can contribute to a more comprehensive understanding of the disease mechanism and new discovery of potential risk factors. However, given the burden of multiple testing and low transcript capture rate in scRNA-seq experiments, DEG identification may suffer from low power. Co-expressed genes and unobserved confounders can also lead to an inflated Type-I error. A new method for DEG identification is introduced in scRNA-seq data analysis under the knockoff framework to overcome these difficulties. The method starts by imputing missing gene expressions by taking advantage of correlations among genes, and then it generates model-X knockoffs in a computationally efficient way. By incorporating widely used marginal screening tests for scRNA-seq data, we implement a knockoff filter for DEG identification that can control the false discovery rate (FDR) at the nominal level. On a range of synthetic and real data sets, FDR control and power gain of the new approach are illustrated. The method is also applied to the single-cell transcriptomic analysis of Alzheimer's disease. The results demonstrate that the new method can identify genes with weaker effects that are missed by conventional approaches.

E0607: **Fine-mapping gene-based associations via knockoff analysis of biobank-scale data**
Presenter: **Shiyang Ma**, Shanghai Jiao Tong University, China

Gene-based tests are important tools for elucidating the genetic basis of complex traits. Despite substantial recent efforts in this direction, the existing tests are still limited owing to low power and detection of false positive signals due to the confounding effects of linkage disequilibrium and co-regulation. BIGKnock (BIobank-scale Gene-based association test via Knockoffs) is proposed, a computationally efficient gene-based testing approach for biobank-scale data that leverages long-range chromatin interaction data and performs conditional genome-wide testing via knockoffs. BIGKnock can prioritize causal genes over proxy associations at a locus. BIGKnock is applied to the UK Biobank data with 405,296 participants for multiple binary and quantitative traits and shows that relative to conventional gene-based tests, BIGKnock produces smaller sets of significant genes that contain the causal gene(s) with high probability.

E0810: **BLEND: Bayesian cellular deconvolution with reference selection**
Presenter: **Jiebiao Wang**, University of Pittsburgh, United States

Cellular deconvolution aims to estimate cell type fractions from bulk transcriptomic and other omic data. While many estimators have been proposed, most of them fail to account for the heterogeneity in cell type-specific (CTS) expression across bulk samples, ignore discrepancies between CTS expression in bulk data and cell type references, and provide no guidance on cell type reference selection or integration. To address these issues, BLEND is introduced, a hierarchical Bayesian method that deconvolves bulk RNA data by leveraging multiple reference datasets. BLEND uses the data to learn the most suitable references for each sample by exploring the convex hulls of references and employs a bag-of-words representation for bulk count data for deconvolution. A Gibbs sampler is derived for posterior computation and an algorithm that maximizes the posterior distribution to speed up computation. The benchmarking studies on both simulated and real human brain data highlight BLEND's superior performance in a variety of scenarios. Requiring no data transformation, cell type marker gene selection, or reference quality evaluation, BLEND facilitates cellular deconvolution with its superior accuracy and robustness.

E0956: **Gene regulatory networks analysis from single cell multi-omics data**
Presenter: **Zhana Duren**, Clemson University, United States
Co-authors: Qiuyue Yuan

Existing methods for gene regulatory networks (GRNs) inference rely on gene expression data alone, or on lower resolution bulk data. Despite recent integration of ATAC-seq and RNA-seq data, learning complex mechanisms from limited independent data points still presents a daunting challenge. Here we present LINGER (LIfelong neural Network for GEne Regulation), a machine learning method to infer GRNs from single-cell paired gene expression and chromatin accessibility data. LINGER incorporates both atlas-scale external bulk data across diverse cellular contexts and prior knowledge of transcription factor (TF) motifs as a manifold regularization. LINGER achieves 4-7-fold relative increase in accuracy over existing methods and reveals a complex regulatory landscape of genome-wide association studies, enabling enhanced interpretation of disease-associated variants and genes. Following the GRN inference from a reference single-cell multiome data, LINGER allows for the estimation of TF activity solely from bulk or single-cell gene expression data, leveraging the abundance of available gene expression data to identify driver regulators from case-control studies.

E0953: **Beyond guilty by association at scale: Searching for causal variants on the basis of genome-wide summary statistics**
Presenter: **Zihuai He**, Stanford University, United States

Understanding the causal genetic architecture of complex phenotypes is essential for future research into disease mechanisms and potential therapies. The aim is to present a novel framework for genome-wide detection of sets of variants that carry non-redundant information on the phenotypes and are therefore more likely to be causal in a biological sense. Crucially, the framework requires only summary statistics obtained from standard genome-wide marginal association testing. The described approach, implemented in open-source software, is also computationally efficient, requiring less than 15 minutes on a single CPU to perform genome-wide analysis. Through extensive genome-wide simulation studies, it is shown that the method can substantially outperform usual two-stage marginal association testing and fine-mapping procedures in precision and recall. In applications to a meta-analysis of ten large-scale genetic studies of Alzheimer's disease (AD), 82 loci associated with AD are identified, including 37 additional loci missed by conventional GWAS pipeline. The identified putative causal variants achieve state-of-the-art agreement with massively parallel reporter assays and CRISPR-Cas9 experiments.

EO255 Room 204 PROGRESS IN ANALYZING CENSORED EVENT TIMES: A CONTEMPORARY PERSPECTIVE Chair: Wenyu Gao

E0656: **Semiparametric analysis of multivariate panel count data with informative observation processes**
Presenter: **Xin He**, University of Maryland, United States
Co-authors: Chang Chen
Multivariate panel count data arise in studies involving several related types of recurrent events in which the study subjects are examined periodically over time. The observation times may vary from subject to subject and carry information about the underlying recurrent event processes of interest. A joint modeling approach is proposed to account for the informative observation processes using bivariate shared frailty models. Estimating equations and an EM algorithm are developed for the parameter estimation, and the resulting estimators are shown to be consistent and asymptotically normal. The proposed methods are evaluated through simulation studies and illustrated with an application to data from a clinical trial of skin cancer.

E0629: **Semiparametric Bayesian kernel survival model using multilevel learning**
Presenter: **Inyoung Kim**, Virginia Tech, United States
Motivated by a breast cancer gene-pathway data set, which exhibits the "small n, large p" characteristics, a semiparametric variable selection method is proposed for the Bayesian kernel survival model to simultaneously study the effects of both clinical covariates and gene expression levels within a pathway on survival time and also identify important variables associated to survival time. The unknown high-dimension functions of pathways are modeled via the Gaussian kernel machine to consider the possibility that genes within the same pathway interact with each other. To address the multiple comparisons problem under a full Bayesian setting, a similarity-dependent procedure is proposed based on the Bayes factor to control the family-wise error rate. The outperformance of the approach is demonstrated under various simulation settings and pathways data.

E0198: **Multivariate degradation modeling with inverse Gaussian processes**
Presenter: **Guanqi Fang**, Zhejiang Gongshang University, China
Many engineering products have more than one failure mode, and the evolution of each mode can be monitored by measuring a performance characteristic (PC). It is found that multi-dimensional degradation processes have often been observed in engineering practice. A novel multivariate degradation model built upon inverse Gaussian processes is introduced. The model is able to account for 1) the stochastic nature of each individual PC, 2) the heterogeneity among different units, and, more importantly, 3) any possible dependence among these PCs. Along with the model, some mathematically tractable properties are discussed, including the joint and conditional distribution functions that could facilitate future degradation prediction and lifetime estimation. In addition, a statistical inference method and model validation tools are provided. Finally, the proposed methodology is demonstrated using simulation studies and illustrative examples.

E0199: **Variational inference for spatial correlated failure time data under Bayesian framework**
Presenter: **Yueyao Wang**, Zhejiang Gongshang University, China
In modern reliability analysis, geographically referenced time-to-event data are often collected. For such reliable data, the spatial dependence on the failure time needs to be properly accounted for in the model. In the literature, spatial random effect models, such as the cumulative exposure model, are often used for analysis with the Bayesian approach as model inference. However, the inference problem is often high dimensional with respect to the number of spatial locations. Consequently, the conventional Markov Chain Monte Carlo (MCMC) methods for sampling the posterior can be very time-consuming when the number of spatial locations is large. Thus, the capability of variational inference (VI) for the inference of spatial survival models is investigated, and a good balance between estimation accuracy and computational efficiency is aimed. Specifically, two divergence metrics, alpha-divergence and the Kullback-Leibler (KL) divergence, are used in the VI methods for the spatial cumulative exposure model. The numerical study compares the MCMC and VI methods under two spatial GPU lifetime data. The comparison results show that the VI method has comparable performance to the MCMC approach but with much more efficient computational time.

E0204: **Regression analysis of semiparametric Cox-Aalen transformation models with partly interval-censored data**
Presenter: **Yinghao Pan**, University of North Carolina at Charlotte, United States
Co-authors: Xi Ning, Yanqing Sun, Peter Gilbert
Partly interval-censored data, comprising exact and interval-censored observations, are prevalent in biomedical, clinical, and epidemiological studies. A flexible class of so-called Cox-Aalen transformation models is introduced for regression analysis of such data. These models offer a versatile framework by accommodating multiplicative and additive covariate effects within a transformation while allowing for potentially time-dependent covariates. Moreover, this class of models includes many popular models, such as the Cox-Aalen and transformation models, as special cases. To facilitate efficient computation, a set of estimating equations is formulated, and an expectation-solving (ES) algorithm that guarantees stability and rapid convergence is proposed. Under mild regularity assumptions, the resulting estimator is shown to be consistent and asymptotically normal, with its variance consistently estimated by weighted bootstrap. Finally, the proposed method is evaluated through comprehensive simulations and applied to analyze data from a randomized HIV/AIDS trial.

EO017 Room 209 NEW STATISTICAL METHODS FOR COMPLEX DATA Chair: Tianxi Li

E0194: **Data-driven label-poisoning backdoor attack**
Presenter: **Xuan Bi**, University of Minnesota, United States
Backdoor attacks, which aim to disrupt or paralyze classifiers on specific tasks, are becoming an emerging concern in several learning scenarios, e.g., machine learning as a service. Various backdoor attacks have been introduced in the literature, including perturbation-based methods, which modify a subset of training data, and clean-sample methods, which relabel only a proportion of training samples. Indeed, clean-sample attacks can be particularly stealthy since they never require modifying the samples at the training and test stages. However, the state-of-the-art clean-sample attack of relabeling training data based on their semantic meanings could be ineffective and inefficient in test performances due to heuristic selections of semantic patterns. A new type of clean-sample backdoor attack is introduced, named a DLP backdoor attack, allowing attackers to backdoor effectively, as measured by test performances, for an arbitrary backdoor sample size. The critical component of DLP is a data-driven backdoor scoring mechanism embedded in a multi-task formulation, which enables attackers to perform well on the normal learning tasks and the backdoor tasks simultaneously. Systematic empirical evaluations show the superior performance of the proposed DLP to state-of-the-art clean-sample attacks.

E0677: **Causal clustering: Design of cluster experiments under network interference**
Presenter: **Lihua Lei**, Stanford University, United States
The design of cluster experiments is studied to estimate the global treatment effect in the presence of network spillovers. A framework is provided to choose the clustering that minimizes the worst-case mean-squared error of the estimated global effect. It is shown that optimal clustering solves a novel penalized min-cut optimization problem computed via off-the-shelf semi-definite programming algorithms. The analysis also characterizes simple conditions to choose between any two cluster designs, including choosing between a cluster or individual-level randomization. The method's properties are illustrated using unique network data from the universe of Facebook's users and existing data from a field experiment.

E1035: **Proximal MCMC for Bayesian inference of constrained and regularized estimation**
Presenter: **Eric Chi**, Rice University, United States

Proximal Markov Chain Monte Carlo (MCMC) is a flexible and general Bayesian inference framework for constrained or regularized parametric estimation. The basic idea of proximal MCMC is to approximate non-smooth regularization terms via the Moreau-Yosida envelope. Initial proximal MCMC strategies, however, fixed nuisance and regularization parameters as constants and relied on the Langevin algorithm for the posterior sampling. Proximal MCMC is extended to a fully Bayesian framework with modeling and data-adaptive estimation of all parameters, including regularization parameters. More efficient sampling algorithms, such as the Hamiltonian Monte Carlo, are employed to scale proximal MCMC to high-dimensional problems. The proposed proximal MCMC offers a versatile and modularized procedure for the inference of constrained and non-smooth problems that are mostly tuning parameter-free. Its utility is illustrated in various statistical estimation and machine-learning tasks.

E0529: **Optimal rates of convergence for sliced inverse regression with differential privacy**
Presenter: **Wenbiao Zhao**, Beijing Institute of Technology, China

Sliced inverse regression (SIR) is a highly efficient paradigm used for dimension reduction by replacing high-dimensional covariates with a limited number of linear combinations. The focus is on the implementation of the classical SIR approach integrated with a Gaussian differential privacy mechanism to estimate the central space while preserving privacy. The tradeoff between statistical accuracy and privacy in sufficient dimension reduction problems is illustrated under both the classical low-dimensional and modern high-dimensional settings. Additionally, the minimax rate of the proposed estimator is achieved with Gaussian differential privacy constraint, and this rate is illustrated to be optimal for multiple index models with a bounded dimension of the central space. Extensive numerical studies on synthetic data sets are conducted to assess the effectiveness of the proposed technique in finite sample scenarios, and real data analysis is presented to showcase its practical application.

E0539: **Moment deviation subspaces of dimension reduction for high-dimensional data with change structure**
Presenter: **Luoyao Yu**, Xián Jiaotong Unversity, China

The notion of moment deviation subspaces of dimension reduction is introduced for high-dimensional data with a change structure. A novel estimation method is proposed to identify subspaces by combining the Mahalanobis matrix and the pooled covariance matrix. The theoretical properties are investigated to show that the change point detection and clustering can be equivalently implemented in the dimension reduction subspaces, whether the data structure is dense or sparse, whenever the dimension divided by the sample size goes to zero. An iterative algorithm is proposed based on dimension reduction subspaces that can be applied for data clustering of high-dimensional data. The numerical studies on synthetic and real data sets suggest that the dimension-reduction versions of existing methods of change point detection and clustering methods significantly improve the performances of existing approaches in finite sample scenarios.

EO192 Room 210 STATISTICAL INFERENCE FOR COMPLEX DATA Chair: Robert Lunde

E0360: **Semi-supervised U-statistics**
Presenter: **Ilmun Kim**, Yonsei University, Korea, South
Co-authors: Larry Wasserman, Sivaraman Balakrishnan, Matey Neykov

Semi-supervised datasets are ubiquitous across diverse domains where obtaining fully labeled data is costly or time-consuming. The prevalence of such datasets has consistently driven the demand for new tools and methods that exploit the potential of unlabeled data. In response to this demand, semi-supervised U-statistics enhanced by the abundance of unlabeled data are introduced, and their statistical properties are investigated. The proposed approach is shown to be asymptotically Normal and exhibits notable efficiency gains over classical U-statistics by effectively integrating various powerful prediction tools into the framework. To understand the fundamental difficulty of the problem, minimax lower bounds are derived in semi-supervised settings, and the procedure is showcased to be semi-parametrically efficient under regularity conditions. Moreover, tailored to bivariate kernels, a refined approach is proposed that outperforms the classical U-statistic across all degeneracy regimes and demonstrates its optimality properties.

E0668: **On varimax asymptotics in network models and spectral methods for dimensionality reduction**
Presenter: **Joshua Cape**, University of Wisconsin, Madison, United States

Varimax factor rotations, while popular among practitioners in psychology and statistics since being introduced in a past study, have historically been viewed with scepticism and suspicion by some theoreticians and mathematical statisticians. Currently, a recent study provides new, fundamental insight: varimax rotations provably perform statistical estimation in certain classes of latent variable models when paired with spectral-based matrix truncations for dimensionality reduction. This newfound understanding of varimax rotations is built by developing further connections to network analysis and spectral methods rooted in entrywise matrix perturbation analysis. Concretely, the asymptotic multivariate normality of vectors is established in varimax-transformed Euclidean point clouds that represent low-dimensional node embeddings in certain latent space random graph models. Related concepts, including network sparsity, data denoising, and the role of matrix rank, are addressed in latent variable parameterizations. Collectively, these findings, at the confluence of classical and contemporary multivariate analysis, reinforce methodology and inference procedures grounded in matrix factorization-based nonparametric techniques. Numerical examples illustrate the findings and supplement the discussion.

E0747: **Temporal spatial model via trend filtering**
Presenter: **Carlos Misael Madrid Padilla**, University of Notre Dame, United States
Co-authors: Oscar Hernan Madrid, Daren Wang

The focus is on the estimation of a nonparametric regression function in the presence of data with temporal-spatial dependencies. In such a context, trend filtering, a nonparametric estimator, is studied. To the best of knowledge, this estimator has not previously been examined in a similar context. For univariate settings, the signals considered are assumed to have a kth weak derivative with bounded total variation, allowing for a general degree of smoothness. In the multivariate setting, we study a variant of the K-nearest neighbor fused lasso estimator. For this case, the function is required to have bounded variation and satisfy a property that extends a piecewise Lipschitz continuity criterion, or the function is assumed to be piecewise Lipschitz. An ADMM algorithm is developed for practical computation. By aligning with lower bounds, the minimax optimality of the univariate and multivariate estimators is shown. A unique phase transition phenomenon, previously unprecedented in trend filtering studies, emerges through the analysis. Both simulation studies and real data applications underscore the superior performance of the method when compared with established techniques in the existing literature.

E1108: **Dynamic subgroup analysis on heterogeneous regression model**
Presenter: **Haowen Zhou**, University of Virginia, United States

In recent years, the heterogeneous effect model, rather than a conventional homogeneous effect model, has become prevalent in various areas such as precision medicine and market segmentation. Yet it remains challenging to deal with such heterogeneity changing over time. To fill this gap, we propose a dynamic subgrouping framework on a heterogeneous regression model, which can capture the temporal pattern on heterogeneous covariates-effects. We impose the novel multidirectional separation penalty on the individualized covariates-effects to pursue subgroups of individuals dynamically while leveraging the temporal pattern of subpopulations by modeling the subgroup centers with smoothing splines. In contrast to

all existing approaches, we allow the individuals to change their underlying subgroup memberships over time. We lay out the theoretical framework for the proposed model and estimates. An efficient ADMM algorithm with computational scalability is developed for model estimation. The outperformance of the proposed model has been validated by simulation studies and empirical data analysis in the stock market.

E1109: **Matrix completion with model-free weighting**
Presenter: **Jiayi Wang**, The University of Texas at Dallas, United States
Co-authors: Raymond Ka Wai Wong, Xiaojun Mao, Kwun Chuen Gary Chan
A novel method is proposed for matrix completion under general non-uniform missing structures. By controlling an upper bound of a novel balancing error, we construct weights that can actively adjust for the non-uniformity in the empirical risk without explicitly modeling the observation probabilities, and can be computed efficiently via convex optimization. The recovered matrix based on the proposed weighted empirical risk enjoys appealing theoretical guarantees. In particular, the proposed method achieves a stronger guarantee than existing work in terms of scaling with respect to the observation probabilities under asymptotically heterogeneous missing settings (where entry-wise observation probabilities can be of different orders). These settings can be regarded as a better theoretical model of missing patterns with highly varying probabilities. We also provide a new minimax lower bound under a class of heterogeneous settings. Numerical experiments are also provided to demonstrate the effectiveness of the proposed method.

EO106 Room 307 RECENT ADVANCES IN STATISTICAL MODELING FOR NEUROIMAGING DATA Chair: Shuo Chen

E0228: **Causal mediation analysis for multilevel and functional data**
Presenter: **Xi Luo**, Univ of Texas Health Science Center at Houston, United States
Co-authors: Yi Zhao, Brian Caffo, Martin Lindquist, Michael Sobel
Causal mediation analysis typically involves conditions that may not be applicable in neuroimaging studies. A multilevel causal mediation framework is introduced to overcome this limitation and more accurately quantify information flow in brain pathways. The framework is designed to tackle several challenges: unmeasured mediator outcome confounding, multilevel time series analysis, and the estimation of functional causal effects. The approach is grounded in multilevel structural equation modeling, complemented by relaxed likelihood estimation methods. Interestingly, certain causal estimates, typically unobtainable in simpler data structures, become identifiable in the more complex data setting. Proof of the asymptotic properties of the estimators is provided, and the numerical properties are illustrated through empirical analysis. Additionally, real fMRI data is utilized to demonstrate the practical effectiveness of the proposed framework.

E0424: **Improving statistical power of multi-modal associations via de-variation**
Presenter: **Jun Young Park**, University of Toronto, Canada
Co-authors: Ruyi Pan, Yinqiu He
Understanding the interplay between different modalities of brain MRI data is crucial for unravelling the complexities of brain structure and function. Existing statistical association tests for two random vectors are often limited in fully capturing dependencies between modalities, particularly by overlooking correlation structures within each modality, leading to the potential loss of statistical power. A novel approach termed de-variation is proposed to address this limitation. De-variation is a simple yet effective preprocessing method that leverages a penalized low-rank factor model to capture within-modality dependencies. Theoretical analyses and simulation studies show (i) its powerful performance when within-modality correlations impact signal-to-noise ratios and (ii) its robustness when these are absent. De-variation is then applied to brain imaging-driven phenotypes (IDPs) derived from functional, structural, and diffusion MRI from the UK Biobank to show its promising performance.

E0625: **Bayesian inference on brain-computer interfaces via GLASS**
Presenter: **Jian Kang**, University of Michigan, United States
Brain-computer interfaces (BCIs), specifically the P300 BCI, enable direct brain-computer communication. Classifying target vs. non-target stimuli from electroencephalogram (EEG) signals, with their low signal-to-noise ratio and complex correlations, is challenging, particularly for users with severe physical disabilities. The Gaussian latent channel model is proposed with sparse time-varying effects (GLASS) within a Bayesian framework, designed to improve classification in imbalanced data. GLASS mitigates spatial correlations through latent channel decomposition and employs a soft-thresholder Gaussian process for sparse, smooth temporal effects. Demonstrated improvements in ALS participants, GLASS highlights critical EEG channels in parietal and occipital regions, corroborating literature. An efficient gradient-based variational inference algorithm and a user-friendly Python module are also introduced for computational ease.

E0914: **Machine learning to the mean and its correction: An application to imaging-based brain age prediction**
Presenter: **Shuo Chen**, University of Maryland, United States
Co-authors: Hwiyoung Lee
A commonly observed issue in machine learning models predicting continuous outcomes is reported, referred to as "machine learning to the mean". When applying a machine learning model built on a training dataset to a testing dataset, the difference between the predicted continuous outcome and the true value is often negatively correlated with the true value. For observations with continuous outcomes much smaller or greater than the mean, the predicted values tend to be automatically warped toward the mean. It is shown that this issue can be caused by the commonly objective function of minimizing the square loss. A new constrained strategy is proposed to correct the bias and develop computationally efficient algorithms for implementation. The new approach is applied to predicting brain age by brain imaging data while addressing the well-known issue of chronological age-related bias in brain age prediction.

E0983: **Advancing statistical analyses for living systematic reviews**
Presenter: **Lifeng Lin**, University of Arizona, United States
A living systematic review (LSR) is a progressive approach aimed at providing ongoing updates and instant synthesis of evidence. Trial sequential analysis (TSA) is an important tool for evaluating the sufficiency of evidence gathered in an LSR. It utilizes trial sequential monitoring boundaries to assess the effectiveness of an intervention and futility boundaries to determine if the intervention does not significantly differ from the control. While TSAs have been increasingly popular, their reproducibility is currently limited due to a lack of detailed information on their assumptions. Moreover, existing TSA methods face challenges due to their significant reliance on interim analyses of randomized controlled trials, which typically involve more homogeneous participant groups than those found in meta-analyses. The purpose is to introduce new methods aimed at preventing premature terminations of LSRs, thereby enabling more robust evidence syntheses. Numerical studies indicate that these proposed methods are more reliable than current methods.

EO027 Room 313 RESEARCH FRONTIERS ON TIME SERIES AND MULTIVARIATE DATA Chair: Ting Zhang

E0582: **Inference for quantile change points in high-dimensional time series**
Presenter: **Mengyu Xu**, University of Central Florida, United States
Co-authors: Likai Chen, Jiaqi Li
Change-point detection methods based on quantiles can effectively detect changes in extreme values. A novel change-point detection scheme that utilizes fixed quantiles of moving sums from high-dimensional time series data is proposed. The approach employs a moving sum (MOSUM) test statistic aggregating the component series by the ℓ^∞ norm. The asymptotic properties of the proposed test statistic are investigated in the context of weak temporal-dependent high-dimensional time series while also allowing for strong and weak cross-sectional dependence. The analysis relies on a powerful uniform Bahadur representation result. Specifically, the existing uniform Bahadur representation is extended to the high-dimensional setting for dependent data. A simulation study demonstrates the effectiveness of the approach.

E0664: **Clustering multivariate extremes**
Presenter: **Shuyang Bai**, University of Georgia, United States
Co-authors: He Tang, Shiyuan Deng
The estimation of multivariate extreme models with a discrete spectral measure is investigated via clustering techniques such as spherical k-means. A method is proposed to consistently select the number of clusters. Large deviation analysis is also discussed, which assesses the quality of convergence of spectral estimation and how to convert spectral measure estimation to coefficient estimation of extremal factor models.

E0762: **High-dimensional clustering via latent semiparametric mixture models**
Presenter: **Boxiang Wang**, University of Iowa, United States
Cluster analysis is a fundamental task in machine learning. Several clustering algorithms have been extended to handle high-dimensional data by incorporating a sparsity constraint in estimating a mixture of Gaussian models. Though it makes some neat theoretical analysis possible, this approach is arguably restrictive for many applications. A novel latent variable transformation mixture model is introduced for clustering in which a mixture of Gaussians is assumed after some unknown monotone data transformation. A new clustering algorithm named CESME is developed for high-dimensional clustering under the assumption that optimal clustering admits a sparsity structure. The use of unspecified transformation makes the model far more flexible than the classical mixture of Gaussians. On the other hand, the transformation also brings quite a few technical challenges to the model estimation as well as the theoretical analysis of CESME. A comprehensive analysis of CESME is presented, including identifiability, initialization, algorithmic convergence, and statistical guarantees on clustering. Leveraging such a transition, a data-adaptive procedure is developed and substantially improves the computational efficiency of CESME. Extensive numerical study and real data analysis show that CESME outperforms the existing high-dimensional clustering algorithms.

E0764: **Matrix denoising and completion based on Kronecker product approximation**
Presenter: **Han Xiao**, Rutgers University, United States
The problem of matrix denoising and completion is considered induced by the Kronecker product decomposition. Specifically, an approximation to a given matrix is proposed by the sum of a few Kronecker products of matrices, which is referred to as the Kronecker product approximation (KoPA). Because the Kronecker product is an extension of the outer product from vectors to matrices, KoPA extends the low-rank matrix approximation and includes it as a special case. Compared with the latter, KoPA also offers greater flexibility since it allows the user to choose the configuration, which are the dimensions of the two smaller matrices forming the Kronecker product. On the other hand, the configuration to be used is usually unknown and needs to be determined from the data in order to achieve the optimal balance between accuracy and parsimony. The use of extended information criteria is proposed to select the configuration. Under the paradigm of high dimensional analysis, it is shown that the proposed procedure is able to select the true configuration with probability tending to one, under suitable conditions on the signal-to-noise ratio. The superiority of KoPA is demonstrated over the low-rank approximations through numerical studies and several benchmark image examples.

E0785: **Revisiting Poisson autoregressive models: Structure and statistical inference**
Presenter: **Dong Li**, Tsinghua University, China
The first-order stationary Poisson autoregression (PAR) is one of the most classical count time series models and has been widely studied. However, few researchers pay attention to nonstationary PAR. PAR is revisited, and some novel results are provided on asymptotical behaviors of the intensity process under nonstationarity. Further, the maximum likelihood estimation is considered in a unified framework of stationary and nonstationary cases, and its asymptotics are established. Monte Carlo simulation studies are conducted to assess the finite-sample performance of the MLE.

EO154 Room 405 STATISTICAL LEARNING AND NONPARAMETRIC METHODS: THEORY AND PRACTICE Chair: Guannan Wang

E0254: **Inference for quantile mediation effects in the presence of complex confounding via deep neural networks**
Presenter: **Shuoyang Wang**, University of Louisville, United States
Co-authors: Yuan Huang, Runze Li
Traditional mediation analysis methods face challenges when dealing with a large number of mediators. In practice, these challenges can be compounded by outliers and the complex relationships introduced by confounders. To address these issues, a novel quantile-based partially linear mediation analysis method is proposed that can handle high-dimensional mediators and introduce deep neural network techniques to model intricate relationships in confounders. Unlike most existing works that focus on mediator selection, inference on mediation effects is emphasized. Theoretical analysis shows that the proposed procedure controls type I error rates for hypothesis testing on mediation effects. When the dimension of the mediator is high, the proposed method consistently selects important features in the outcome model. Numerical studies show that the proposed method outperforms existing approaches under a variety of settings, demonstrating its versatility and reliability as a modeling tool for complex data. The application of the proposed method to study DNA methylation's mediation effect of childhood trauma on cortisol stress reactivity reveals previously undiscovered relationships by providing a comprehensive profile of the relationship at various quantiles.

E0462: **Nonparametric biomarker based treatment selection with reproducibility data**
Presenter: **Xiao Song**, University of Georgia, United States
Biomarkers for treatment selection are considered for evaluation under assay modification. Survival outcome, treatment, and Affymetrix gene expression data were attained from cancer patients. Consider migrating a gene expression biomarker to the Illumina platform. A recent novel approach allows a quick evaluation of the migrated biomarker with only a reproducibility study needed to compare the two platforms, achieved by treating the original biomarker as an error-contaminated observation of the migrated biomarker. However, its assumptions of a classical measurement error model and a linear predictor for the outcome may not hold. Ignoring such model deviations may lead to sub-optimal treatment selection or failure to identify effective biomarkers. To overcome such limitations, a nonparametric logistic regression is adopted to model the relationship between the event rate and the biomarker, and the deduced marker-based treatment selection is optimal. A nonparametric relationship is further assumed between the migrated and original biomarkers, and it is shown that the error-contaminated biomarker leads to sub-optimal treatment

selection compared to the error-free biomarker. The estimation via B-spline approximation is obtained. The approach is assessed by simulation studies and demonstrated through application to lung cancer data.

E0528: **Penalized deep partially linear cox models with application to CT scans of lung cancer patients**
Presenter: **Yuming Sun**, College of William and Mary, United States
Co-authors: Yi Li, Jian Kang
Partially linear Cox models have gained popularity for survival analysis by dissecting the hazard function into parametric and nonparametric components, allowing for the effective incorporation of both well-established risk factors (such as age and clinical variables) and emerging risk factors (e.g., image features) within a unified framework. However, when the dimension of parametric components exceeds the sample size, the task of model fitting becomes formidable, while nonparametric modeling grapples with the curse of dimensionality. A novel penalized deep partially linear Cox model (penalized DPLC) is proposed, which incorporates the smoothly clipped absolute deviation (SCAD) penalty to select important texture features and employs a deep neural network to estimate the nonparametric component of the model. The convergence and asymptotic properties of the estimator are proven and compared to other methods through extensive simulation studies, evaluating its performance in risk prediction and feature selection. The proposed method is applied to the National Lung Screening Trial dataset to uncover the effects of key clinical and imaging risk factors on patients' survival. Findings provide valuable insights into the relationship between these factors and survival outcomes.

E0592: **Sparse functional linear discriminant analysis for high-dimensional predictors**
Presenter: **Limeng Liu**, University of Minnesota, United States
Co-authors: Guannan Wang, Sandra Safo
Functional data analysis (FDA) aims to analyze functional datasets where observations are not individual data points but functions. Most FDA approaches are for continuous time domains with a single variable. However, multivariate data measured at dense or sparse time points are collected in biomedical research with a typical goal of finding time-variant profiles to distinguish between classes. Fisher's linear discriminant analysis (LDA) is a popular multivariate dimension reduction method for finding linear combinations of variables that optimally separate classes. Most existing LDA methods for the FDA only apply to a single variable or binary classes and cannot identify variables discriminating between classes over time. Sparse functional linear discriminant analysis (SFLDA) is proposed to find linear combinations of multiple functional predictors that optimally discriminate between two or more classes over time and identify functional predictors. Simulations are used to demonstrate the effectiveness of SFLDA. SFLDA is applied to the inflammatory bowel disease study, and a personal omics profiling dataset is integrated to identify longitudinal biomarkers of disease progression.

E0760: **DeepIDA-GRU: A deep learning pipeline for integrative discriminant analysis of cross-sectional and longitudinal**
Presenter: **Sandra Safo**, University of Minnesota, United States
Biomedical research now commonly integrates diverse data types or views from the same individuals to better understand the pathobiology of complex diseases, but the challenge lies in meaningfully integrating these diverse views. Existing methods often require the same type of data from all views (cross-sectional data only or longitudinal data only) or do not consider any class outcome in the integration method, presenting limitations. To overcome these limitations, a pipeline that harnesses the power of statistical and deep learning methods is developed to integrate cross-sectional and longitudinal data from multiple sources. Additionally, it identifies key variables contributing to the association between views and the separation among classes, providing deeper biological insights. This pipeline includes variable selection/ranking using linear and nonlinear methods, feature extraction using functional principal component analysis and Euler characteristics, and joint integration and classification using dense feed-forward networks and recurrent neural networks. This pipeline is applied to cross-sectional and longitudinal multiomics data (metagenomics, transcriptomics, and metabolomics) from an inflammatory bowel disease (IBD) study and microbial pathways, metabolites, and genes are identified that discriminate by IBD status, providing information on the etiology of IBD. Simulations are conducted to compare the two feature extraction methods.

EO137 Room 406 RECENT ADVANCES IN DIMENSION REDUCTION Chair: Jun Song

E0414: **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 the evaluation of 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 graphical model is based on conditional independence, given a set of sufficient predictors with a substantially reduced dimension. In this way, the curse of dimensionality that comes with a high-dimensional kernel is avoided. The population-level properties, convergence rate, and variable selection consistency of the estimate are developed. By simulation comparisons and an analysis of the DREAM 4 Challenge data set, the method is demonstrated to outperform the existing methods when the Gaussian or copula Gaussian assumptions are violated, and its performance remains excellent in the high-dimensional setting.

E0428: **Some theory about efficient dimension reduction with respect to interaction between two responses**
Presenter: **Wei Luo**, Zhejiang University, China
Efficient dimension reduction with respect to the interaction between two response variables, which facilitates statistical analysis in multiple important application scenarios, was initially discussed in a recent study. The efficient dimension reduction subspaces were introduced, and, under mild conditions on the predictor, they were equated with the family of dual inverse regression subspaces. Besides the general framework, however, limited theory has been proposed to uncover the mystery of these spaces. A thorough characterization of the family of dual inverse regression subspaces is proposed, including their uniform lower and upper bounds, their explicit forms, their consistent and exhaustive estimation, some interesting special cases, and certain subfamilies that have desired sparsity. In addition, some of these results are extended to provide more insights into the efficient dimension reduction subspaces under the general settings, including their uniform lower and upper bounds and the sufficient and necessary conditions for the uniqueness of the space that are assessable in practice. These results largely complete the theoretical foundation for the new type of dimension reduction and, as such, enhance its applicability in statistical problems such as missing data analysis and causal inference; the latter is illustrated by simulation studies and a real data example at the end.

E0469: **Sparse sufficient dimension reduction via penalized gradient learning for composite quantile regressions**
Presenter: **Seogyoung Lee**, Korea University, Korea, South
Co-authors: Seung Jun Shin
It is essential to identify informative features in machine learning. Among many others, sufficient dimension reduction (SDR) has gained great attention due to its promising performance in extracting a small number of features in both regression and classification problems. SDR, however, often suffers from poor interpretability, especially in high dimensions, since all predictors have non-zero loading. A novel sparse SDR method, called sparse-qOPG, is proposed by extending the idea of qOPG. The qOPG elegantly employs a series of quantile regressions with different quantile levels to conduct SDR, i.e., estimate the central subspace. The qOPG objective function is first re-expressed using the gradient of the

quantile functions in the reproducing kernel Hilbert space. A non-convex group penalty is then employed to the gradient term to obtain a sparse basis estimator for the central subspace. The selection consistency of the sparse qOPG is established, and its promising performance is demonstrated in both synthetic and real data applications.

E0987: **Sufficient dimension reduction for high dimensional nonlinear vector autoregressive models**
Presenter: **Jiaying Weng**, Bentley University, United States
The vector autoregressive model is a fundamental tool for modeling multivariate time series data. It is widely used in various fields, such as economics, finance, and climate studies. One of the challenges in modeling high-dimensional time series data is the curse of dimensionality, particularly when incorporating multiple time series and increasing the order of the vector autoregressive model. The purpose is to explore sufficient dimension reduction in nonlinear vector autoregressive models, where the present vector is influenced by multiple indices defined on past lags. The proposed sufficient dimension reduction approaches aim to identify these indices from the past information of the multivariate time series. Specifically, several linear combinations of the covariate vector are sought, comprising past lags, such that the current response vector is conditionally independent of the covariate vector given the linear combination. The linear combinations depict the time series' central subspace, the ultimate goal of sufficient dimension reduction. A time series martingale difference divergence matrix method is proposed for the nonlinear vector autoregressive models to estimate the central subspace. For high-dimensional time series, a sparse estimation procedure is developed to identify the central subspace using a penalized optimization problem.

E1071: **Beta regression for double-bounded response with correlated high-dimensional covariates**
Presenter: **Jianxuan Liu**, Syracuse University, United States
Continuous responses measured on a standard unit interval (0,1) are ubiquitous in many scientific disciplines. Statistical models built upon a normal error structure do not generally work because they can produce biased estimates or result in predictions outside either bound. In real-life applications, data are often high-dimensional, correlated, and consist of a mixture of various data types. Little literature is available to address the unique data challenge. A semiparametric approach is proposed to analyze the association between a double-bounded response and high-dimensional correlated covariates of mixed types. The proposed method makes full use of all available data through one or several linear combinations of the covariates without losing information from the data. The only assumption made is that the response variable follows a beta distribution, and no additional assumption is required. The resulting estimators are consistent and efficient. The proposed method is illustrated in simulation studies and is demonstrated in a real-life data application. The semiparametric approach contributes to the sufficient dimension reduction literature for its novelty in investigating double-bounded response, which is absent in the current literature. A new tool is also provided for data practitioners to analyze the association between a popular unit interval response and mixed types of high-dimensional correlated covariates.

EO079 Room 408 ADVANCEMENT IN STATISTICAL GENETICS AND GENOMICS STUDY Chair: Xinyi Li

E0545: **Advancing graph neural networks for disease classification and feature selection in high-dimensional data**
Presenter: **Tiantian Yang**, University of Idaho, United States
Omics data play crucial roles in exploring disease pathways, forecasting clinical outcomes, and gaining insights for disease classification. However, the significant challenge of dealing with a relatively small number of samples and a large number of features complicates the development of predictive models for omics data analysis. This challenge arises from inherent sparsity in biological networks and unknown feature interactions, adding further complexities. The advent of graph neural networks (GNN) helps alleviate these challenges by incorporating known functional relationships into a graph. However, many existing GNN models utilize graphs either from existing networks or generated ones alone, limiting model effectiveness. To overcome this restriction, an innovative GNN model is proposed that integrates information from both externally and internally generated feature graphs. The model is extensively tested through simulations and real data applications, confirming its superior performance in classification tasks compared to existing state-of-the-art baseline models. Furthermore, the GNN model can select features with meaningful interpretations in the biomedical context.

E0621: **Pseudotime analysis for time-series single-cell sequencing and imaging data**
Presenter: **Gang Li**, University of Washington at Seattle, United States
Many single-cell RNA-sequencing studies have collected time-series data to investigate transcriptional changes concerning various notions of biological time, such as cell differentiation, embryonic development, and response to stimulus. Accordingly, several unsupervised and supervised computational methods have been developed to construct single-cell pseudotime embeddings for extracting the temporal order of transcriptional cell states from these time-series scRNA-seq datasets. However, existing methods, such as psupertime, suffer from low predictive accuracy, and this problem becomes even worse when we try to generalize to other data types, such as scATAC-seq or microscopy images. Sceptic, a support vector machine model is proposed for supervised pseudotime analysis. Sceptic is demonstrated to achieve significantly improved prediction power (accuracy improved by 1.4 38.9%) for six publicly available scRNA-seq data sets over state-of-the-art methods, and Sceptic also works well for single-nucleus image data.

E0593: **A unified quantile framework reveals nonlinear heterogeneous transcriptome-wide associations**
Presenter: **Tianying Wang**, Colorado State University, United States
Co-authors: Iuliana Ionita-Laza, Ying Wei
Transcriptome-wide association studies (TWAS) are powerful tools for identifying putative causal genes by integrating genome-wide association studies and gene expression data. However, most TWAS methods focus on linear associations between genes and traits, ignoring the complex nonlinear relationships that exist in biological systems. To address this limitation, a novel quantile transcriptomics framework, QTWAS, is proposed that takes into account the nonlinear and heterogeneous nature of gene-trait associations. The approach integrates a quantile-based gene expression model into the TWAS model, which allows for the discovery of nonlinear and heterogeneous gene-trait associations. By conducting comprehensive simulations and examining various psychiatric and neurodegenerative traits, it is demonstrated that the proposed model outperforms traditional techniques in identifying gene-trait associations. Additionally, QTWAS can uncover important insights into nonlinear relationships between gene expression levels and phenotypes, complementing traditional TWAS approaches. Applications are further shown to 100 continuous traits from the UK Biobank and ten binary traits related to brain disorders.

E0660: **Additive tree flows for density estimation and two-sample comparison**
Presenter: **Naoki Awaya**, Stanford University, United States
Co-authors: Li Ma
A new nonparametric method is proposed for two fundamental unsupervised learning tasks: density estimation and two-sample comparison, which are known to be challenging in high-dimensional settings. Motivated by the recent success in normalizing flow methods in the machine learning community, a new class of flow models consisting of "trees", i.e., conditional density functions is introduced, defined on recursive partition structures. The novelty of the proposed method is the introduction of a new efficient sequential algorithm that works like the boosting algorithm typically used for supervised learning. As in the classical boosting algorithm, the proposed algorithm repeatedly transforms the observations ("residuals") and fits a new density function to the residuals. It is shown that the empirical performance of our proposed method is competitive

with the deep neural network methods, but the computational cost is drastically improved. Its application is also presented to biological data such as microbiome data.

E1061: **Double weighting scheme for k-nearest neighbors for binary classification of high-dimensional gene expression data**
Presenter: **Zardad Khan**, United Arab Emirates University, United Arab Emirates
Co-authors: Saeed Aldahmani, Amjad Ali, Hailiang Du
The accurate classification of tissue samples in high-dimensional gene expression datasets can be challenging due to the large number of genes, many of which do not significantly contribute to the classification. To address this issue, a new method called double-weighted k-nearest neighbors (DW-k-NN) is introduced. This method is specifically designed for gene expression data and incorporates feature weights that are derived from the differential expression of genes between classes. By using an exponential function to calculate estimated weighted distances, informative features have a greater impact, while less or non-informative features have a decreased impact. The test point is assigned the class label with the largest sum of outputs for both classes separately. DW-k-NN aims to achieve robust classification results for high-dimensional gene expression datasets by considering the proposed weighting scheme based on genes' capability to express differentially. Experimental evaluations have demonstrated the effectiveness of DW-k-NN in accurately classifying gene expression datasets when compared to several other k-NN-based methods. Overall, DW-k-NN presents a promising approach to gene expression data analysis through the two-fold weighted distance calculation strategy.

EO052 Room 411 (Virtual sessions) ADVANCES IN MACRO- AND FINANCIAL ECONOMETRICS Chair: Toshiaki Watanabe

E0269: **Estimating trend inflation in a regime-switching Phillips curve**
Presenter: **Jouchi Nakajima**, Hitotsubashi University, Japan
A regime-switching Phillips curve model is developed to estimate trend inflation. Extending the earlier work, trend inflation, the slope of the Phillips curve, and the oil price pass-through rate are allowed to follow a regime-switching process. An empirical analysis using Japan's consumer price index illustrates that including the oil price and its time-varying pass-through rate improves the model's ability to forecast inflation. The empirical results also show that the obtained trend inflation highly correlates with firms' inflation expectations.

E0393: **Multivariate realized stochastic volatility model using time varying coefficient characteristic factor regression**
Presenter: **Tsunehiro Ishihara**, Takasaki City University of Economics, Japan
The computational cost to estimate a high-dimensional time-varying correlation volatility model is often expensive. A multivariate stochastic volatility model is proposed with observed characteristic factors and their realized covariance. To reduce the computational time, the high-dimensional model is split into conditional univariate models and low-dimensional characteristic factor multivariate models and estimated in parallel. For conditional univariate models, the time-varying coefficient characteristic factor regression model is proposed with stochastic volatility, and their realized measurements are introduced into the model. For the low-dimensional multivariate stochastic volatility model for characteristic factors, the matrix exponential realized stochastic volatility model is used. As an illustrative example, the model to 33-dimensional Japanese sector indices and market are applied, as well as value and size factors. Model comparison is conducted with other multivariate models.

E0604: **Stochastic volatility in mean: Efficient analysis by a generalized mixture sampler**
Presenter: **Daichi Hiraki**, University of Tokyo, Japan
Co-authors: Siddhartha Chib, Yasuhiro Omori
The simulation-based Bayesian analysis of stochastic volatility is considered in mean (SVM) models. Extending the highly efficient Markov chain Monte Carlo mixture sampler for the SV model proposed in prior studies, an accurate approximation of the non-central chi-squared distribution is developed as a mixture of thirty normal distributions. Under this mixture representation, the parameters and latent volatilities are sampled in one block. A correction of the small approximation error is also detailed by using additional Metropolis-Hastings steps. The proposed method is extended to the SVM model with leverage. The methodology and models are applied to excess holding yields in empirical studies, and the SVM model with leverage is shown to outperform competing volatility models based on marginal likelihoods.

E0564: **Cross-sectional analysis of stock returns using option-implied tail risk**
Presenter: **Masato Ubukata**, Meiji Gakuin University, Japan
The purpose is to investigate whether an option-implied market tail risk has substantial predictive power for the cross-section of average returns in the Japanese stock market. A time-varying option-implied jump variation is calculated using the Nikkei 225 options data from January 2006 to March 2024. Monthly predictive regressions are run for each stock. Stocks are then sorted into several portfolios based on their estimated tail risk loadings. Average monthly value- and equal-weighted portfolio returns are tracked, and the hypothesis that tail risk helps explain differences in expected returns across stocks is tested.

E0569: **Dynamic Bayesian networks with conditional dynamics in edge addition and deletion**
Presenter: **Mike So**, The Hong Kong University of Science and Technology, Hong Kong
Co-authors: Shun Hin Chan, Amanda Chu
A dynamic Bayesian network framework is presented that facilitates intuitive gradual edge changes. Two conditional dynamics are used to model the edge addition and deletion, as well as edge selection separately. Unlike previous research that uses a mixture network approach, which restricts the number of possible edge changes or structural priors to induce gradual changes, which can lead to unclear network evolution, the model induces more frequent and intuitive edge change dynamics. Markov chain Monte Carlo (MCMC) sampling is employed to estimate the model structures and parameters and demonstrate the model's effectiveness in a portfolio selection application.

EC281 Room 207 CAUSAL INFERENCE Chair: Cy Sin

E0751: **Individual treatment rule estimation with M-learning**
Presenter: **Bo Lu**, The Ohio State University, United States
Individualized treatment rules (ITRs) tailor treatments to individuals based on their unique characteristics to optimize clinical outcomes. Current approaches use outcome modeling or propensity score weighting to control confounding in complex medical data. To avoid model misspecification and the impact of extreme weights, matched-learning (M-learning) was recently proposed for continuous outcomes. We expand the existing M-learning methodology to estimate optimal ITRs under the right censored data. Matched sets are constructed for individuals by comparing observed times, and an inverse probability censoring weight is incorporated into the value function to handle censored observations. Additionally, a full matching design is proposed in M-learning to reduce the potential overuse of a single subject when matching with replacement. The proposed value function is demonstrated to be unbiased for the true value function without censoring. To assess the method's performance, an extensive simulation study is conducted to compare the proposed method with the existing M-learning approach and a weighted learning approach. Results are evaluated based on winning probabilities and estimated values. The simulation reveals that all methods are generally fine in the absence of unmeasured confounders, and different methods show somewhat different performance under various scenarios. But their performances drop substantially in the presence of unmeasured confounders.

E0947: **Double robust Bayesian inference on average treatment effects**
Presenter: **Ruixuan Liu**, Chinese University of Hong Kong, Hong Kong

A double robust Bayesian inference procedure is proposed on the average treatment effect under unconfoundedness. The robust Bayesian approach involves two important modifications: first, the prior distributions of the conditional mean function are adjusted; second, the posterior distribution of the resulting ATE is corrected. Both adjustments make use of pilot estimators motivated by the semiparametric influence function for ATE estimation. Asymptotic equivalence of the Bayesian procedure and efficient frequentist ATE estimators are proven by establishing a new semiparametric Bernstein-von Mises theorem under double robustness, i.e., the lack of smoothness of conditional mean functions can be compensated by high regularity of the propensity score and vice versa. Consequently, the resulting Bayesian credible sets form confidence intervals with asymptotically exact coverage probability. In simulations, our double robust Bayesian procedure leads to significant bias reduction of point estimation over conventional Bayesian methods and more accurate coverage of confidence intervals compared to existing frequentist methods. The method is illustrated in an application to the national supported work demonstration.

E1031: **Randomization inference on policy assignments**
Presenter: **EunYi Chung**, University of Illinois at Urbana Champaign, United States

Randomization inference is quickly becoming a widely used statistical approach in the social, behavioral, and natural sciences. In the setting of regression kink designs, propose a randomization test that is constructed based on random kink points assigned by a policy. The limitation of their method is that researchers are assumed to know the policy data-generating process that selects the kink point and use that distribution to simulate critical values for the test. Although the randomization test has an exact size under such an assumption, the test is no longer valid, even asymptotically, if the researcher misspecifies the policy assignment distribution. The first contribution is to provide a general framework for randomization tests based on policy assignments of individuals into treatment and control groups. The framework includes not only regression discontinuity and kink designs but also bunching and difference-in-differences models. The proposed test controls size in large samples even when the researcher does not know the policy assignment distribution; it retains the exactness property of the randomization test when the policy distribution is known. Simulations show desirable finite sample properties, and an empirical application illustrates the procedure.

E1043: **Doubly robust counterfactual classification**
Presenter: **Kwangho Kim**, Korea University, Korea, South
Co-authors: Edward Kennedy, Jose Zubizarreta

Counterfactual classification is studied as a new tool for decision-making under hypothetical (contrary to fact) scenarios. A doubly robust nonparametric estimator is proposed for a general counterfactual classifier, where flexible constraints can be incorporated by casting the classification problem as a nonlinear mathematical program involving counterfactuals. Next, the rates of convergence of the estimator are analyzed, and a closed-form expression is provided for its asymptotic distribution. The analysis shows that the proposed estimator is robust against nuisance model misspecification and can attain fast root-n rates with tractable inference even when using nonparametric machine learning approaches. The empirical performance of the methods is studied by simulation and application on recidivism risk prediction.

E1111: **Rematching estimators for average treatment effects**
Presenter: **Lam Lam Hui**, The Chinese University of Hong Kong, Hong Kong
Co-authors: Kin Wai Chan

Matching estimators are widely applied in practice for their great intuitive appeal. However, simple matching estimators with a fixed number of matches (M_0) are generally inefficient. Matching estimators with a variable number of matches are proposed to gain efficiency via rematching. Rather than increasing M_0 to gain precision, which introduces a substantial increase in bias, the key is to rematch the treated units from the opposite direction to utilize unmatched control units. The proposed rematching estimators are applicable to both the average treatment effect and its counterpart for the treated population. They are proven asymptotically valid and uniformly more efficient than matching estimators. Simulation results confirm that the proposed rematching estimators substantially improve the simple matching estimators in finite samples. As an empirical illustration, we apply the estimators proposed in this article to the National Supported Work data.

Thursday 18.07.2024 14:00 - 15:40 Parallel Session H – EcoSta2024

EI006 Room 106 RECENT ADVANCES IN ECONOMETRICS Chair: Massimiliano Caporin

E0156: **Ups and (draw)downs**
Presenter: **Tommaso Proietti**, University of Roma Tor Vergata, Italy
The concept of drawdown quantifies the potential loss in the value of a financial asset when it deviates from its historical peak. It plays an important role in evaluating market risk, portfolio construction, assessing risk-adjusted performance, and trading strategies. A novel measurement framework is introduced that produces, along with the drawdown and its dual (the drawup), two Markov chain processes representing the current lead time with respect to the running maximum and minimum, i.e., the number of time units elapsed from the most recent peak and trough. Under relatively unrestrictive assumptions regarding the returns process, the chains are homogeneous and ergodic. It is shown that, together with the distribution of asset returns, they determine the properties of the drawdown and drawup time series in terms of size, serial correlation, persistence, and duration. Furthermore, they form the foundation of a new algorithm for dating peaks and troughs of the price process, delimiting bear and bull market phases. Other contributions deal with out-of-sample prediction and robust estimation of the drawdown.

E0157: **Understanding fluctuations through multivariate circulant singular spectrum analysis**
Presenter: **Pilar Poncela**, Universidad Autonoma de Madrid, Spain
Co-authors: Eva Senra, Juan Bogalo
The decomposition of multichannel signals into their underlying components is a problem of key interest in many disciplines. It involves the analysis of nonlinear and nonstationary time series. After extracting the underlying latent components, the need is to identify their frequencies of oscillation. Many successful multivariate methods identify the frequencies after the decomposition problem has been solved. However, they might face difficulties matching extracted components with frequencies. Multivariate circulant singular spectrum analysis (MCiSSA) is developed, a self-identifying procedure for the frequencies without the need to further process the data. Seeing MSSA procedures as a generalization of principal components that includes series and their lags, the key is the use of block circulant matrices instead of the variance-covariance matrix, which matches eigenvalues with frequencies. MCiSSA performs a double diagonalization of that block circulant matrix that uncovers cross-section relations per frequency. The procedure works well in synthetic examples with latent signals modulated in amplitude and/or frequency. Finally, MCiSSA is applied to real data, firstly to energy commodity prices that are dominated by European gas at all frequencies. The second application shows the growing trend of temperatures in 12 South European cities and their changing pattern in seasonality.

E0277: **Variational inference for a Bayesian nonparametric model with structural breaks**
Presenter: **Yong Song**, University of Melbourne, Australia
Co-authors: John Maheu, Xuan Vu
A variational inference (VI) algorithm is proposed for a dynamic Bayesian nonparametric framework, in which the standard Markov chain Monte Carlo (MCMC) method is impractical for large data sets due to its high computational cost. Because this dynamic nonparametric framework includes parameter structural breaks, the conditional analytic solution from the exponential family cannot be applied. Instead, a novel structured VI through Rao-Blackwellisation is proposed. Through both the simulation study and the application to the US banking data, the usefulness of the new method is shown.

EO125 Room 102 ADVANCES IN HIGH-DIMENSIONAL ECONOMETRICS Chair: Manabu Asai

E0196: **Scalable estimation of multinomial response models with uncertain consideration sets**
Presenter: **Kenichi Shimizu**, University of Alberta, Canada
A standard assumption in fitting unordered multinomial response models for J mutually exclusive nominal categories on cross-sectional or longitudinal data is that the responses arise from the same set of J categories between subjects. However, when responses measure a choice made by the subject, it is more appropriate to assume that the distribution of multinomial responses is conditioned on a subject-specific consideration set, where this consideration set is drawn from the power set of the set of J categories. Because the cardinality of this power set is exponential in J, estimation is infeasible in general. An approach to overcoming this problem is provided. A key step in the approach is a probability model over consideration sets based on a general representation of probability distributions on contingency tables, which results in mixtures of independent consideration models. Although the support of this distribution is exponentially large, the posterior distribution over consideration sets given parameters is typically sparse and is easily sampled in an MCMC scheme. Posterior consistency of the parameters of the conditional response model and the distribution of consideration sets are shown. The methodology's effectiveness is documented in simulated longitudinal data sets with $J = 100$ categories and real data from the cereal market with $J = 68$ brands.

E0299: **Sparse factor models of high dimension**
Presenter: **Benjamin Poignard**, Osaka University, Japan
Co-authors: Yoshikazu Terada
The estimation of the factor model-based variance-covariance matrix is considered when the factor loading matrix is assumed sparse. The estimation problem is recast as a penalized M-estimation criterion where the identification issue of the factor loading matrix is accounted for while fostering sparsity in potentially all its entries. Consistency and recovery of the true zero entries are established when the number of parameters is diverging. These theoretical results are supported by simulation experiments, and the relevance of the proposed method is illustrated by real data applications.

E0565: **High-dimensional sparse factor multivariate stochastic volatility models**
Presenter: **Manabu Asai**, Soka University, Japan
Co-authors: Benjamin Poignard
Factor models are useful for reducing the dimension of variables. Factor structure is accommodated on high-dimensional multivariate stochastic volatility (MSV) models to construct the space factor MSV (fMSV) model. Two sparse factor models are considered; one assumes sparsity for the covariance matrix of the idiosyncratic errors, while the other has sparsity on the factor loading matrix. Some theoretical results are provided for the fMSV models. Using simulated and real data, the in-sample and out-of-sample forecasting performance is examined, comparing the fMSV models with DCC and BEKK models.

E0693: **Testing heteroskedasticity in high-dimensional linear regression**
Presenter: **Akira Shinkyu**, Shiga University, Japan
A new testing procedure of heteroskedasticity is proposed in high-dimensional linear regression, where the number of covariates can be larger than the sample size. The testing procedure is based on residuals of the Lasso. The test statistic is demonstrated to have asymptotic normality under the null hypothesis of homoscedasticity. Simulation results show that the proposed testing procedure obtains accurate empirical sizes and powers.

EO226 Room 103 RECENT ADVANCES IN LARGE-SCALE DATA ANALYSIS Chair: Yaowu Liu

E0562: **Nonparametric learning for 3D point cloud data**
Presenter: **Xinyi Li**, Clemson University, United States
Co-authors: Shan Yu, Yueying Wang, Guannan Wang, Lily Wang, Ming-Jun Lai
In recent years, there has been an exponentially increased amount of point clouds collected with irregular shapes in various areas. Motivated by the importance of solid modeling for point clouds, a novel and efficient smoothing tool is developed based on multivariate splines over the triangulation to extract the underlying signal and build up a 3D solid model from the point cloud. The proposed method can denoise or deblur the point cloud effectively, provide a multi-resolution reconstruction of the actual signal, and handle sparse and irregularly distributed point clouds to recover the underlying trajectory. In addition, the method provides a natural way of reducing numerosity data. The theoretical guarantees of the proposed method are established, including the convergence rate and asymptotic normality of the estimator, and show that the convergence rate achieves optimal nonparametric convergence. A bootstrap method is also introduced to quantify the uncertainty of the estimators. Through extensive simulation studies and a real data example, the superiority of the proposed method is demonstrated over traditional smoothing methods in terms of estimation accuracy and data reduction efficiency.

E0599: **A robust integrated mean variance correlation and its use in high dimensional data analysis**
Presenter: **Wei Xiong**, University of International Business and Economics, China
A robust measure of independence is proposed between two random variables, named integrated mean-variance correlation (IMVC). It has several appealing properties, (a) it lies between zero and one, is zero if and only if the variables are independent and is one if and only if one variable is a measurable function of the other, (b) it is invariant to monotone transformations and is robust to the presence of outliers, (c) it is able to measure the degree of any functional dependencies, including both global and local dependence, (d) its estimation does not require moment conditions on both variables and has a relative low computational complexity. Several important applications of IMVC are considered. First, a distribution-free IMVC independence test is developed, and its explicit asymptotic null distribution is derived, which facilitates the fast calculation of p-values. Second, the IMVC is utilized as a marginal utility to identify active predictors in a high dimensional setting by introducing an IMVC-based model-free feature screening method. This framework can naturally handle censored data arising in survival analysis. To further control the false discovery rate, an IMVC-based local false discovery rate method is proposed that simultaneously exploits commonalities and heterogeneities among predictors, thus improving upon existing methods. The superior performance of the proposed procedures is demonstrated by exhaustive simulation examples and real data applications.

E0686: **Identifying temporal pathways using biomarkers in the presence of latent non-Gaussian components**
Presenter: **Shanghong Xie**, Southwestern University of Finance and Economics, China
Time series data collected from a network of random variables are useful for identifying temporal pathways among the network nodes. Observed measurements may contain multiple sources of signals and noises, including Gaussian signals of interest and non-Gaussian noises, including artefacts, structured noise, and other unobserved factors (e.g., genetic risk factors, disease susceptibility). Existing methods, including vector autoregression (VAR) and dynamic causal modelling, do not account for unobserved non-Gaussian components. Furthermore, existing methods cannot effectively distinguish contemporaneous relationships from temporal relations. A novel method is proposed to identify latent temporal pathways using time series biomarker data collected from multiple subjects. The model adjusts for the non-Gaussian components and separates the temporal network from the contemporaneous network. The algorithm is fast and can easily scale up. The identifiability and the asymptotic properties of the temporal and contemporaneous networks are derived. Superior performance of the method is demonstrated by extensive simulations and an application to a study of attention-deficit/hyperactivity disorder (ADHD).

E0756: **A new regularization method for high-dimensional portfolio selection**
Presenter: **Songshan Yang**, Renmin University of China, China
With the global financial market experiencing continuous expansion and escalating volatility, the development of efficient strategies for high-dimensional portfolio selection has become critically important. Previous approaches to high-dimensional portfolio selection have mainly focused on large-cap companies, presenting challenges when confronted with datasets such as the Russell 2000 index. The aim is to address portfolio optimization challenges within this context, using the 2020-2021 U.S. stock market as a case study. A Dantzig-type portfolio optimization (DPO) model is proposed, and an efficient parallel computing algorithm is presented based on asset-splitting. Through empirical analysis of the S&P 500 and Russell 2000 indices, the consistent outperformance of the DPO portfolios is demonstrated over corresponding ETFs in terms of Sharpe and Sortino ratios, especially for the Russell 2000 index. A new, effective approach is provided for investors seeking to optimize their portfolios in complex market environments.

EO085 Room 104 RECENT ADVANCES IN TIME SERIES AND PANEL DATA ECONOMETRICS Chair: Tingting Cheng

E0374: **Time-varying vector error-correction models: Estimation and inference**
Presenter: **Yayi Yan**, Shanghai University of Finance and Economics, China
Co-authors: Jiti Gao , Bin Peng
A time-varying vector error-correction model is considered to allow for different time series behaviours (e.g., unit-root and locally stationary processes) to interact with each other and co-exist. From a practical perspective, this framework can be used to estimate shifts in the predictability of non-stationary variables, test whether economic theories hold periodically, etc. A time-varying Granger representation theorem is first developed, which facilitates the establishment of asymptotic properties for the model, and then estimation and inferential methods and theory are proposed for both short-run and long-run coefficients. An information criterion is also proposed to estimate the lag length, a singular-value ratio test to determine the cointegration rank, and a hypothesis test to examine the parameter stability. To validate the theoretical findings, extensive simulations are conducted. Finally, the empirical relevance is demonstrated by applying the framework to investigate the rational expectations hypothesis of the U.S. term structure.

E0373: **Time-varying generalized network autoregressive models**
Presenter: **Boyao Wu**, University of International Business and Economics, China
A novel class of time-varying network autoregression models is considered to extend popular network autoregressive models by allowing for general network structures, time-varying model coefficients, and the cross-sectionally dependent error term. A local linear method is proposed to estimate time-varying coefficients, and a recursive-design bootstrap procedure is developed to construct valid confidence intervals for time-varying coefficients in the presence of the cross-sectional dependent error term. Asymptotic theories are established on the estimate and the bootstrap procedure under mild conditions. The proposed estimation and bootstrap procedure are illustrated using simulated and real data. The main contribution is to linear models with the network effect, and light is shed on bootstrap inferences and locally stationary processes.

E0377: **A localized neural network with dependent data: Estimation and inference**
Presenter: **Fei Liu**, Nankai University, China
Co-authors: Jiti Gao , Bin Peng, Yanrong Yang
A Localized Neural Network (LNN) model is proposed. LNN-based estimation and inference are developed under treatment effects. The use of identification restrictions is explored and an estimation theory is established for the LNN setting under mild conditions. A variable selection technique is adopted to further reduce the number of effective parameters. The asymptotic distributions are derived, and the bootstrap procedures are provided accordingly to construct valid inferences. Finally, the theoretical findings are examined through extensive numerical studies.

E0371: **GMM estimation for high-dimensional panel data models**
Presenter: **Tingting Cheng**, Nankai University, China
A class of high dimensional moment restriction panel data models is studied with interactive effects, where factors are unobserved, and factor loadings are nonparametrically unknown smooth functions of individual characteristics variables. The dimension of the parameter vector and the number of moment conditions are allowed to diverge with sample size. This is a very general framework and includes many existing linear and nonlinear panel data models as special cases. In order to estimate the unknown parameters, factors, and factor loadings, a sieve-based generalized method, as well as the moments estimation method, is proposed. It is shown that under a set of simple identification conditions, all those unknown quantities can be consistently estimated. Further, asymptotic distributions of the proposed estimators are established. In addition, tests for over-identification are proposed, factor loading functions are specified, and large sample properties are established. Moreover, a number of simulation studies are conducted to examine the performance of the proposed estimators and test statistics in finite samples. An empirical example of stock return prediction is studied to demonstrate the usefulness of the proposed framework and corresponding estimation methods and testing procedures.

EO211 Room 105 RECENT DEVELOPMENTS IN POINT PROCESSES Chair: Shizhe Chen

E0580: **Bias-correction and test for mark-point dependence with replicated marked point processes**
Presenter: **Yongtao Guan**, Chinese University of Hong Kong, Shenzhen, China
Mark-point dependence plays a critical role in research problems that can be fitted into the general framework of marked-point processes. The focus is on adjusting for mark-point dependence when estimating the mean and covariance functions of the marking process, given independent replicates of the marked-point process. It is assumed that the mark process is a Gaussian process and the point process is a log-Gaussian Cox process, where the mark-point dependence is generated through the dependence between two latent Gaussian processes. Under this framework, naive local linear estimators ignoring the mark-point dependence can be severely biased. It is shown that this bias can be corrected using a local linear estimator of the cross-covariance function, and uniform convergence rates are established for the bias-corrected estimators. Furthermore, a test statistic based on local linear estimators for mark-point independence is proposed, and it is shown to converge to an asymptotic normal distribution in a parametric root n convergence rate. Model diagnostics tools are developed for key model assumptions, and a robust functional permutation test is proposed for a more general class of mark-point processes. The effectiveness of the proposed methods is demonstrated using extensive simulations and applications to two real data examples.

E0691: **Is score matching suitable for estimating point processes**
Presenter: **Feng Zhou**, Renmin University of China, China
Score-matching estimators for point processes have gained widespread attention in recent years because they do not require the calculation of intensity integrals, thereby effectively addressing the computational challenges in maximum likelihood estimation (MLE). Some existing works have proposed score-matching estimators for point processes. However, it is demonstrated that the incompleteness of the estimators proposed in those works renders them applicable only to specific problems, and they fail for more general point processes. To address this issue, the weighted score matching estimator is introduced to point processes. Theoretically, the consistency of the estimator proposed is proven. Experimental results indicate that the estimator accurately estimates model parameters on synthetic data and yields results consistent with MLE on real data. In contrast, existing score-matching estimators fail to perform effectively.

E0794: **Extend the ETAS model step-by-step**
Presenter: **Jiancang Zhuang**, Institute of Statistical Mathematics, Japan
The space-time epidemic-type aftershock sequence (ETAS) model, which is a special type of marked Hawkes process, has been widely used as a standard model to analyze local, regional, or even global seismicity. The assumptions of this model include: (1) The magnitudes are identically independently distributed and also independent from other components. (2) The background seismicity is stationary in time but nonhomogeneous in space. (3) Each event, no matter whether it is a background event or is triggered by other shocks, triggers its own offspring independently according to some probability rules. Many features of seismicity that are not included in the ETAS model have been revealed by analysis of seismic data from different regions, indicating that the ETAS model should be extended to incorporate more information on seismicity. The current developments on the extension of the ETAS model are summarized, including (1) Non-stationary background rates, such as long-term trend and seasonality, (2) Earthquake depth, (3) Geometry of earthquake rupture, (4) Earthquake focal mechanisms, (5) Magnitude dependence in triggering, and (6) Depth dependent clustering. In the implementations of the extensions, stochastic declustering and stochastic reconstruction have been used as basic tools to estimate the non-parametric parts. The current achievements are outlined, and a brief perspective of possible future developments is given.

E0814: **Simultaneous estimation and clustering of additive shape invariant models for neural data**
Presenter: **Shizhe Chen**, University of California, Davis, United States
Co-authors: Shizhe Chen, Zitong Zhang
Technological advancements have enabled the recording of spiking activities from large neuron ensembles, presenting an exciting yet challenging opportunity for statistical analysis. The focus is on the challenges from a common type of neuroscience experiments, where randomized interventions are applied over the course of each trial. The objective is to identify groups of neurons with unique stimulation responses and estimate these responses. The observed data, however, comprise superpositions of neural responses to all stimuli, which is further complicated by varying firing latencies across neurons. A novel additive shape invariant model is introduced, capable of simultaneously accommodating multiple clusters, additive components, and unknown time shifts. Conditions for the identifiability of model parameters are established, offering guidance for the design of future experiments. The properties of the proposed algorithm are examined through simulation studies, and the proposed method is applied to neural data collected in mice.

EO251 Room 108 RECENT ADVANCES IN STATISTICAL LEARNING Chair: Guo Yu

E0335: **Bi-level offline reinforcement learning with limited exploration**
Presenter: **Wenzhuo Zhou**, University of California Irvine, United States
Offline reinforcement learning (RL), which seeks to learn a good policy based on a fixed, pre-collected dataset, is studied. A fundamental challenge behind this task is the distributional shift due to the dataset lacking sufficient exploration, especially under function approximation. To tackle this issue, a bi-level structured policy optimization algorithm is proposed that models a hierarchical interaction between the policy (upper level) and the value function (lower level). The lower level focuses on constructing a confidence set of value estimates that maintain sufficiently small weighted average Bellman errors while controlling uncertainty arising from distribution mismatch. Subsequently, at the upper level, the policy aims to maximize a conservative value estimate from the confidence set formed at the lower level. This novel formulation preserves the maximum flexibility of the implicitly induced exploratory data distribution, enabling the power of model extrapolation. In practice, it can be solved through a computationally efficient, penalized adversarial estimation procedure. The theoretical regret guarantees do not rely on any data-coverage and completeness-type assumptions, only requiring realizability. These guarantees also demonstrate that the learned policy represents the best effort among all policies, as no other policies can outperform it.

E0264: **Completely pivotal estimation in multivariate response linear regression models**
Presenter: **Guo Yu**, University of California Santa Barbara, United States
Despite the vast literature on sparse multivariate response linear regression models, most current methods require a known or explicit estimate of the dependence structure among the random errors. As a result, these methods hinge on computationally expensive methods (e.g., cross-validation) to determine the proper level of regularization. A completely pivotal framework is proposed for the sparse multivariate response linear regression model. The method estimates the coefficient matrix using a model-agnostic regularization parameter that does not depend on either the covariance matrix or the tail conditions of the random errors. In this sense, the proposal is completely tuning-free. Computationally, the estimator is a solution to a convex second-order cone program, which can be solved efficiently. Theoretically, the proposed estimator achieves favorable estimation error rates under mild conditions and could use a second-stage enhancement with non-convex penalties. Through comprehensive numerical studies, the method demonstrates promising statistical performance. Remarkably, the method exhibits strong robustness to the violation of the Gaussian assumption and significantly outperforms competing methods in the heavy-tailed settings.

E0973: **Doubly robust interval estimation for optimal policy evaluation in online learning**
Presenter: **Hengrui Cai**, University of California Irvine, United States
Evaluating the performance of an ongoing policy plays a vital role in many areas, such as medicine and economics, to provide crucial instruction on the early stop of the online experiment and timely feedback from the environment. Policy evaluation in online learning thus attracts increasing attention by inferring the mean outcome of the optimal policy (i.e., the value) in real time. Yet, such a problem is particularly challenging due to the dependent data generated in the online environment, the unknown optimal policy, and the complex exploration and exploitation trade-off in the adaptive experiment. The aim is to overcome these difficulties in policy evaluation for online learning. The probability of exploration that quantifies the probability of exploring the non-optimal actions under commonly used bandit algorithms is explicitly derived. This probability is used to conduct valid inference on the online conditional mean estimator under each action and develop the doubly robust interval estimation (DREAM) method to infer the value under the estimated optimal policy in online learning. The proposed value estimator provides double protection on the consistency and is asymptotically normal with a Wald-type confidence interval provided. Extensive simulations and real data applications are conducted to demonstrate the empirical validity of the proposed DREAM method.

E0986: **Variance estimation for multivariate high-dimensional random effects models under heteroskedasticity**
Presenter: **Xiaodong Li**, UC Davis, United States
Co-authors: Xiaohan Hu, Zhentao Li
Variance estimation in high-dimensional random effects models has recently been widely used in genomics for model-based heritability estimation, and extensions to multivariate traits have also attracted much attention in various phenotypically rich studies. The purpose is to introduce the recent work on making inferences about certain variance parameters, e.g. signal-to-noise ratios, in high-dimensional random effects models with multivariate responses under heteroskedasticity. Two methods are considered: a method of moments and a likelihood-based aggregated estimating equation method. For each method, the consistency and asymptotic distribution of the estimator is established. In particular, the results characterize how the standard errors of the estimators depend on the multivariate noise heteroskedasticity.

EO243 Room 109 RECENT ADVANCES IN DEEP LEARNING: THEORY, ALGORITHMS AND APPLICATIONS Chair: Puyu Wang

E0476: **Nonlinear functional regression by functional deep neural network with kernel embedding**
Presenter: **Zhongjie Shi**, The University of Hong Kong, Hong Kong
With the rapid development of deep learning in various fields of science and technology, such as speech recognition, image classification, and natural language processing, recently, it has also been widely applied in functional data analysis (FDA) with some empirical success. However, due to the infinite-dimensional input, we need a powerful dimension reduction method for functional learning tasks, especially for nonlinear functional regression. Based on the idea of smooth kernel integral transformation, a functional deep neural network is proposed with an efficient and fully-data-dependent dimension reduction method. The architecture of the functional net consists of a kernel embedding step, an integral transformation with a data-dependent smooth kernel; a projection step, a dimension reduction by projection with eigenfunction basis based on the embedding kernel; and finally, an expressive deep ReLU neural network for the prediction. The utilization of smooth kernel embedding enables the functional net to be discretization invariant, efficient, and robust to noisy observations, capable of utilizing information in both input functions and response data and have a low requirement on the number of discrete points for unimpaired generalization performance. Theoretical analysis is conducted, including approximation error and generalization error analysis and numerical simulations to verify these advantages of the functional net.

E0638: **Pairwise learning with deep neural networks**
Presenter: **Junyu Zhou**, University of Sydney, Australia
Pairwise learning refers to learning tasks where the loss function considers a pair of samples simultaneously. Specific examples include ranking, pairwise least squares regression, and metric learning. The generalization analysis of pairwise learning with general losses is delved into by leveraging the specific structure of the target functions. Specifically, the target function of pairwise learning is demonstrated to exhibit (anti)symmetry if the loss function is (anti)symmetric. Building on this observation, structured (anti)symmetric deep neural networks are constructed to approximate the target function. Considering the hypothesis space consisting of these structured deep neural networks, an oracle-type inequality of the empirical minimizer is developed for pairwise learning. These results are applied to concrete examples such as ranking, pairwise least squares regression, and metric learning for illustration.

E0935: **Exploring the benefits of visual prompting in differential privacy**
Presenter: **Xuebin Ren**, Xián Jiaotong University, China

Visual prompting (VP) is an emerging and powerful technique that allows sample-efficient adaptation to downstream tasks by engineering a well-trained frozen source model. The benefits of VP are explored in constructing compelling neural network classifiers with differential privacy (DP). VP is explored and integrated into canonical DP training methods, and its simplicity and efficiency are demonstrated. In particular, it is discovered that VP, in tandem with PATE (i.e., a state-of-the-art DP training method that leverages the knowledge transfer from an ensemble of teachers), achieves the state-of-the-art privacy-utility tradeoff with minimum expenditure of privacy budget. Moreover, additional experiments on cross-domain image classification are conducted with a sufficient domain gap to further unveil the advantage of VP in DP. Lastly, extensive ablation studies are also conducted to validate the effectiveness and contribution of VP under DP consideration.

E0571: **Helmholtz machine with differential privacy**
Presenter: **Junying Hu**, Northwest University, China

Helmholtz machine(HM) is the classic hierarchical probabilistic model for building the probability distribution of perception data, and the wake-sleep(WS) algorithm has been widely used as a training algorithm. To prevent the attacker from restoring the training set data by using the trained HM model, a Gaussian mechanism is introduced to the WS algorithm to propose a wake-sleep algorithm based on differential privacy (DP-WS) and use DPWS to train HM to get the HM model with privacy protection, named DP-HM. Rigorous proof of the privacy guarantee is provided. In addition, the experiments on MNIST and Bio-ID face datasets show that the DP-HM model can be trained under a modest privacy budget and still have acceptable model quality.

EO129 Room 110 RECENT ADVANCES IN DESIGN THEORIES OF EXPERIMENTS Chair: Qi Zhou

E0792: **Capturing column information of a design for models with uncertainty**
Presenter: **Qi Zhou**, Tianjin University of Finance And Economics, China

Model-robust designs have been studied extensively in the literature. These designs are employed to achieve robustness over a set of possible models, which are usually assumed to have an equal probability of being the true model. However, they may not be appropriate for experiments where prior knowledge indicates that effects involving certain factors are more likely to be significant than others. In such cases, it is important to select designs that have superior estimation capacity and information capacity over a subset of the model space that contains models that are more likely to be important. This can be achieved by strategically assigning the factors to the columns of a fractional factorial design. The individual estimation capacity (iEC) and individual information capacity (iIC) are proposed, and these criteria are used to distinguish columns of a factorial design. For a given design, the maximum number of columns, g, is tabulated for which iEC=100%. A new class of designs is proposed that maximizes g, and the trade-offs between the number of runs, the number of columns, and g are evaluated. The emphasis is on the model space that consists of models containing a subset of main effects and their associated two-factor interactions.

E0772: **Column-orthogonal designs with multi-dimensional stratifications**
Presenter: **Xue Yang**, School of Statistics, Tianjin University of Finance and Economics, China

The orthogonal Latin hypercube design and its relaxation and column orthogonal design are two kinds of orthogonal designs for computer experiments. However, they usually do not achieve maximum stratifications in multi-dimensional margins. Some methods are proposed to construct column orthogonal designs with multi-dimensional stratifications by rotating symmetric and asymmetric orthogonal arrays. The newly constructed column orthogonal designs ensure that the estimates of all linear effects are uncorrelated with each other and even uncorrelated with the estimates of all second-order effects (quadratic effects and bilinear effects) when the rotated orthogonal arrays have strength larger than two. Besides orthogonality, the resulting designs also preserve better space-filling properties than those constructed by using the existing methods. In addition, a method to construct a new class of orthogonal Latin hypercube designs is provided with multi-dimensional stratifications by rotating regular factorial designs. Some newly constructed orthogonal Latin hypercube designs are tabulated for practical use.

E0737: **Construction of group doubly coupled designs**
Presenter: **Weiping Zhou**, Guilin University of Electronic Technology, China

For computer experiments with both qualitative and quantitative factors, doubly coupled designs (DCDs) are well suited to investigate the interaction effects between any two qualitative factors and all quantitative factors. The restriction of a DCD is that the number of qualitative factors cannot exceed the number of its factor levels. To break this restriction, group doubly coupled designs (GDCDs) are proposed, which can accommodate more qualitative factors than DCDs. Several methods are introduced to construct GDCDs with low-dimensional space-filling property for quantitative factors.

E0766: **Bayesian optimal designs for generalized linear mixed models based on the penalized quasi-likelihood method**
Presenter: **Yao Shi**, Qingdao University, China

Generalized linear mixed models are widely used in data analysis, while the complexity of the information matrices for such models makes optimal design questions challenging. Moreover, based on some previous research, locally optimal designs for such models can be sensitive to the local parameters. The focus is on the Bayesian optimality criterion to get a robust optimal design. The evaluation of the Bayesian criterion is based on the penalized quasi-likelihood method and on a non-informative prior, to get rid of the influence from the prior choices. Bayesian designs found by a particle swarm optimization algorithm are presented and discussed. The robustness of such a Bayesian design is studied by comparing it with locally optimal designs. Finally, as an illustration, Bayesian optimal designs are derived for a real study.

EO173 Room 111 STATISTICAL MODELING AND COMPUTING METHODS FOR COMPLEX DATA Chair: Victor Hugo Lachos Davila

E0184: **A robust factor analysis model utilizing the canonical fundamental skew-t distribution**
Presenter: **Tsung-I Lin**, National Chung Hsing University, Taiwan
Co-authors: Wan-Lun Wang, I-An Chen

The traditional factor analysis, which relied on the assumption of multivariate normality, has been extended by jointly incorporating the restricted multivariate skew-t (rMST) distribution for the unobserved factors and errors. However, the limited utility of the rMST distribution in capturing skewness concentrated in a single direction prompted the development of a more adaptable and robust factor analysis model. A more flexible, robust factor analysis model is introduced based on the broader canonical fundamental skew-t (CFUST) distribution, called the CFUSTFA model. The proposed new model can account for more complex features of skewness in multiple directions. An efficient alternating expectation conditional maximization algorithm fabricated under several reduced complete-data spaces is developed to estimate parameters under the maximum likelihood (ML) perspective. To assess the variability of parameter estimates, an information-based approach is presented to approximate the asymptotic covariance matrix of the ML estimators. The efficacy and practicality of the proposed techniques are demonstrated through the analysis of simulated and real datasets.

E0185: **Multivariate contaminated normal censored regression model: Properties and maximum likelihood inference**
Presenter: **Wan-Lun Wang**, National Cheng Kung University, Taiwan
The multivariate contaminated normal (MCN) distribution, which contains two extra parameters with respect to parameters of the multivariate normal distribution, one for controlling the proportion of mild outliers and the other for specifying the degree of contamination, has been widely applied in robust statistics in the case of elliptically heavy-tailed empirical distributions. The MCN model is extended to data with possibly censored values due to limits of quantification, referred to as the MCN with censoring (MCN-C) model, and further establishes the censored multivariate linear regression model where the random errors have the MCN distribution, named as the MCN censored regression (MCN-CR) model. Two computationally feasible expectation conditional maximization (ECM) algorithms are developed for the maximum likelihood estimation of MCN-C and MCN-CR models. An information-based method is used to approximate the standard errors of location parameters and regression coefficients. The capability and superiority of the proposed models are illustrated by a real-data example and simulation studies.

E0258: **Exploring metric performance for binary classification in unbalanced data: A comparative study**
Presenter: **Jorge Luis Bazan**, University of Sao Paulo, Brazil
Co-authors: Alex de la Cruz Huayanay
Addressing the recurring challenge in statistical modeling, binary classification, becomes even more complex when encountering imbalanced data in response categories. Extensive literature presents various models and algorithms for binary classification, each leveraging explanatory variables differently. To assess their effectiveness, numerous computing methods have been proposed, utilizing predictive measures to evaluate model performance. The focus is on the performance analysis of prominent metrics found in literature, and its application is extended to the realm of econometrics. Through the investigation, the aim is to provide insights into the efficacy of these metrics and their applicability in real-world scenarios.

E0260: **On the matrix-variate normal distribution for interval-censored and missing data**
Presenter: **Victor Hugo Lachos Davila**, University of Connecticut, United States
Co-authors: Salvatore Daniele Tomarchio, Salvatore Ingrassia, Antonio Punzo
Matrix-variate distributions are powerful tools for modeling three-way datasets that often arise in longitudinal and multidimensional spatiotemporal studies. However, observations in these datasets can be missing or subject to some detection limits because of the restriction of the experimental apparatus. A novel matrix-variate normal distribution for interval-censored and/or missing data is proposed. An analytical yet efficient EM-type algorithm is developed to conduct maximum likelihood estimation of the parameters, having closed-form expressions that rely on truncated moments. Results obtained from the analysis of both simulated data and real case studies concerning water quality monitoring are reported to demonstrate the effectiveness of the proposed method.

E0218: **Censored autoregressive regression models with Student-t innovations**
Presenter: **Christian Eduardo Galarza Morales**, Escuela Superior Politecnica del Litoral, Ecuador
Co-authors: Fernanda Schumacher, Katherine Andreina Loor Valeriano, Larissa Avila Matos
Data collected over time is common in applications and may contain censored or missing observations, making it difficult to use standard statistical procedures. The focus is on proposing an algorithm to estimate the parameters of a censored linear regression model with errors serially correlated and innovations following a Student-t distribution. This distribution is widely used in the statistical modelling of data containing outliers since its longer-than-normal tails provide a robust approach to handling such data. The maximum likelihood estimates of the proposed model are obtained through a stochastic approximation of the EM algorithm. The methods are applied to an environmental dataset regarding ammonia-nitrogen concentration, which is subjected to a limit of detection (left censoring) and contains missing observations. Additionally, two simulation studies are conducted to examine the asymptotic properties of the estimates and the robustness of the model.

EO258 Room 212 NEW DEVELOPMENTS OF CAUSAL INFERENCES AND ITS APPLICATIONS Chair: Jinzhu Jia

E0710: **Nonlinear Mendelian randomization**
Presenter: **Jinzhu Jia**, Peking University, China
Using the Mendelian randomization (MR) approach to explore the causal relationship between exposure and outcome can effectively avoid confounding bias. However, most of the current MR methods are only suitable for cases where the effect of exposure on the outcome is linear. Two approaches to nonlinear MR are proposed, either by fitting a linear model of exposure-instrument and quadratic model of outcome-instrument (QC method) or by fitting a linear model of exposure-instrument and quadratic model of outcome-predicted exposure (two-stage method), the quadratic causality of exposure could be identified and estimated on the outcome. A series of simulations showed that the QC method and two-stage method had high power and low type I error. In real data applications, the QC method and two-stage method found that body mass index had a J-shaped effect on basal metabolic rate and had an inverted J-shaped effect on the level of high-density lipoprotein cholesterol in UK Biobank participants.

E0758: **Interaction tests with covariate-adaptive randomization**
Presenter: **Wei Ma**, Renmin University of China, China
Co-authors: Likun Zhang
Treatment-covariate interaction tests are commonly applied by researchers to examine whether the treatment effect varies across patient subgroups defined by baseline characteristics. The objective is to explore treatment-covariate interaction tests involving covariate-adaptive randomization. Without assuming a parametric data-generating model, usual interaction tests are investigated and it is observed that they tend to be conservative. Specifically, their limiting rejection probabilities under the null hypothesis do not exceed the nominal level and are typically strictly lower than it. Modifications are proposed to the usual tests to obtain corresponding valid tests to address this problem. Moreover, a novel class of stratified-adjusted interaction tests are introduced that are simple, more powerful than the usual and modified tests, and broadly applicable to most covariate-adaptive randomization methods. The results encompass two types of interaction tests: one involving stratification covariates and the other involving additional covariates that are not used for randomization. The application of interaction tests in clinical trials is clarified, and valuable tools are offered to reveal treatment heterogeneity, which is crucial for advancing personalized medicine.

E0916: **Root-n consistent semiparametric learning with high-dimensional nuisance functions under minimal sparsity**
Presenter: **Yuhao Wang**, Tsinghua University, China
Co-authors: Lin Liu
Treatment effect estimation under unconfoundedness is a fundamental task in causal inference. In response to the challenge of analyzing high-dimensional datasets collected in substantive fields such as epidemiology, genetics, economics, and social sciences, many methods for treatment effect estimation with high-dimensional nuisance parameters (the outcome regression and the propensity score) have been developed in recent years. However, it is still unclear what the necessary and sufficient sparsity condition on the nuisance parameters for the treatment effect to be $\sqrt{n}$-estimable. A new Double-Calibration strategy that corrects the estimation bias of the nuisance parameter estimates computed by regularized high-dimensional techniques is proposed. The corresponding Doubly-Calibrated estimator is demonstrated to achieve $1/\sqrt{n}$ rate as long as one of the nuisance parameters is sparse with sparsity below $\sqrt{n}/\log p$, where p denotes the ambient dimension of the covariates. In contrast, the

other nuisance parameter can be arbitrarily complex and completely misspecified. The Double-Calibration strategy can also be applied to settings other than treatment effect estimation, e.g. regression coefficient estimation in the presence of a diverging number of controls in a semiparametric partially linear model.

E0990: **Robust Mendelian randomization coupled with Alphafold2 for drug target discovery**
Presenter: **Zhonghua Liu**, Columbia University, United States
Mendelian randomization (MR) uses genetic variants as instrumental variables (IVs) to infer the causal effect of a modifiable exposure on the outcome of interest by removing unmeasured confounding bias. However, some genetic variants might be invalid IVs due to violations of core IV assumptions. MR analysis with invalid IVs might lead to biased causal effect estimates and misleading scientific conclusions. To address this challenge, a novel MR method is proposed to select valid genetic IVs and then perform post-selection inference (MR-SPI) based on two-sample genome-wide summary statistics. Nine hundred twelve plasma proteins were analyzed using the large-scale UK Biobank proteomics data in 54,306 participants, and seven proteins were identified as significantly associated with the risk of Alzheimer's disease. AlphaFold2 is employed to predict the 3D structural alterations of these seven proteins due to missense genetic variations, providing new insights into their biological functions in disease etiology.

EO067 Room 202 STATISTICAL METHODOLOGIES FOR NEUROIMAGING DATA Chair: Yi Zhao

E0864: **Functional support vector machine with applications to brain imaging data**
Presenter: **Todd Ogden**, Columbia University, United States
Co-authors: Shanghong Xie
Linear and generalized linear scalar-on-function modelling have been commonly used to understand the relationship between a scalar response variable (e.g., continuous or binary outcome) and functional or image-valued predictors. Such techniques are sensitive to model misspecification when the relationship between the response variable and the functional predictors is complex. On the other hand, support vector machines (SVMs) are among the most robust prediction models but do not take into account the high correlations between repeated measurements and cannot be used for irregular data. A novel method is proposed to integrate functional principal component analysis (FPCA) with SVM techniques for classification and regression to account for the continuous nature of functional data and the nonlinear relationship between the scalar response variable and the functional predictors. The performance of the method is demonstrated through extensive simulation experiments and through the problem of classification of alcoholics using electroencephalography (EEG) signals.

E0641: **Longitudinal prediction of brain markers using Gaussian process deep kernel learning**
Presenter: **Haochang Shou**, University of Pennsylvania, United States
Longitudinal prediction of brain markers is of high importance for early diagnosis and prognosis for better understanding and differentiating complex human diseases such as ageing and Alzheimer's disease (AD). The aim is to propose to predict the longitudinal changes of the structural MRI markers by leveraging multimodal features and spatial dependency across brain regions. Using the expressivity of deep kernel learning with Gaussian processes (GP), a personalized and reliable prediction model is presented for noisy, longitudinal data, which can provide individually tailored predictions of longitudinal biomarkers. The model includes the population component that captures the global trend across a set of diverse subjects and the personalized component that adapts the predictions using the history of each subject. The proposed model utilizes a deep neural network to learn complex global trends from a large number of patients and Gaussian processes (GP) to probabilistically quantify the uncertainty of the predictions and model the individual trends of each subject. The model is evaluated on multiple diverse and heterogeneous longitudinal imaging cohorts, and volumetric data for individual regions of interest (ROIs) is predicted, as well as machine learning-based composite scores for brain atrophy. The predicted trajectories are further demonstrated to differentiate between subjects with different disease progression statuses.

E0466: **Sparse partial generalized tensor regression**
Presenter: **Dayu Sun**, Indiana University, United States
Tensor data, often characterized as multi-dimensional arrays, have become increasingly prevalent in biomedical studies, particularly in neuroimaging applications. Analyzing these complex datasets can be challenging due to the high dimensionality and inherent structures within tensors. The sparse partial generalized tensor regression (SPGTR) method is proposed for modeling general types of outcomes involving both tensor and vector/scalar predictors. The novel mode-wise penalized manifold optimization techniques enable the achievement of dimension reduction and sparsity in tensor coefficient estimation, improving the overall prediction performance. The asymptotic behavior of the proposed estimation is established. It demonstrates the effectiveness of the SPGTR through extensive simulation studies, and its application is showcased in investigating the association between posttraumatic stress disorder (PTSD) and brain connectivity matrices derived from functional magnetic resonance imaging (fMRI) data.

E1029: **Beyond massive univariate tests: Covariance regression reveals complex patterns of brain functional connectivity**
Presenter: **Yi Zhao**, Indiana University, United States
Studies of brain functional connectivity typically involve massive univariate tests, performing statistical analysis on each individual connection. The problem of regressing covariance matrices is considered on associated covariates. The goal is to use covariates to explain variation in covariance matrices across units. As such, covariate-assisted principal (CAP) regression is introduced, an optimization-based method for identifying components associated with the covariates using a generalized linear model approach. For high-dimensional data, a well-conditioned linear shrinkage estimator of the covariance matrix is introduced. With multiple covariance matrices, the shrinkage coefficients are proposed to be common across matrices. Theoretical studies demonstrate that the proposed covariance matrix estimator is optimal in achieving the uniform minimum quadratic loss asymptotically among all linear combinations of the identity matrix and the sample covariance matrix. Under regularity conditions, the proposed estimator of the model parameters is consistent. Computationally efficient algorithms are developed to jointly search for common linear projections of the covariance matrices, as well as the regression coefficients. The superior performance of the proposed approach over existing methods is illustrated through simulation studies and a fMRI dataset.

EO234 Room 204 BAYESIAN ANALYSIS WITH DIFFERENT REAL-WORLD APPLICATIONS Chair: Chong Zhong

E0315: **Bayesian group lasso for spatial autoregressive model with convex combinations of different spatial weights**
Presenter: **Jianchao Zhuo**, Xiamen University, China
Co-authors: Zhengzheng Cai, Xiaoyi Han

The spatial autoregressive (SAR) models with convex combinations of different spatial weights have been employed to capture spatial spillovers from different channels and identify the relative importance of each channel. The scenario is considered where different spillover channels exhibit group structure, with multiple spillover channels (spatial weights) being classified into several different groups. A new Bayesian group lasso prior is proposed to detect groups with and without spillover effects and to tackle the multicollinearity issue among multiple spatial weights within the same group. Simulation results suggest that the newly proposed prior performs well in parameter estimation. Lastly, the model is applied with the proposed prior to investigating the sovereign risk spillover effects among developed and developing countries and finding the co-existence of multiple transmission channels in the presence of contemporaneous risk spillover. Among possible spillover channels, the socioeconomic proximity index within informational channels plays the most crucial role, which confirms the empirical findings of a prior study. Additionally, marginal effect analysis suggests that the indirect effect of state spending is almost as large as the direct effect in view of future diffusion impacts.

E0670: **Bayesian integrative region segmentation in spatially resolved transcriptomic studies**
Presenter: **Yinqiao Yan**, Renmin University of China, China

The spatially resolved transcriptomic study is a recently developed biological experiment that can measure gene expressions and retain spatial information simultaneously, opening a new avenue to characterize fine-grained tissue structures. A nonparametric Bayesian method, named BINRES, is proposed to carry out the region segmentation for a tissue section by integrating all three types of data generated during the study: gene expressions, spatial coordinates, and the histology image. BINRES is able to capture more subtle regions than existing statistical partitioning models that only partially make use of the three data modes and is more interpretable than neural-network-based region segmentation approaches. Specifically, due to a nonparametric spatial prior, BINRES does not require a prespecified region number and can learn it automatically. BINRES also combines the image and the gene expressions in the Bayesian consensus clustering framework and thus flexibly adjusts their label alignment contribution weights in a data-adaptive manner. A computationally scalable extension is developed for large-scale studies. Both simulation studies and the real application to three mouse spatial transcriptomic datasets demonstrate that BINRES outperforms the competing methods and easily achieves the uncertainty quantification of the integrative partition.

E0729: **Posterior sampling from truncated Ferguson-Klass representation of normalized completely random measure mixtures**
Presenter: **Junyi Zhang**, The Hong Kong Polytechnic University, Hong Kong
Co-authors: Angelos Dassios

The finite approximation of the completely random measure (CRM) is introduced by truncating its Ferguson-Klass representation. The approximation is obtained by keeping the N largest atom weights of the CRM unchanged and combining the smaller atom weights into a single term. The simulation algorithms are developed for the approximation and characterize its posterior distribution, for which a blocked Gibbs sampler is devised. The approximation is used in two models. The first assumes such an approximation as the mixing distribution of a Bayesian nonparametric mixture model and leads to a finite approximation to the model posterior. The second concerns the finite approximation to the Caron-Fox model. Examples and numerical implementations are given based on the gamma, stable and generalized gamma processes.

E0722: **On posterior mixing under unidentified nonparametric models**
Presenter: **Chong Zhong**, The Hong Kong Polytechnic University, Hong Kong
Co-authors: Jin Yang, Junshan Shen, Catherine Liu, Zhaohai Li

Poor mixing is a stumbling block to Bayesian prediction under nonparametric models with multiple unidentified infinite dimensional parameters, incurring poor estimation of the posterior predictive distribution. Motivated by the prediction under unidentified transformation models, we address poor mixing through a criterion that quantifies the informativeness of nonparametric priors to help the sampler correctly explore the posterior. We first recast the transformation model to its equivalence with a compressed parameter space and propose a set of nonparametric priors. Under general conditions, we formulate the relationship between the asymptotic posterior variation and the nonparametric priors, allowing people to adjust the informativeness of nonparametric priors to fulfill the criterion to obtain mixed posterior. We also find an interesting result that the posterior is still proper even if the prior for the finite-dimensional parameter is improper. Comprehensive simulations and real-world data analysis illustrate our method in addressing poor mixing and demonstrate the superiority in predicting survival outcomes compared with contemporary methods.

EO215 Room 207 RECENT ADVANCES IN PRECISION MEDICINE Chair: Xinzhou Guo

E0516: **ImgKnock: Knockoff inference with optic disc images for glaucoma diagnosis and risk factors**
Presenter: **Zhe Fei**, UC Riverside, United States

ImgKnock is introduced, an innovative pipeline that leverages the power of knockoff inference and deep learning for the analysis of optic disc images in glaucoma. Knockoff inference, known for its ability to control false discovery rates in high-dimensional data, is adapted here to handle complex image data. ImgKnock uniquely generates knockoff images of fundus photographs, enabling the prediction of glaucoma-related outcomes and the identification and testing of important image regions for accurate diagnosis. The method extends knockoff generation to non-tabular data, maintaining key features like the swapping property and ensuring robust feature selection. Central to ImgKnock is a novel testing procedure based on deep neural networks (DNNs), which allows for controlled false discovery rates at the pixel level. This approach facilitates precise inference of feature importance, which is crucial for understanding the disease. ImgKnock has been successfully applied to various image datasets, including MNIST and CIFAR-10, demonstrating its versatility. Significantly, the analysis of optic disc images from glaucoma patients at UCLA Stein Eye Institute has led to novel clinical insights, pinpointing specific fundus regions that are predictive of glaucoma. This advancement in glaucoma research highlights ImgKnock's potential to contribute to improved diagnostic methods and a deeper understanding of the disease.

E0544: **Air pollution and mortality at the intersection of race and social class**
Presenter: **Xiao Wu**, Columbia University, United States

Black Americans are exposed to higher annual levels of air pollution than White Americans and may be more susceptible to its health effects. Low-income Americans may also be more susceptible to air pollution than high-income Americans. Relying on stratum-specific mortality data, with strata defined jointly according to individual-level factors (age, sex, race, and Medicaid eligibility), a causal inference method is developed for continuous exposures to estimate exposure-response curves using stratum-specific mortality data adjusted for observed confounders. The method incorporates an outcome model and an estimated inverse probability of exposure weight to adjust for confounding, which produces a doubly robust estimator. In an extensive analysis of 623 million person-years of Medicare data covering 73 million individuals aged 65 and older from 2000 through 2016, the relationship between annual PM2.5 exposure and mortality across subpopulations defined by race (Black vs. White) and income (Medicaid eligible vs. ineligible) is examined. It is found that higher-income Black persons, low-income White persons, and low-income Black persons may benefit more from lower PM2.5 levels than higher-income White persons. These findings underscore the importance of considering racial identity and income together when assessing health inequities.

E0549: **A general framework for incorporating identification uncertainty in individualized treatment rules**
Presenter: **Muxuan Liang**, University of Florida, United States
Co-authors: Yingqi Zhao, Ting Ye
Estimating individualized treatment rules (ITRs) from observational data or clinical trials with non-adherence is challenging due to possible unmeasured confounding bias. Partial identification approaches using an instrumental variable (IV) provide characterizations on possible values of the conditional average treatment effects (CATEs). A new class of 'optimal' ITRs is developed to guide treatment decisions when the CATEs are only partially identified. A novel value function is defined, allowing a reject option in treatment decisions under partial identification, and that value function is used to define a class of IV-optimal ITRs with a reject option. The reject option informs those susceptible to identification uncertainty and allows the use of alternative ITRs derived from other studies or outcomes for these patients. To estimate the IV-optimal ITRs with a reject option, a weighted classification framework is developed with a modified hinge loss function, where the weights are non-smooth transformations of nuisance parameters. Simulations and real data analysis are conducted to demonstrate the superiority of the developed framework and estimation procedure.

E0877: **Prediction of relapse in pediatric chronic disease using compound medical records**
Presenter: **Jiasheng Shi**, The Chinese University of Hong Kong, Hong Kong
In pediatric chronic disease studies, the recorded data often presents compound data features, e.g., longitudinal data with fragment medical records and time series medical records interlaced with each other. These compound data features provide ample information but require a hybrid statistical analyzing procedure. To address this difficulty, a novel formulation is proposed for medical data with such a compound data structure and an efficient algorithm is proposed to tackle the extrapolation and clustering loop within the overall estimation procedure. An application to pediatric ulcerative colitis chronic disease is presented for the estimation and relapse risk prediction, which is particularly useful for patients' disease management profiles.

EO218 Room 209 STATISTICAL METHODS FOR COMPLEX DATA Chair: Archer Yang

E0866: **Safe feature identification rule for fused Lasso by an extra dual variable**
Presenter: **Pan Shang**, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, China
Co-authors: Huangyue Chen, Lingchen Kong
Fused Lasso was proposed to characterize the sparsity of the coefficients and the sparsity of their successive differences for the linear regression. Due to its wide applications, there are many existing algorithms to solve fused Lasso. However, the computation of this model is time-consuming in high-dimensional data sets. To accelerate the calculation of fused Lasso in high-dimension data sets, the safe feature identification rule is built up by introducing an extra dual variable. With a low computational cost, this rule can eliminate inactive features with zero coefficients and identify adjacent features with the same coefficients in the solution. To the best of our knowledge, existing screening rules cannot be applied to speed up the computation of fused Lasso, and this is the first time to deal with this problem. To emphasize, the rule is a unique result that is capable of identifying adjacent features with the same coefficients; the result is named the safe feature identification rule. Numerical experiments on simulation and real data illustrate the efficiency of the rule, which means this rule can reduce the computational time of fused Lasso. In addition, the rule can be embedded into any efficient algorithm to speed up the computational process of fused Lasso.

E0644: **Extreme marginal quantile treatment effect for high dimensional data**
Presenter: **Jing Zhou**, University of East Anglia, United Kingdom
The estimation and inference of the marginal quantile treatment effects are investigated for high-dimensional data when the quantile level approaches 0 or 1. When the quantile level approaches the ends, quantile regression cannot accurately model the tail distributions. To overcome this limitation, an alternative approach is proposed that uses extreme quantile models to estimate the marginal effect in the presence of a continuous covariate shift. Such models use an extreme value index to model the tail of the distribution function. This method estimates an extreme value index at intermediate quantile levels and extrapolates to the tails where the quantile level is close to zero. By extrapolating, the aim is to estimate the extreme treatment effects consistently and obtain the corresponding asymptotic distribution. Further, to enhance model interpretation, a hypothesis test is proposed to identify the relevant covariates among hundreds of variables for the extreme quantile treatment effects.

E0968: **Inference on derivatives of high dimensional regression function with deep neural network**
Presenter: **Yue Zhao**, University of York, United Kingdom
We study the estimation of the partial derivatives of non-parametric regression functions with many predictors, and a subsequent significance test for the said derivatives. Our derivative estimator is the derivative of the convolution of a regression function estimator and a smoothing kernel, where the regression function estimator is a deep neural network whose structure could scale up as the sample size grows. We demonstrate that in the context of modeling with deep neural networks, derivative estimation is quite different from estimating the regression function itself, and the smoothing operation becomes beneficial. Our subsequent significance test, where the null hypothesis is that a partial derivative is zero, is based on the moment generating function of the aforementioned derivative estimator. To render our estimator and test effective when in high dimensions, we assume that the high-dimensional predictors can serve as the proxies for certain latent, lower-dimensional factors; moreover, we estimate and test the partial derivatives in a coordinate-wise manner, similar to a screening procedure, after controlling for the latent factors. We also finely adjust the regression function estimator to achieve the desired asymptotic normality under the null hypothesis. We demonstrate the excellent performance of our test in a simulation study and real world applications. This is a joint work with Prof. Jianqing Fan at Princeton and Prof. Weining Wang at Groningen.

E1034: **Linear discriminant regularized regression**
Presenter: **Xin Bing**, University of Toronto, Canada
Linear discriminant analysis (LDA) is an important classification approach. Its simple linear form makes it easy to interpret, and it is capable of handling multi-class responses. It is closely related to other classical multivariate statistical techniques, such as Fisher's discriminant analysis, canonical correlation analysis and linear regression. Its connection to multivariate response regression is strengthened by characterizing the explicit relationship between the discriminant directions and the regression coefficient matrix. This key characterization leads to a new regression-based multi-class classification procedure that is flexible enough to deploy any existing structured, regularized, and even non-parametric regression methods. Moreover, the new formulation is generically easy to analyze compared to existing regression-based LDA procedures. In particular, complete theoretical guarantees are provided for using the widely used ℓ_1-regularization that has not yet been fully analyzed in the LDA context. The theoretical findings are corroborated by extensive simulation studies and real data analysis.

EO203 Room 210 RECENT ADVANCES IN STATISTICAL INFERENCE AND COMPLEX DATA ANALYSIS Chair: Yaqing Chen

E0293: **A distance and kernel-based framework for global and local two-sample conditional distribution testing**
Presenter: **Xianyang Zhang**, Texas A&M University, United States
Co-authors: Jian Yan, Zhuoxi Li

Testing for the equality of two conditional distributions is critical in numerous modern applications such as transfer learning and program evaluation. However, this fundamental problem has surprisingly received little attention in the literature. The primary objective is to establish a distance and kernel-based framework for two-sample conditional distribution testing that is adaptable to multivariate distributions and allows for heterogeneity in the marginal distributions. Two metrics are proposed, the conditional generalized energy distance and the conditional maximum mean discrepancy, which completely characterize the homogeneity of two conditional distributions. Utilizing these metrics, local and global tests are developed that can identify local and global discrepancies between two conditional distributions. In theory, the convergence rates, as well as the asymptotic distributions of the local and global tests, are derived under both the null and alternative hypotheses. To approximate the finite-sample distributions of the test statistics, a novel local bootstrap procedure is employed. The proposed local and global two-sample conditional distribution tests demonstrate reliable performance through simulations and real data analysis.

E0331: **Statistical inference for non-Euclidean valued time series**
Presenter: **Xiaofeng Shao**, University of Illinois at Urbana-Champaign, United States
Co-authors: Feiyu Jiang, Changbo Zhu

Data objects taking value in a general metric space have become increasingly common in modern data analysis. The focus is on two important statistical inference problems, namely, two-sample testing and change-point detection, for such non-Euclidean data under temporal dependence. Typical examples of non-Euclidean valued time series include yearly mortality distributions, time-varying networks, and covariance matrix time series. To accommodate unknown temporal dependence, the self-normalization (SN) technique is advanced to the inference of non-Euclidean time series, which is substantially different from the existing SN-based inference for functional time series that reside in Hilbert space. Theoretically, new regularity conditions are proposed that could be easier to check than those in the recent literature and derive the limiting distributions of the proposed test statistics under both null and local alternatives. For the change-point detection problem, the consistency for the change-point location estimator is also derived, and the proposed change-point test is combined with wild binary segmentation to perform multiple change-point estimation. Numerical simulations and real data illustrations demonstrate the effectiveness and robustness of the proposed tests compared with existing methods in the literature.

E0340: **Region-based functional genome-wide association detection for imaging response**
Presenter: **Wenliang Pan**, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, China

Advancements in data acquisition technology have fueled the growth of brain imaging genetic studies, which seek to uncover connections between brain images and genetic markers. Nevertheless, persistent challenges such as misalignment, region heterogeneity, and registration errors necessitate innovative solutions. The region-based functional genome-wide association detection (rfGWAS) method is introduced to address these issues. Focusing on small regions rather than individual voxels, rfGWAS streamlines computation while preserving the detection of meaningful associations. Theoretical analysis confirms that rfGWAS adheres to the independence-zero equivalence principle and reliably identifies significant region sets. Moreover, its test statistic effectively controls Type-I errors under null hypotheses and attains a probability of 1 for rejecting alternative hypotheses. Simulation results underscore the method's effectiveness, and its application to hippocampus surface data from the ADNI study demonstrates its potential. rfGWAS emerges as a promising solution for uncovering region-based associations in brain imaging studies, mitigating critical shortcomings of existing approaches.

E0354: **Dynamic matrix recovery**
Presenter: **Ying Yang**, Chinese Academy of Sciences, China
Co-authors: Ziyuan Chen, Fang Yao

Matrix recovery from sparse observations is an extensively studied topic emerging in various applications, such as recommendation systems and signal processing, which includes matrix completion and compressed sensing models as special cases. A general framework is proposed for dynamic matrix recovery of low-rank matrices that evolve smoothly over time. Starting from the case that the observations are independent across time, it is extended to the setting that both the design matrix and noise possess certain temporal correlations by modified concentration inequalities. By combining neighboring observations, sharp estimation error bounds of both settings are obtained, showing the influence of the underlying smoothness, the dependence and effective samples. A dynamic, fast iterative shrinkage thresholding algorithm that is computationally efficient and characterizes the interplay between algorithmic and statistical convergence is proposed. Simulated and real data examples are provided to support such findings.

EO181 Room 307 FUNCTIONAL DATA ANALYSIS AND ITS APPLICATIONS Chair: Eliana Christou

E0548: **Empirical likelihood inference for functional mean models with application to human cognitive impairment**
Presenter: **Honglang Wang**, Indiana University Indianapolis, United States
Co-authors: Xiang Wang

The two-step-refining inference is considered for the mean function of sparse functional data to account for within-subject correlation. The refined estimator improves the efficiency of the local kernel smoothing estimator, which assumes a working independence correlation structure. The empirical likelihood (EL) based inference is proposed for the mean function of functional data with a bias-correlated estimating equation derived from the two-step-refining procedure. It not only establishes the asymptotic normality of the refined estimator but also derives Wilk's theorem for the empirical likelihood ratio test. The proposed methods perform favorably in finite sample applications from the simulation studies, as well as real data application to the Alzheimer's disease neuroimaging initiative (ADNI) study.

E0904: **Dimension reduction for the conditional quantiles of functional data with categorical predictors**
Presenter: **Shanshan Wang**, The University of North Carolina at Charlotte, United States
Co-authors: Eliana Christou, Eftychia Solea, Jun Song

Functional data analysis has received significant attention due to its frequent occurrence in modern applications, such as in the medical field, where electrocardiograms or electroencephalograms can be used for a better understanding of various medical conditions. Due to the infinite-dimensional nature of functional elements, the focus is on dimension-reduction techniques. The focus is shifted to modeling the conditional quantiles of functional data, noting that existing works are limited to quantitative predictors. Consequently, the first approach is introduced to partial dimension reduction for the conditional quantiles under the presence of both functional and categorical predictors. The proposed algorithm is presented, and the convergence rates of the estimators are derived. Moreover, the finite sample performance of the method is demonstrated using simulation examples and a real data set based on functional magnetic resonance imaging.

E0901: **Functional motif discovery in stock market prices**
Presenter: **Marzia Cremona**, Universite Laval, Canada
Co-authors: Lyubov Doroshenko, Federico Severino
Financial asset prices display recurrent patterns over time. However, such time series are usually noisy and volatile, making the identification of repetitive patterns particularly difficult. These motifs are rarely exploited for price prediction, even though some of them, such as the surge of a financial bubble, occur periodically and feature similar shapes. Asset prices are embedded in a functional data analysis framework by extending and using probabilistic K-means with local alignment to discover functional motifs in stock price time series. The information of the discovered motifs is then exploited to perform the price forecasts with a novel motif-based (MB) algorithm introduced. After illustrating the technique on simulations of mixed causal-noncausal autoregressive processes, it is applied to the prices of S&P 500 top components and motif-based forecasting is performed. Finally, its performance is compared to traditional forecasting models.

E0911: **Robust inverse regression for multivariate elliptical functional data**
Presenter: **Eftychia Solea**, Queen Mary University of London, United Kingdom
Co-authors: Eliana Christou, Jun Song
Functional data have received significant attention as they frequently appear in modern applications, such as functional magnetic resonance imaging (fMRI) and natural language processing. The infinite-dimensional nature of functional data makes it necessary to use dimension-reduction techniques. Most existing techniques, however, rely on the covariance operator, which can be affected by heavy-tailed data and unusual observations. Thus, a robust sliced inverse regression is considered for multivariate elliptical functional data. Therefore, a new statistical linear operator is introduced, called the conditional spatial sign Kendall's tau covariance operator, which can be seen as an extension of the multivariate Kendall's tau to both the conditional and functional settings. The new operator is robust to heavy-tailed data and outliers and can provide a robust estimate of sufficient predictors. The convergence rates of the proposed estimators are also derived for both completely and partially observed data. Finally, the finite sample performance of the estimator is demonstrated using simulation examples and a real dataset based on fMRI.

EO163 Room 313 ADVANCES IN COMPLEX TIME SERIES Chair: Wai-keung Li

E0359: **Inference for the panel ARMA-GARCH model when both N and T are large**
Presenter: **Bing Su**, The University of Hong Kong, China
Co-authors: Ke Zhu
A panel ARMA-GARCH model is proposed to capture the dynamics of large panel data with N individuals over T time periods. For this model, a two-step estimation procedure is provided to estimate the ARMA parameters and GARCH parameters stepwisely. Under some regular conditions, it is shown that all of the proposed estimators are asymptotically normal with the convergence rate $1/\sqrt{(NT)}$, and they have asymptotic biases when both N and T diverge to infinity at the same rate. Particularly, the asymptotic biases result is found from the fixed effect, estimation effect, and unobservable initial values. To correct the biases, the bias-corrected version is further proposed by estimators by using either the analytical asymptotics or the jackknife method. The asymptotic results are based on a new central limit theorem for the linear-quadratic form in the martingale difference sequence when the weight matrix is uniformly bounded in rows and columns. Simulations and one real example are given to demonstrate the usefulness of the panel ARMA-GARCH model.

E0483: **Multi-view dynamic social network modeling**
Presenter: **Shun Hin Chan**, The Hong Kong University of Science and Technology, Hong Kong
Co-authors: Amanda Chu, Mike So
A flexible multi-view dynamic social network model is developed using a regression-like structure, incorporating exogenous and endogenous variables from the lagged networks to model edge changes. The model does not rely on latent space, simplifying network estimation and prediction. Furthermore, it integrates a multi-view feature to represent various relationship types at each time point. The proposed model offers an intuitive interpretation of the estimation. Bayesian model averaging method is also applied to predict networks.

E0574: **Encoding recurrence into transformers**
Presenter: **Yuxi Cai**, The University of Hong Kong, Hong Kong
Co-authors: Feiqing Huang, Kexin Lu, Zhen Qin, Yanwen Fang, Guangjian Tian, Guodong Li
The purpose is to break down with ignorable loss an RNN layer into a sequence of simple RNNs, each of which can be further rewritten into a lightweight positional encoding matrix of a self-attention, named the recurrence encoding matrix (REM). Thus, recurrent dynamics introduced by the RNN layer can be encapsulated into the positional encodings of a multi-head self-attention, and this makes it possible to seamlessly incorporate these recurrent dynamics into a transformer, leading to a new module, self-attention with recurrence (RSA). The proposed module can leverage the recurrent inductive bias of REMs to achieve a better sample efficiency than its corresponding baseline transformer, while self-attention is used to model the remaining non-recurrent signals. The relative proportions of these two components are controlled by a data-driven gated mechanism, and the effectiveness of RSA modules is demonstrated by time series forecasting tasks.

E0725: **Semi-strong double-autoregressive models: Structure and estimation**
Presenter: **Xuqin Wang**, Xiamen University, China
The first-order semi-strong double autoregressive model is investigated, where semi-strong means that the errors are not required to be independent over time. A sufficient condition is firstly obtained for a unique, strictly stationary, and ergodic solution of the model without the need to check irreducibility. Consistency and asymptotic normality of the quasi-maximum likelihood estimated parameters are also studied under some mild conditions. In contrast to the existing literature, the innovation variable is not required to be Gaussian or independent over time. Then, extensions to the higher-order semi-strong double autoregressive model are also discussed. Finally, the finite sample performance of the quasi-maximum likelihood estimator is assessed through Monte Carlo simulations.

EO201 Room 405 STOCHASTIC MODELS IN STATISTICS Chair: Soudeep Deb

E0318: **Within game forecasting in T20 cricket, a time series classification approach using parallelism**
Presenter: **Rishideep Roy**, University of Essex, United Kingdom
A novel approach is introduced to within-game forecasting by treating each match as a dynamic time series. The unique contribution lies in the application of a sophisticated time series clustering method, which leverages the parallelism of trend functions. The aim is also to study the changing influence of covariates within a match towards a match outcome.

E0319: **Robust and efficient parameter estimation for discretely observed stochastic processes**
Presenter: **Abhik Ghosh**, Indian Statistical Institute, India
Co-authors: Rohan Hore
In various practical situations, data is encountered from stochastic processes, which can be efficiently modelled using an appropriate parametric model for subsequent statistical analyses. Unfortunately, the most common estimation and inference methods based on the maximum likelihood (ML) principle are susceptible to minor deviations from assumed model or data contamination due to their well-known lack of robustness. Since the alternative non-parametric procedures often lose significant efficiency, a robust parameter estimation procedure is developed for discretely observed data from a parametric stochastic process model, which exploits the properties of the popular density power divergence measure in the framework of minimum distance inference. In particular, the minimum density power divergence estimators (MDPDE) are defined for the independent increment and the Markov processes. The asymptotic consistency and distributional results are established for the proposed MDPDEs in these dependent stochastic process set-ups, and their benefits are illustrated over the usual ML estimator for common examples like the Poisson process, drifted Brownian motion and auto-regressive models.

E0605: **Limited theorems and phase transitions in tensor Curie-Weiss Ising and Potts models**
Presenter: **Somabha Mukherjee**, National University of Singapore, Singapore
The purpose is to discuss some recent developments in the area of asymptotics of magnetization and the inverse temperature/magnetic field estimates in p-spin Curie-Weiss Ising and Potts models, two well-known models in statistical physics that capture multi-body interactions. The results are characterized by surprising phase transition phenomena and limit theorems of the empirical magnetization on different regions of the parameter space, the existence of a critical curve in the interior of this space on which the estimators have mixture-limiting distributions and a surprising super-efficiency phenomenon at the boundary point(s) of this critical curve.

E0994: **A test for counting sequences of integer-valued autoregressive models**
Presenter: **Yuichi Goto**, Kyushu University, Japan
Co-authors: Kou Fujimori
Non-negative integer-valued time series have been intensively studied for decades. Integer autoregressive (INAR) models are one of the most popular models in this field. To define the INAR model, the binomial thinning operator or, more generally, the generalized Steutel and van Harn operator guarantees the integer-valued nature. However, the distributions of the counting sequences used in the operators have been determined by the preference of analysts without statistical verification so far. The aim is to propose a test for counting sequences of INAR models and show that the test has an asymptotically correct and consistent size.

EO066 Room 406 NEW ADVANCES IN STATISTICAL ESTIMATION, TESTING AND CLASSIFICATION Chair: Ke Yang

E0292: **Sparse positive-definite estimation for covariance matrices with repeated measurements**
Presenter: **Yuedong Wang**, University of California - Santa Barbara, United States
Repeated measurements arise in many areas, such as epidemiology, medicine, psychology, and neuroscience, where random variables are measured multiple times across different subjects. In such settings, dependence structures among random variables that are between subjects and within a subject may be different. Ignoring this fact may lead to misleading and questionable analytic results. The problem of simultaneous sparse and positive-definite estimation is studied for the between-subject and within-subject covariance matrices. The convergence rates are established for the proposed between-subject and within-subject covariance matrix estimators under some regularity conditions. In general, the convergence rate for the within-subject covariance matrix estimator depends on the total number of observations, while the convergence rate for our between-subject covariance matrix estimator is affected by the number of groups and is insensitive to the imbalance of the data. The finite-sample performance of the proposed estimators is illustrated numerically in comprehensive simulations and a real data application.

E0162: **Hierarchical Neyman-Pearson classification for prioritizing severe disease categories in COVID-19 patient data**
Presenter: **Lijia Wang**, City University of Hong Kong, Hong Kong
COVID-19 has a spectrum of disease severity, ranging from asymptomatic to requiring hospitalization. Understanding the mechanisms driving disease severity is crucial for developing effective treatments and reducing mortality rates. One way to gain such understanding is using a multi-class classification framework, in which patients' biological features are used to predict patients' severity classes. In this severity classification problem, it is beneficial to prioritize identifying more severe classes and control the under-classification errors, in which patients are misclassified into less severe categories. The Neyman-Pearson classification paradigm has been developed to prioritize the designated type of error. However, current NP procedures are either for binary classification or do not provide high probability controls on the prioritized errors in multi-class classification. A hierarchical NP framework and an umbrella algorithm that generally adapts to popular classification methods and controls the under-classification errors with high probability are proposed.

E0378: **A cross-validation approach for distribution-free two-sample testing with high-dimensional data**
Presenter: **Shunan Yao**, Hong Kong Baptist University, Hong Kong
Two-sample testing is a cornerstone of modern statistics. In the realm of high-dimensional two-sample testing, traditional methods often rely on distributional assumptions, e.g., Hotelling's T-test, or use a subset of the data for dimension reduction, thereby not utilizing the full dataset for the actual test. A novel cross-validation style approach is introduced to two-sample testing that is computationally tractable and capable of using the full dataset for testing. Specifically, (a) it is shown that the method achieves super-uniformity in finite samples under the null hypothesis, where both samples originate from an identical distribution, (b) establish the asymptotic properties of the test statistic, and (c) the performance is demonstrated of the method through numerical applications.

E0201: **An intrinsic measure for quantifying the heterogeneity in meta-analysis**
Presenter: **Ke Yang**, Beijing University of Technology, China
Co-authors: Enxuan Lin, Wangli Xu, Liping Zhu, Tiejun Tong
Quantifying the heterogeneity is an important issue in meta-analysis, and among the existing measures, the I2 statistic is most commonly used. A motivating example is first presented to demonstrate that the I2 statistic heavily depends on the study sample sizes. It is further shown, by a connection between analysis of variance and meta-analysis, that the I2 statistic was defined to quantify the heterogeneity between the observed

effect sizes. Inspired by this, a new measure is introduced that aims to directly quantify the heterogeneity between the study populations involved in the meta-analysis in a way that avoids the influence of sample sizes through the observed effect sizes. More importantly, a new estimator, namely the IQ statistic, is also proposed to estimate the newly defined intrinsic measure for quantifying the heterogeneity. For practical use, the exact formulas for the IQ statistic are also specified under three different scenarios, including the mean, the mean difference, and the standardized mean difference. Simulations and real data analysis demonstrate that the IQ statistic provides an asymptotically unbiased estimator of the true heterogeneity between the study populations, and it does not depend on the study sample sizes as expected.

EO102 Room 408 AT THE INTERSECTION OF STATISTICAL LEARNING AND MACHINE LEARNING Chair: Yichen Cheng

E0276: **Bayesian-frequentist hybrid inference in clinical and genomic research applications**
Presenter: **Gang Han**, Texas A&M University, United States

The Bayesian-frequentist hybrid model and associated inference can combine the advantages of both Bayesian and frequentist methods and avoid their limitations. However, except for a few special cases in existing literature, the computation under the hybrid model is generally nontrivial or even unsolvable. A computation algorithm for hybrid inference is developed under any general loss functions. Three simulation examples demonstrate that hybrid inference can improve upon frequentist inference by incorporating valuable prior information. Bayesian inference based on non-informative priors is also improved, where the latter leads to biased estimates for the small sample sizes used in inference. The proposed method is illustrated in research applications, including a biomechanical engineering design for knee prostheses, surgical treatment of acral lentiginous melanoma, modelling HIV viral load dynamics, and an analysis of RNA single-cell sequencing data incorporating cell probabilities for identifying protein-coding genes associated with pulmonary fibrosis.

E0597: **Resolving conflicts in crowds: An earnings forecasts application**
Presenter: **Houping Xiao**, Robinson College of Business/GSU, United States
Co-authors: Shiyu Wang

Recently, investors can obtain earnings forecast information through traditional venues, such as Wall Street, Institutional Brokers' Estimate System (IBES), as well as modern social media platforms like Estimize, which generates consensus estimates based on the forecasts from individuals with different backgrounds. As a result, this will inevitably lead to conflicts in the earnings forecast. The aim is to present a novel and effective optimization-based approach to resolve such conflicts in earnings forecast data and further generate an accurate and robust earnings forecast consensus. Consistent with the wisdom-of-crowds effect, the new earnings forecast consensus is more accurate than the Wall Street consensus (67.5% of estimations with error less than the Wall Street) and IBES consensus (67.4% of estimations with error less than the IBES) of the time. Moreover, the new earnings forecast consensus can provide incrementally useful information in forecasting earnings, and the incremental information is further priced in the market after the earnings announcement.

E0695: **Linear shrinkage convexification of penalized linear regression with missing data**
Presenter: **Johan Lim**, Seoul National University, Korea, South

One of the common challenges faced by researchers in recent data analysis is missing values. In the context of penalized linear regression, which has been extensively explored over several decades, missing values introduce bias and yield a non-positive definite covariance matrix of the covariates, rendering the least square loss function non-convex. A novel procedure called the linear shrinkage positive definite (LPD) modification is proposed to address this issue. The LPD modification aims to modify the covariance matrix of the covariates to ensure consistency and positive definiteness. Employing the new covariance estimator, the penalized regression problem can be transformed into a convex one, thereby facilitating the identification of sparse solutions. Notably, the LPD modification is computationally efficient and can be expressed analytically. In the presence of missing values, the selection consistency is established, and the convergence rate of the l1-penalized regression estimator is proven with LPD, showing an optimal l2-error convergence rate. To further evaluate the effectiveness of the approach, real data is analyzed from the genomics of drug sensitivity in cancer (GDSC) dataset. This dataset provides incomplete measurements of drug sensitivities of cell lines and their protein expressions. A series of penalized linear regression models are conducted, with each sensitivity value serving as a response variable and protein expressions as explanatory variables.

E0874: **Inference of single-cell gene regulatory networks via statistical learning methods**
Presenter: **Min Chen**, University of Texas at Dallas, United States

Gene regulatory networks (GRNs) are crucial for understanding the complex relationships between genes and their regulators, which are fundamental to all cellular processes. By deciphering GRNs, scientists can gain insights into the regulatory crosstalk that drives various diseases, potentially leading to new treatments and therapies. Inferring gene regulatory networks using single-cell RNA sequencing (scRNA-seq) is particularly important because scRNA-seq provides a high-resolution view of gene expression in individual cells, revealing the heterogeneity within a population that bulk RNA-seq might miss. Most existing methods are developed for bulk mRNA experiments. The existing methods are reviewed and demonstrated through simulation and real data the power of statistical learning methods in reconstructing gene regulation networks for single-cell data.

EO240 Room 411 (Virtual sessions) RECENT ADVANCES IN MANIFOLD-RELATED STATISTICAL INFERENCE Chair: Rong Tang

E0538: **Random fixed boundary flows: A twin sister of principal flow**
Presenter: **Zhigang Yao**, National University of Singapore, Singapore

The focus is on fixed boundary flows with canonical interpretability as principal components extended on non-linear Riemannian manifolds. The aim is to find a flow with fixed starting and ending points for noisy multivariate data sets lying near an embedded non-linear Riemannian manifold. In geometric terms, the fixed boundary flow is defined as an optimal curve that moves in the data cloud with two fixed endpoints. At any point in the flow, the inner product of the vector field is maximized, which is calculated locally, and the tangent vector of the flow. The rigorous definition is derived from an optimization problem using the intrinsic metric on the manifolds. For random data sets, the fixed boundary flow is named the random fixed boundary flow, and its limiting behavior is analyzed under noisy observed samples. It is shown that the fixed boundary flow yields a concatenate of three segments, one of which coincides with the usual principal flow when the manifold is reduced to the Euclidean space. It is further proven that the random fixed boundary flow converges largely to the population fixed boundary flow with high probability. Finally, it illustrates how the random fixed boundary flow can be used and interpreted and demonstrates its application in real data sets.

E0384: **Mean-field variational inference via Wasserstein gradient flow**
Presenter: **Rentian Yao**, University of Illinois at Urbana Champaign, United States

Variational inference, such as the Mean-Field (MF), requires certain conjugacy structures for efficient computation. These can impose unnecessary restrictions on the viable prior distribution family and further constraints on the variational approximation family. A computational framework is introduced to implement MF variational inference for Bayesian models, with or without latent variables, using the Wasserstein gradient flow (WGF), a modern mathematical technique for performing a gradient flow over the space of probability measures. Theoretically, the algorithmic convergence of the proposed approaches is analyzed, providing an explicit expression for the contraction factor. Existing results on MF variational posterior concentration from a polynomial to an exponential contraction are also strengthened by utilizing the fixed point equation of the time-

discretized WGF. Computationally, a new constraint-free function approximation method is proposed using neural networks to numerically realize the algorithm. This method is shown to be more precise and efficient than traditional particle approximation methods based on Langevin dynamics.

E0272: **Wasserstein convergence of persistence diagrams on generic manifolds**

Presenter: **Vincent Divol**, Universite Paris Dauphine PSL, France

Persistence diagrams (PDs) are routinely used in Topological Data Analysis to describe the topology of a sample in a multiscale fashion. They consist of a multiset of points in the upper half-plane, where each point in the PD intuitively corresponds to a topological feature of the underlying point cloud. When the sample lies on a submanifold of the Euclidean space, the PD of the sample (with respect to the Cech filtration) is known to be separated into two parts. A small number of points in the PD, which lie far away from the diagonal of the upper half-plane, correspond to the PD of the underlying manifold. On the other hand, a large collection of points lying close to the diagonal informally represents "topological noise". A complete asymptotic description of the structure of this topological noise is provided in the case where the sample lies on a generic submanifold. In particular, limit laws are offered for the total persistence of such PDs and prove convergence results with respect to Wasserstein distances. This generalizes previous results proven in another study in the case of points sampled in the cube $[0,1]^m$.

E0275: **Logistic regression models for elastic shape of curves**

Presenter: **Min Ho Cho**, Inha University, Korea, South

Shape analysis is widely used in many areas, such as computer vision and medical and biological studies. One challenge in analyzing the shape of an object in an image is its invariant property to shape-preserving transformations. To measure the distance or dissimilarity between two shapes, the square-root velocity function (SRVF) representation and the elastic metric are used. Since shapes are inherently high-dimensional in a nonlinear space, a tangent space is adopted at the mean shape and a few Principal Components (PCs) on the linearized space. Classification methods are proposed based on logistic regression using these PCs and tangent vectors with the elastic net penalty. Its performance compared with other model-based methods for shape classification is assessed on the shape of algae in watersheds, as well as simulated data generated by the mixture of von Mises-Fisher distributions.

Thursday 18.07.2024 16:10 - 18:15 Parallel Session I – EcoSta2024

EV268 Room 411 (Virtual sessions) METHODOLOGICAL STATISTICS AND ECONOMETRICS Chair: Stefano Rizzelli

E0288: **Forecast combination and interpretability using random subspace**
Presenter: **Boris Kozyrev**, Halle Institute for Economic Research (IWH), Germany
Forecast aggregation is investigated via random subspace regressions (RS), and the potential link between RS and the Shapley value decomposition (SVD) is explored using the US GDP growth rates. This combination of techniques enables handling high-dimensional data and reveals the relative importance of each individual forecast. First, the possibility of enhancing forecasting performance in certain practical instances is demonstrated by randomly selecting smaller subsets of individual forecasts and obtaining a new set of predictions based on a regression-based weighting scheme. The optimal value of selected individual forecasts is also empirically studied. Then, a connection between RS and the SVD is proposed, enabling the examination of each individual forecast's contribution to the final prediction, even when the number of forecasts is relatively large. This approach is model-agnostic (can be applied to any set of forecasts) and facilitates understanding of how the aggregated prediction is obtained based on individual forecasts, which is crucial for decision-makers.

E0931: **Spectral CLTs for large language and large multimodal models**
Presenter: **Andrej Srakar**, Institute for Economic Research Ljubljana, Slovenia
Since the past pioneering studies, central and noncentral limit theorems have been constantly refined and extended. Recently, another study extended this to spectral central limit theorems that are valid for additive functionals of isotropic and stationary Gaussian fields. It uses the Malliavin-Stein method and Fourier analysis techniques when Y_t admits Gaussian fluctuations in a long memory context. In another recent article, existing language models are augmented with long-term memory. It proposed a framework of language models augmented with long-term memory, which enables LLMs to memorize long histories. The focus is to develop spectral central limit theorems in the context of augmented large language models, as well as to present extensions of LLM labelled, large multimodal models. The main stochastic calculus tools are derived from the Malliavin-Stein method, Fourier analysis, and free probability. Applications and extensions of the work are possible in multiple areas in probability theory, statistics, data science and econometrics, such as stochastic geometry, spherical random fields, deep neural networks and graph neural networks, causal AI and functional data analysis. Applications on datasets from finance and medical imaging are presented. In conclusion, possible Bayesian extensions are discussed.

E0978: **Assessing the probability of museum opening choices and its spatial distribution through multilevel logit kriging**
Presenter: **Sabrina Maggio**, University of Salento, Italy
Co-authors: Sandra De Iaco
Museums are extensively distributed all over the Italian territory and have a very important role in education and recreation through research activities and exhibition space management, as well as they are the major repositories of the national cultural heritage, whose protection, valorization and development are endorsed and assured. An innovative methodology that combines multilevel multinomial ordered models with spatial continuity models is proposed. In particular, the spatial components of the data are treated as different units of the multilevel multinomial model in order to assess the complex pattern of variability at each level, as well as the resulting spatially indexed probability, are interpreted as a finite realization of a stationary random field, in order to examine its continuity over the territory. Thus, a multilevel multinomial ordered model is defined for the estimation of the probability of the museums opening decisions, by taking into account the variation of such probability both at regional and provincial levels for some peculiar museums characteristics. Then, the spatial continuity of the propensity, measured by the logit, of remaining open all over the year or at least seasonally, is modelled and a new form of kriging, called multilevel logit kriging, is used for interpolation purposes. The ISTAT microdata concerning the Italian survey on the museums and cultural institutions will be considered.

EO092 Room 102 CONTRIBUTIONS IN THEORETICAL AND APPLIED ECONOMETRICS Chair: Massimiliano Caporin

E0513: **Responsible investing: ESG risk budgeting**
Presenter: **Runfeng Yang**, CaFoscari University of Venice, Italy
A new risk budgeting framework is proposed based on the ESG risk factor, a risk factor which captures market realization of ESG scores, in addition to the traditional setting based on ESG scores. Empirical analysis of the European stock market from 2013 to 2022 reveals limited evidence of ESG impact on the Sharpe ratio. Significant effects are found in specific sector/ESG groups and during extreme market events. Notably, actively pursuing ESG risk exposure may significantly impair risk-return trade-offs, with the impact influenced by the choice of ESG score datasets. These findings point out the complexities of including ESG factors in investment and emphasize the need for careful risk management in handling these issues.

E0402: **Nonparametric estimation and forecasting of time-varying parameter models**
Presenter: **Yu Bai**, Monash University, Australia
Issues of using a local estimator in a forecasting model affected by parameter instability are addressed. The choices of kernel weighting function are analyzed, and the bandwidth parameters are associated with the local estimator. The asymptotic optimality of the bandwidth selection procedure is proven, and an analytical criterion is provided for the choice of the kernel weighting function. The theoretical results are examined through an extensive Monte Carlo study and an empirical application on bond return predictability.

E0403: **Assessing nonlinear impact of ESG on company performance**
Presenter: **Yufeng Mao**, University of Padova, Italy
The prevalent linear models employed to gauge the influence of ESG criteria are assessed on corporate performance within the existing literature. The effects of ESG factors are examined on both accounting-based and market-based profitability proxies. The nonlinear effect is first examined by incorporating a quadratic ESG score term into a fixed effects panel data model. This is followed by an analysis of asymmetric influences, distinguishing between the effects of positive and negative ESG score fluctuations. Lastly, a quantile fixed effects model is implemented to account for potential variability in ESG impact across different quantiles of corporate performance metrics. The findings reveal that ESG scores exhibit increasing marginal effects. Additionally, asymmetrical impacts of positive versus negative ESG score changes are uncovered. Results from quantile regression suggest that ESG scores' influence on a market-based profitability proxy is not uniform across quantiles.

E1007: **Estimating the natural rate of interest and the risk appetite in the US: An accelerating score-driven state space model**
Presenter: **Tibor Pal**, University of Salerno, Italy
Co-authors: Giuseppe Storti
The aim is to develop a novel accelerating score-driven state-space model for estimating the natural rate of interest and risk appetite. The proposed model extends the class of score-driven state-space models proposed in another study by assigning a time-varying weight to the conditional likelihood score. The flexible parameters are driven by autocorrelations and cross-correlations of past score innovations, resulting in an accelerated updating mechanism. The model is used to study and estimate the US natural rate of interest in the Laubach-Williams framework, where the IS

and Phillips curve relationships become time-varying. Furthermore, a financial component is introduced into the model to account for the dynamic effect of risk attitude in the financial intermediation sector. In addition to estimating the natural rate of interest, the proposed extensions to the baseline Laubach-Williams framework allow the time-varying nature of the relationships involved to be analyzed and risk appetite to be estimated, thus providing an additional yardstick for monetary policy.

E0226: **Quantile spillover indexes: Simulation-based evidence, confidence intervals and a decomposition**
Presenter: **Massimiliano Caporin**, University of Padova, Italy
Co-authors: Giovanni Bonaccolto, Jawad Shahzad

Quantile-spillover indexes have recently become popular for analyzing tail interdependence. An extensive simulation study shows that the estimation of spillover indexes is affected by a positive distortion when the parameters of the underlying fitted models are not evaluated with respect to their statistical significance. The distortion is reduced by filtering out non-significant parameters and also for increasing sample sizes, thanks to the consistency of estimators, but is not fully disappearing due to type I error. Another step is introducing a simulation-based approach to recovering confidence intervals from quantile spillover indexes. In addition, an algebraic decomposition of quantile spillover is put forward, separating the dynamic interdependence from the contemporaneous interdependence (due to residual correlation). Empirical evidence shows that distortions in real data are sizable, and the decomposition points out that most of the spillover is due to contemporary effects. All of the results extend and are confirmed for the spillover index of a prior study.

EO058 Room 103 NEW ADVANCES IN PANEL DATA MODELS Chair: Alexandra Soberon

E0239: **Estimating latent-variable panel data models using parameter-expanded SEM methods**
Presenter: **Siqi Wei**, IE University, Spain

New estimation algorithms are presented for three types of dynamic panel data models with latent variables: factor models, discrete choice models, and persistent-transitory quantile processes. The new methods combine the parameter expansion (PX) ideas in a prior study with the stochastic expectation-maximization (SEM) algorithm in likelihood and moment-based contexts. The goal is to facilitate convergence in models with a large space of latent variables by improving algorithmic efficiency. This is achieved by specifying expanded models within the M step. Effectively, new estimators for the pseudo-data are proposed within iterations that take into account the fact that the model of interest is misspecified for draws based on parameter values far from the truth. Conditions are provided under which the new algorithm dominates SEM in terms of the global rate of convergence and characterizes the asymptotic distributions of the estimators based on PX-SEM algorithms. Finally, in simulations, it is shown that the new algorithms significantly improve the convergence speed relative to standard SEM algorithms, sometimes dramatically so.

E0252: **Empirical likelihood based testing device for a semiparametric panel data model with cross-sectional dependence**
Presenter: **Luis Antonio Arteaga Molina**, Universidad de Cantabria, Spain
Co-authors: Juan Manuel Rodriguez-Poo

Empirical-likelihood-based inference for nonparametric panel data models with cross-sectional dependence is investigated. A common factor structure is used to characterize the cross-sectional dependence. The empirical likelihood is employed to formulate a functional form specification test. The procedure is based on a comparison with kernel smoothing estimators. To obtain the estimators, an empirical likelihood ratio is first developed to obtain a maximum empirical likelihood local linear common correlated effects estimator. To show the feasibility of the technique and to analyze its small sample properties, a Monte Carlo simulation exercise is implemented, and the proposed technique is also illustrated in an empirical analysis of the environmental Kuznets curve hypothesis.

E0283: **Estimation and inference of panel data models with a generalized factor structure**
Presenter: **Juan Manuel Rodriguez-Poo**, Universidad de Cantabria, Spain
Co-authors: Alexandra Soberon, Stefan Sperlich

A new class of panel data models are introduced where unobserved factors and factor loadings are introduced in a nonlinear fashion. To estimate the parameters of interest in this class of models, a consistent and asymptotically normal root-NT estimator is proposed. The approach shares the same philosophy as the common correlated effects estimation technique: it removes the unknown factors by transforming the model, but our proposal covers a wider set of models. Since a conditional independence assumption is the vault key of the estimation procedure, a consistent specification test is also proposed to check whether this assumption is fulfilled or not. The test relies on combining the methodology of conditional moments tests and nonparametric estimation techniques. Using degenerate and nondegenerate theories of U-statistics, the test is asymptotically distributed under the null while it diverges under the alternative at a rate that is arbitrarily close to the square root of NT. The finite sample performance is evaluated by the new estimators and test statistics with simulated data examples, and a real empirical problem is provided about the effect of the EU ETS on the economic development of the EU countries.

E0291: **Estimation and testing for varying coefficient multidimensional panel data models: A differencing approach**
Presenter: **Alexandra Soberon**, Universidad de Cantabria, Spain
Co-authors: Daniel Henderson, Christopher Parmeter

An estimation method and an array of hypothesis tests are presented for varying coefficient multidimensional panel data regression models. The asymptotic distribution of the proposed nonparametric estimator is derived, and the necessary central limit theory is developed to conduct inference and to construct valid tests. The presence of multiple effects over differing dimensions requires nontrivial changes to the well-known central limit theory for U statistics. The types of inference conducted offer a diverse array of hypotheses for applied work, and test statistics are explicitly presented for some of the most important hypothesis tests. To illustrate the usefulness of the proposed suite of tests, an empirical application is provided focusing on the gravity model of international trade. It is found that the standard linear in parameters model is misspecified, suggesting the presence of modeled nonlinearities that could potentially prove useful for policy recommendations. Lastly, an appendix contains a detailed set of simulations that supports the asymptotic developments and reveals that the testing infrastructure possesses correct asymptotic size and high power.

E0350: **Tests for many treatment effects in regression discontinuity panel data models**
Presenter: **Georg Keilbar**, Humboldt-University of Berlin, Germany
Co-authors: Likai Chen, Liangjun Su, Weining Wang

Numerous studies use regression discontinuity design (RDD) for panel data by assuming that the treatment effects are homogeneous across all individuals/groups and pooling the data to- together. It is unclear how to test for the significance of treatment effects when the treatments vary across individuals/groups, and the error terms may exhibit complicated dependence structures. The estimation and inference of multiple treatment effects are examined when the errors are not independent and identically distributed, and the treatment effects vary across individuals/groups. A simple analytical expression is derived for approximating the variance-covariance structure of the treatment effect estimators under general dependence conditions and proposes two test statistics: one is to test for the overall significance of the treatment effect and the other for the homogeneity of the treatment effects. It is found that in the Gaussian approximations of the test statistics, the dependence structures in the data can be safely ignored due to the localized nature of the statistics. This has the important implication that the simulated critical values can be

easily obtained. Simulations demonstrate the tests have superb size control and reasonable power performance in finite samples regardless of the presence of strong cross-section dependence or/and weak serial dependence in the data. The tests on two datasets are applied, and significant overall treatment effects are found in each case.

EO327 Room 104 CONTRIBUTIONS TO THE ESTIMATION PROBLEM IN STOCHASTIC SYSTEMS Chair: Raquel Caballero-Aguila

E0296: **Signal estimation from quantized measurements with random matrices, time-correlated noises and miscellaneous attacks**
Presenter: **Raquel Caballero-Aguila**, Universidad de Jaen, Spain
Co-authors: Jun Hu, Josefa Linares-Perez
The least-squares (LS) linear estimation problem of stochastic signals using quantized measurements with random parameter matrices and time-correlated additive noises is addressed. This scenario is examined under the influence of mixed network attacks, including random deception attacks and denial-of-service (DoS) attacks, using Bernoulli random variables to model the stochastic nature of these attacks. Through an innovation approach, LS centralized fusion filtering and smoothing algorithms are derived, using a covariance-based methodology and a prediction compensation strategy to mitigate the effects of DoS attacks. A simulation example is presented to illustrate the broad applicability of employing random parameter matrices, which effectively cover a wide variety of network-induced uncertainties and random failures, thus offering a more faithful representation of engineering realities. The numerical simulations further corroborate the effectiveness of the proposed estimation scheme and shed light on the impact of random attack probabilities on the estimation accuracy. In sum, the proposed algorithms contribute to the advancement of signal processing and network security research, particularly in scenarios involving quantized measurements, mixed uncertainties, time-correlated noises and miscellaneous network attacks.

E0510: **Centralized fusion prediction in hypercomplex systems with random packet dropouts under properness conditions**
Presenter: **Rosa Maria Fernandez-Alcala**, University of Jaen, Spain
Co-authors: Jose Domingo Jimenez-Lopez, Jesus Navarro-Moreno , Juan Carlos Ruiz-Molina
The centralized fusion prediction problem is investigated for multi-sensor stochastic systems affected by multiple random packet dropouts. The packet dropouts are described as a sequence of independent Bernoulli random variables with known probabilities. The problem is tackled in the tessarine domain and examined under specific properness conditions linked to the vanishing of certain pseudo-correlation functions. Concretely, assuming that the state and the observations are jointly Tk proper, a linear least-mean-square centralized fusion prediction algorithm is devised for correlated state and observations noises. The proposed methodology leads to a dimensionality reduction in the processes involved, resulting in significant computational savings. A numerical example is given to show the implementation of the presented algorithm as well as its better performance over its counterpart in the quaternion domain in different uncertainty scenarios.

E0775: **Event-triggered state estimation for time-varying systems under coder-decoder mechanism**
Presenter: **Hongxu Zhang**, Harbin University of Science and Technology, China
This note discusses the recursive filtering problem for a set of time-varying systems. In order to avoid data collisions, an event-triggered mechanism is employed to determine whether the sensor signal is received by the predictor side. For the traditional event-triggered mechanisms, the sensors determine whether to transmit the sequence by judging the triggering conditions, and the estimator updates the value using the received signal. For the proposed event-triggering mechanism, the predictor determines whether to receive the transmitted information based on its own judgment conditions. In addition, the predicted signal is encoded when transmitted to the estimator side, and then the estimator updates the signal by utilizing the decoded predictor. The aim is to propose a robust state estimation algorithm for addressed systems such that sufficient conditions are derived to ensure the existence of the gain matrices by optimizing the upper bound of the estimation error covariance. Finally, a numerical example is utilized to illustrate the validity and correctness of the developed optimal state estimation strategy.

E0789: **Distributed fusion filtering for multi-sensor multi-rate systems with stochastic nonlinearities and Packet disorders**
Presenter: **Hui Yu**, Harbin University of Science and Technology, China
Co-authors: Pengyun Yue, Jun Hu, Long Xu
The distributed fusion filtering problem is investigated for a class of multi-sensor multi-rate systems (MSMRSs) with stochastic nonlinearities and packet disorders (PDs). The PDs caused by random transmission delay are described by a set of random variables with known probability distributions. In order to facilitate the design of the filtering scheme, a virtual measurement method is introduced to convert multi-rate systems into single-rate systems. The purpose is to design a new local filtering scheme such that, for the presence of stochastic nonlinearities and PDs, the upper bound on each local filtering error covariance (LFEC) is minimized by selecting an appropriate gain matrix. By using mathematical induction, a sufficient condition is obtained to ensure that each LFEC is uniformly bounded. In addition, local estimates are fused based on the inverse covariance intersection (ICI) fusion method. Finally, the effectiveness of the proposed fusion filtering algorithm is shown through a simulation example.

E0787: **Dynamic-data-encryption-based distributed state estimation for nonlinear complex networks against DoS attacks**
Presenter: **Jun Hu**, Harbin University of Science and Technology, China
Co-authors: Bingxin Lei, Na Lin, Zhihui Wu
The optimized distributed state estimation (SE) scheme is proposed for time-varying nonlinear complex networks (NCNs) subject to denial of service (DoS) attacks under the dynamic-data-encryption (DDE) scheme. The DDE scheme is adopted to ensure data privacy by dynamically updating a secret key during data transmission. In particular, the compensation-based measurement model against DoS attacks is established to mitigate the effect of DoS attacks. The main objective is to design a distributed SE scheme such that an optimized upper bound (UB) on the estimation error covariance (EEC) is guaranteed by appropriately designing the estimator gain (EG). Besides, a sufficient condition with respect to the uniform boundedness of SE error is presented. Finally, the viability of the proposed SE scheme is illustrated by a numerical simulation.

EO166 Room 105 BIG DATA COMPUTATION AND APPLICATIONS Chair: Feng Li

E0566: **Modeling the genetic architecture of human complex traits based on genome-wide association summary statistics**
Presenter: **Xia Shen**, Fudan University, China
The focus is to provide an in-depth exploration of the advances in statistical modeling for unraveling the genetic architecture of human complex traits, particularly leveraging genome-wide association summary statistics without requiring individual-level data access. It delves into the cutting-edge big data inference methods for estimating genetic parameters in human genetics and their role in determining the causal relationships across various human complex traits. It is examined how the advancements in genomic analysis have expanded our ability to not only decipher the genetic architecture but also to understand the network of genetic correlations and causal links among diverse human traits and diseases. Through real-world examples, it is demonstrated how the inferred genetic architecture enhances the comprehension of the biological underpinnings of human complex traits. Additionally, it is highlighted how the interplay between genetic correlation and causality has significant implications for the fields of genetics, epidemiology, and healthcare.

E0500: **Distributed sparse regression for high dimensional financial big data based on gradient hard thresholding pursuit**
Presenter: **Shanshan Wang**, Beihang University, China

With the rapid development of internet technology, the global volume of data has experienced explosive growth. In the field of finance, many datasets exhibit significant sample sizes and high-dimensional features. Meanwhile, within these datasets, variables often exhibit extremely high levels of correlation, posing significant challenges to effective analysis. The focus is on a distributed sparse regression algorithm framework tailored to the high-dimensional big data. Firstly, through distributed SVD, column orthogonalization of high-dimensional big data is achieved, effectively eliminating inter-variable correlations and addressing the issue of high correlation. Subsequently, by integrating the GraHTP algorithm based on l_0 regularization penalized regression and employing a divide-and-conquer framework, distributed solutions to high-dimensional sparse regression problems are pursued, thereby achieving rapid and efficient variable selection and parameter estimation. Furthermore, its desirable theoretical properties are also presented, including unbiasedness and sparse recovery, and the proposed algorithm is applied to simulated data characterized by high correlation, high dimensionality, and large sample sizes to validate its performance in both theory and simulation. Lastly, the application of the proposed algorithm to predict the annualized returns of 2,588 Chinese A-share stocks from 2019 to 2022 demonstrates its practical utility in the field of finance.

E0706: **Understanding the impact of emotion on customer conversion: Evidence from automotive live streaming**
Presenter: **Ziyu Xiong**, Peking University, China
Co-authors: Yuntao Dong, Jing Zhou, Xuening Zhu, Hansheng Wang

Automotive live streaming is rapidly emerging as a vital platform for the real-time display of vehicles to audiences. However, the conversion of casual viewers into committed followers persists as an intricate challenge, demanding more in-depth scrutiny. Consequently, understanding the factors that influence viewer conversion is crucial. In this setting, based on the emotional contagion theory, a research model is established using real-time data to investigate the effect of the broadcaster's emotion on viewer conversion. To effectively identify the emotion conveyed through the broadcaster's speech, a convolution-neutral network model is developed, and three emotional variables are constructed to represent the broadcaster's emotion, categorized into valence emotion, arousal emotion and emotional fluctuation. Applying a large-scale sample of 5035 live streams, the empirical study finds that pleasure emotion and arousal emotion can bolster conversion, while emotional fluctuation may hinder it. Interestingly, the study unveils that the conversion gains brought about by female broadcasters' emotions tend to surpass those of their male counterparts, indicating a notable "female advantage" regarding emotional influence and viewer conversion. These findings provide valuable guidance for platforms and broadcasters to optimize marketing interventions such as broadcaster emotion enhancement and viewer engagement improvement in live streaming.

E0370: **Local information advantage and stock returns: Evidence from social media**
Presenter: **Feng Li**, Central University of Finance and Economics, China

The information asymmetry between local and nonlocal investors with a large dataset of stock message board postings is examined. The abnormal relative postings of a firm, i.e., unusual changes in the volume of postings from local versus nonlocal investors, are documented to capture locals' information advantage. This measure positively predicts firms' short-term stock returns as well as those of peer firms in the same city. Sentiment analysis shows that posting activities primarily reflect good news, potentially due to social transmission bias and short sales constraints. The information driving return predictability through content-based analysis is identified. Abnormal relative postings also lead to analysts' forecast revisions. Overall, investors' interactions on social media contain valuable geography-based private information.

E1121: **Day-ahead probability forecasting for redispatch 2.0 measures**
Presenter: **Alla Petukhina**, HTW Berlin, Germany
Co-authors: Maria Basangova, Vlad Bolovaneanu, Alexandra Conda, Awdesch Melzer, Christina Erlwein-Sayer

The purpose is to advance a data-driven, day-ahead forecasting model for assessing the probability, direction, and scale of electrical congestions within Germany's complex power grid. We build a two-stage model, with the first stage employing an XGBoost classifier to predict the probability and direction of congestion events, and the second stage using a regression task for redispatch load forecasting. Utilising state-of-the-art machine learning algorithms, including TSMixer (Time-Series Mixer), LLMtime and NBEATSx (Neural basis expansion analysis with exogenous variables), the model is specifically designed to operate on an hourly basis, thereby offering timely insights for grid management. The analysis uncovers compelling evidence that key exogenous variables, such as real-time meteorological conditions, electricity supply-demand indicators, and Brent oil price fluctuations, can be harnessed to make highly reliable predictions concerning grid congestion events. The economic feasibility of the forecast was also analysed. The model can potentially serve as a useful resource for transmission system operators (TSOs) and policymakers interested in grid management and cost mitigation efforts.

EO210 Room 106 RECENT ADVANCES IN RANDOM MATRIX THEORY AND HIGH-DIMENSIONAL STATISTICS Chair: Jiang Hu

E0606: **Random matrix theory for high-frequency data**
Presenter: **Qiang Liu**, Shanghai University of Finance and Economics, China

In empirical random matrix theory, the sample data are often assumed to be independently and identically distributed, while this is not true for high-frequency data. The model of high-frequency data and some key problems in this field are introduced, and the method of using random matrix theory to solve these problems is discussed. Moreover, the findings are presented as some applications for the estimation of spot covariance matrix in high dimension, which are based on its empirical spectral distribution (ESD) and linear spectral statistics (LSS).

E0284: **The asymptotic distribution of the canonical variable in high-dimensional canonical correlation analysis**
Presenter: **Xiaozhuo Zhang**, Northeast Normal university, China

The asymptotic distribution of the canonical variable in the high-dimensional CCA problem is introduced. The canonical variable is actually the eigenvector of the canonical correlation matrix. Hence, the direction of the canonical variable is described by the cosine of the angle between the eigenvectors, which correspond to the spike eigenvalues of the population and sample canonical correlation matrix, respectively, and give the asymptotic distribution of the cosine of the above angle. Simultaneously, in order to relax the assumption of partial normality, the substitution theorem of the quadratic form is also established.

E0687: **Homogeneity test for high dimensional mixture data**
Presenter: **Yiming Liu**, Jinan University, China

Testing homogeneity, that is, testing whether the data come from a homogeneous or a heterogeneous population, is one of the most important problems in multivariate mixture models. If the data come from a homogeneous population, there is no need to apply a mixture model. A new homogeneity test is considered for high-dimensional mixture data.

E0721: **A CLT for LSS of large dimensional sample covariance matrices with diverging spikes**
Presenter: **Zhijun Liu**, Northeast Normal University, China

The central limit theorem (CLT) is established for linear spectral statistics (LSSs) of a large-dimensional sample covariance matrix when the population covariance matrices are involved with diverging spikes. This constitutes a nontrivial extension of the Bai-Silverstein theorem (BST),

a theorem that has strongly influenced the development of high-dimensional statistics, especially in the applications of random matrix theory to statistics. The new CLT accommodates spiked eigenvalues, which may either be bounded or tend to infinity. The new CLT for LSS is then applied to test the hypothesis that the population covariance matrix is the identity matrix or a generalized spiked model. The asymptotic distributions of the corrected likelihood ratio test statistic and the corrected Nagao's trace test statistic are derived under the alternative hypothesis. Moreover, power comparisons are presented between these two LSSs and Roy's largest root test. In particular, it is demonstrated that except for the case in which there is only one spike, the LSSs could exhibit higher asymptotic power than Roy's largest root test.

E0765: **KOO approach for scalable variable selection problem in large-dimensional regression**
Presenter: **Jiang Hu**, Northeast Normal University, China
An important issue in many multivariate regression problems is to eliminate candidate predictors with null predictor vectors. In a large-dimensional (LD) setting where the numbers of responses and predictors are large, model selection encounters the scalability challenge. Knock-one-out (KOO) statistics hold promise to meet this challenge. The almost sure limits and the central limit theorem of the KOO statistics are derived under the LD setting and mild distributional assumptions (finite fourth moments) of the errors. These theoretical results guarantee the strong consistency of a subset selection rule based on the KOO statistics with a general threshold. To enhance the robustness of the selection rule, a bootstrap threshold is also proposed for the KOO approach. Simulation results support the conclusions and demonstrate the selection probabilities by the KOO approach, with the bootstrap threshold outperforming the methods using the Akaike information threshold, Bayesian information threshold, and Mallow's Cp threshold. The proposed KOO approach is compared with those based on information threshold to a chemometrics dataset and a yeast cell-cycle dataset, which suggests the proposed method identifies useful models.

EO055 Room 108 RECENT DEVELOPMENTS OF LEARNING THEORY Chair: Dingxuan Zhou

E1066: **Functional data analysis**
Presenter: **Dirong Chen**, Beihang University, China
In a random sample of functional data, each subject is recorded as one or several functions. The high or infinite dimensional structure of these data is a rich source of information. On the other hand, the high intrinsic dimensionality of the data poses challenges both for theory and computation. The purpose is to introduce some methods in functional data analysis, with emphasis on the estimation of mean functions and covariance functions based on discretely observed data. A data-driven approach is proposed based on the framelet block thresholding. It has the advantages of adaptivity to local spatial, global smoothness and sampling frequency.

E0271: **Sparse online regression algorithm with insensitive loss functions**
Presenter: **Ting Hu**, School of Management, Xán Jiaotong University, China
A class of kernel-based online gradient descent algorithms is presented for addressing regression problems, which generates sparse estimators in an iterative way to reduce the algorithmic complexity for training streaming datasets and model selection in large-scale learning scenarios. In the setting of support vector regression, the sparse online learning algorithm is designed by introducing a sequence of insensitive distance-based loss functions. Consistency and error bounds are proven, quantifying the generalization performance of such algorithms under mild conditions. The theoretical results demonstrate the interplay between statistical accuracy and sparsity property during the learning process. It is shown that the insensitive parameter plays a crucial role in providing sparsity as well as fast convergence rates. The numerical experiments also support the theoretical results.

E0614: **On learning with Gaussian kernel**
Presenter: **Fusheng Lv**, Nankai University, China
Gaussian kernel-based learning has received great success in practice over the last decades. The theory properties of two types of learning algorithms is discussed generated by the Tikhonov regularization scheme and associated with varying Gaussian kernels. It shows that the algorithm with corrector loss achieves the minimax rate of convergence. The convergence rate of regression risk for the modal regression algorithm will also be presented.

E0865: **Theory of deep convolutional neural networks**
Presenter: **Dingxuan Zhou**, University of Sydney, Australia
Deep learning based on deep neural networks with network architectures has been powerful in practical applications but is less understood theoretically. The network structures such as convolutional architectures give essential difficulty, making the theory different from the classical one for fully connected neural networks. A mathematical theory of approximating and learning functions or operators is presented by deep convolutional neural networks and related schemes.

EO235 Room 109 STATISTICAL LEARNING BASED ON LATENT MODELS AND GRAPH APPROACHES Chair: Xu Zhang

E0182: **Learning mixed latent forest models**
Presenter: **Can Zhou**, Nanjing Audit University, China
Co-authors: Xiaofei Wang, Jianhua Guo
A latent forest model is adopted for mining mixed data. Compared to the traditional latent tree model, the adopted model is more flexible with several trees, and observed variables are allowed to be internal nodes of trees. This model can capture more complex potential mechanisms behind data. The latent structural learning and the parameter estimation for this model are addressed. For structural learning, a consistent bottom-up algorithm is designed, and a theoretical guarantee is provided on a finite sample size bound for the exact structural recovery. For parameter estimation, a moment estimator algorithm is suggested, and the estimator is proven asymptotically normal. The simulation studies indicate that the algorithms performed well in learning the mixed latent forest model. Moreover, the learned mixed latent forest model had a better classification performance than the Naive Bayes model. The real data analysis shows that the learned model captured the hierarchical structure and latent information behind the Changchun mayor hotline data.

E0200: **Quasi-maximum likelihood estimation for large dimensional matrix factor models**
Presenter: **Chaofeng Yuan**, Heilongjiang University, China
A novel approach, called the quasi-maximum likelihood estimation (QMLE), is introduced for estimating large dimensional matrix factor models. In contrast to the principal components-based approach, QMLE takes into account the heteroskedasticity of the idiosyncratic error term, which is simultaneously estimated with other parameters. Theoretical analysis shows that the QMLE estimator of the factor loading matrices achieves faster convergence rates than existing estimators under similar conditions. The asymptotic distributions of the QMLE estimators are also presented. Extensive numerical experiments demonstrate that the QMLE method performs better empirically, especially when heteroscedasticity exists. Furthermore, two real examples in finance and macroeconomics reveal factor patterns across rows and columns which coincide with financial, economic, or geographical interpretations.

E0509: **A graph-theoretic approach to Parkinsonian freezing of gait detection from videos**
Presenter: **Rui Luo**, City University of Hong Kong, Hong Kong
Co-authors: Qi Liu, Chuan Shi, Catherine Liu
Parkinson's disease (PD) is a chronic illness marked by motor complications, notably freezing of gait (FOG), which is a temporary inability to move while walking. FOG's unpredictable onset increases fall risk, highlighting the need for precise analysis of gait transitions for PD diagnosis and treatment. Statistical methods, particularly the Frechet mean, are utilized to analyze gait by representing the central tendency of Laplacian matrices from graph transformations. This facilitates image sequence analysis and comparison, aiding in the mathematical prediction of gait sequence similarity and novel FOG detection methods. Two challenges are tackled: identifying a metric to quantify distances between gait graph Laplacians efficiently and adapting the Frechet statistic for detecting multiple FOG-normal gait transitions without compromising test significance. The approach uses the Log-Euclidean metric to create a metric space for Laplacian matrices, with closed-form solutions for the Frechet mean and variance. A multi-change point detection framework is developed using Frechet analysis, incorporating binary segmentation and incremental computation to improve efficiency. This methodology is confirmed using Kinect3D and AlphaPose datasets, representing a breakthrough in gait analysis for PD management.

E0511: **On statistical analysis of high-dimensional factor models**
Presenter: **Zhigen Gao**, Northeast Normal University, China
High-dimensional factor models have been applied in many fields. The principal component analysis is a popular estimation method for factor models to compute and provides consistent estimators for common factors and factor loadings. Several contributions are made to the asymptotic properties of the principal components estimates (PCE) of factor models as both the sample size T and the variable dimension N go to infinity. Firstly, bias-adjusted estimates of variance are presented for both common factors and idiosyncratic errors. Secondly, an interesting result is found that the predictor of common factors is also biased with the bias of $O_p(T^{(-1)})$ under the PC, especially for the case that T is relatively small compared with N. Meanwhile, the minor modification makes estimates of variance for both common factors and idiosyncratic errors, the predictor of common factors unbiased with a theoretical guarantee. Finally, the asymptotic properties of the PCE are established under a novel proof framework. Simulations are carried out to verify these results.

EO114 Room 110 RECENT ADVANCES IN DESIGN OF EXPERIMENTS AND SAMPLING Chair: Jinyu Yang

E0323: **Three-orthogonal designs robust to nonnegligible two-factor interactions**
Presenter: **Jinyu Yang**, Nankai University, China
Considering the robustness of nonnegligible two-factor interactions, two-level orthogonal arrays with clear two-factor interactions are popular choices. However, if some significant quadratic effects exist, two-level designs do not work to detect them. There is a need to consider designs with three levels. Existing three-level designs with clear effects have a large number of runs, which are not economical choices. The concept of clear two-factor interactions is generalized to three-orthogonal designs, especially three-level ones. The constructions of this new class of robust designs are explored, and their desirable properties are illustrated. Specifically, some of the obtained designs can accommodate a large number of factors ($m = n/4+1$). Due to the appealing structure, the obtained designs make themselves attractive choices for robust designs.

E0305: **Selecting strong orthogonal arrays by linear allowable level permutations**
Presenter: **Guanzhou Chen**, Nankai University, China
Co-authors: Boxin Tang
Space-filling designs are widely used in physical and computer experiments when the model between the response and input factors is uncertain. Recently, another study justified the use of strong orthogonal arrays (SOAs) under a broad class of space-filling criteria. However, when allowable level permutations are applied to an SOA, a class of SOAs can be obtained with different geometrical structures, and it is not clear which one should be selected for practical use. This issue is addressed by considering a representative subset of allowable level permutations called linear allowable level permutations. These special-level permutations offer theoretical convenience in classifying various geometrically non-isomorphic SOAs. Based on these results, construction methods are provided to obtain SOAs that are more space-filling than those in the literature.

E0304: **An efficient quasi-random sampling for copulas**
Presenter: **Sumin Wang**, Nankai university, China
An efficient method is examined for quasi-random sampling of copulas in Monte Carlo computations. Traditional methods, like conditional distribution methods (CDM), have limitations when dealing with high-dimensional or implicit copulas, which refer to those that cannot be accurately represented by existing parametric copulas. Instead, generative models, such as generative adversarial networks (GANs), are proposed to generate quasi-random samples for any copula. GANs are a type of implicit generative model used to learn the distribution of complex data, thus facilitating easy sampling. GANs are employed to learn the mapping from a uniform distribution to copulas. Once this mapping is learned, obtaining quasi-random samples from the copula only requires inputting quasi-random samples from the uniform distribution. This approach offers a more flexible method for any copula. Additionally, theoretical analysis of quasi-Monte Carlo estimators is provided based on quasi-random samples of copulas. Through simulated and practical applications, particularly in the field of risk management, the proposed method is validated, and its superiority is demonstrated over various existing methods.

E0577: **Construction of optimal mixed-level uniform designs**
Presenter: **Liuqing Yang**, Central South University, China
The theory of uniform design has received increasing interest because of its wide application in the field of computer experiments. The generalized discrete discrepancy is proposed to evaluate the uniformity of the mixed-level factorial design. The aim is to give a lower bound of the generalized discrete discrepancy and provide some construction methods of optimal mixed-level uniform designs which can achieve this lower bound. These methods are all deterministic construction methods which can avoid the complexity of stochastic algorithms. Both saturated mixed-level uniform designs and supersaturated mixed-level uniform designs can be obtained with these methods. Moreover, the resulting designs are also 2-optimal and minimum moment aberration designs.

EO038 Room 212 STATISTICAL ANALYSES OF COMPLEX DATA STRUCTURES Chair: Abhik Ghosh

E0207: **Robust inference for high-dimensional logistic regression**
Presenter: **Maria Jaenada**, Universidad Complutense Madrid, Spain
Co-authors: Abhik Ghosh, Leandro Pardo

Real-life data often requires differentiating between two binary classes. In this regard, logistic regression has proven to be a valuable classification technique, offering a straightforward probabilistic interpretation during the classification process. However, in many areas of knowledge, there is a growing need to handle high-dimensional data where the number of variables exceeds the number of observations. High-dimensional data present unique challenges, particularly susceptibility to contamination, which arises from the well-known lack of robustness in classical maximum likelihood methods. To address this issue, a robust procedure is proposed, combining a minimum distance estimator based on the density power divergence (DPD) with regularization techniques, including lasso, adaptive lasso, and non-concave procedures. The proposed techniques find significant applications, especially in scenarios dealing with noisy gene expression data, spectra, and spectral data. Indeed, the practical utility of the techniques is demonstrated in gene selection and patient classification by analyzing real datasets.

E0256: **Weighted likelihood methods for torus data**
Presenter: **Claudio Agostinelli**, University of Trento, Italy
Co-authors: Luca Greco, Giovanni Saraceno

Robust estimation of wrapped models to multivariate circular data (torus data) is considered based on the weighted likelihood methodology. Robust model fitting is achieved by a set of estimating equations based on the computation of data-dependent weights aimed to down-weight anomalous values, such as unexpected directions that do not share the main pattern of the bulk of the data. To solve these equations, an algorithm based on a data augmentation approach and a suitable modification of the expectation-maximization (EM) algorithm is proposed. The advantages and disadvantages of a classification EM algorithm are also discussed. Asymptotic properties and robustness features of the estimators under study are presented, whereas their finite sample behavior has been investigated by Monte Carlo numerical experiment and real data examples.

E0257: **A Bayesian quantile joint modeling of multivariate longitudinal and time-to-event data**
Presenter: **Kiranmoy Das**, Beijing Institute of Mathematical Sciences and Applications, China

Linear mixed models are traditionally used for jointly modeling (multivariate) longitudinal outcomes and event-time(s). However, when the outcomes are non-Gaussian, a quantile regression model is more appropriate. In addition, in the presence of some time-varying covariates, it might be of interest to see how the effects of different covariates vary from one quantile level (of outcomes) to the other and, consequently, how the event-time changes across different quantiles. For such analyses, linear quantile mixed models can be used, and an efficient computational algorithm can be developed. A dataset from the acute lymphocytic leukemia (ALL) maintenance study conducted by Tata Medical Center, Kolkata is analyzed. For this dataset, a Bayesian quantile joint model is developed for the three longitudinal biomarkers and time-to-relapse. An asymmetric Laplace distribution (ALD) is considered for each outcome, and the mixture representation of the ALD is exploited to develop a Gibbs sampler algorithm to estimate the regression coefficients. The proposed model allows different quantile levels for different biomarkers, but still simultaneously estimates the regression coefficients corresponding to a particular quantile combination. Simulation studies are performed to assess the effectiveness of the proposed approach.

E0322: **A divide-and-conquer approach for spatiotemporal analysis of large house price data from Greater London**
Presenter: **Soudeep Deb**, Indian Institute of Management Bangalore, India
Co-authors: Kapil Gupta

Statistical research in real estate markets, particularly understanding spatiotemporal dynamics of house prices, has garnered attention in recent times. Albeit Bayesian methods are common in spatiotemporal modelling, standard Markov chain Monte Carlo (MCMC) techniques are usually slow for large datasets such as house price data. A divide-and-conquer spatiotemporal modeling approach is proposed to tackle this problem. The method involves partitioning the data into multiple subsets and utilizing an appropriate Gaussian process model for each subset in parallel. The results from each subset are then combined via the Wasserstein barycenter technique to obtain the global parameters for the original problem. The divide-and-conquer approach allows multiple observations per spatial and time unit, thereby offering added benefit for practitioners. As a real-life application, house price data of 0.65 million properties are analyzed from 983 middle-layer super-output areas in London for a period of eight years. The methodology renders insightful findings about the effects of various amenities, trend patterns, and the relationship of price to carbon emission. Further, as demonstrated from a cross-validation study, it records good predictive accuracy while balancing the computational need and is proved to be more effective than a conventional Bayesian approach.

E0844: **Robust singular value decomposition**
Presenter: **Ayanendranath Basu**, Indian Statistical Institute, India
Co-authors: Subhrajyoty Roy, Abhik Ghosh

Singular value decomposition (SVD) of a data matrix is traditionally based on the least squares principle and, as a consequence, is very sensitive to the presence of outliers. As a result, the different application domains that use classical methods of SVD may experience degraded performance. A robust singular value decomposition technique is proposed based on the minimum density power divergence estimator. Apart from the theoretical properties of the estimator, a practical algorithm based on alternating weighted regression is also proposed to obtain the estimate. Simulation results and a real application of the video surveillance background modelling problem are presented to demonstrate the performance of the method.

EO224 Room 202 METHODS FOR HIGH-DIMENSIONAL AND HIGH FREQUENCY DATA IN ECONOMICS AND FINANCE Chair: Yayi Yan

E0415: **Testing of regression coefficients under over-parameterized model with hidden confounders**
Presenter: **Yeheng Ge**, The Hong Kong Polytechnic University, China
Co-authors: Xingdong Feng, Mengyun Wu, Shuyan Chen, Tao Li

The existing high-dimensional inference methods could be invalid due to the existence of hidden confounders and non-sparse control variables. A novel parameter inference method is proposed based on the ridge estimator for the over-parameterized model, which is consistent and asymptotically normal even as the non-sparse high dimensional control variables and hidden confounders are present. The proposed method has a closed-form solution, which is new in high dimensional inference and hence can be easily applied to the inference for streaming data. The convergence rate of the bias and variance term is established under various data-generating schemes. A phase transition phenomenon is observed under the cases of $n < p$ and $n > p$, and the corresponding asymptotic results are established. The finite sample performance of the proposed method is well illustrated with simulation studies and real data applications on the GTEx project and FRED-QD database.

E0494: **An efficient multivariate volatility model for many assets**
Presenter: **Wenyu Li**, The University of Hong Kong, China

A flexible and computationally efficient multivariate volatility model is developed, which allows for dynamic conditional correlations and volatility spillover effects among financial assets. The new model has desirable properties such as identifiability and computational tractability for many

assets. A sufficient condition of strict stationarity is derived for the new process. Two quasi-maximum likelihood estimation methods are proposed for the new model with and without low-rank constraints on the coefficient matrices, respectively, and the asymptotic properties for both estimators are established. Moreover, a Bayesian information criterion with selection consistency is developed for order selection, and the testing for volatility spillover effects is carefully discussed. The finite sample performance of the proposed methods is evaluated in simulation studies for small and moderate dimensions. The usefulness of the new model and its inference tools is illustrated by two empirical examples for five stock markets and 17 industry portfolios, respectively.

E0703: **Factor overnight GARCH-Ito models**
Presenter: **Xinyu Song**, Shanghai University of Finance and Economics, China
Co-authors: Donggyu Kim, Minseog Oh, Yazhen Wang
A unified factor overnight GARCH-Ito model is introduced for large volatility matrix estimation and prediction. To account for whole-day market dynamics, the proposed model has two different instantaneous factor volatility processes for the open-to-close and close-to-open periods. At the same time, each embeds the discrete-time multivariate GARCH model structure. To estimate latent factor volatility, the low rank plus sparse structure is assumed, and nonparametric estimation procedures are employed. Then, based on the connection between the discrete-time model structure and the continuous-time diffusion process, a weighted least squares estimation procedure is proposed with the nonparametric factor volatility estimator and its asymptotic theorems are established.

E0705: **Multiperiod dynamic portfolio choice: When high dimensionality meets return predictability**
Presenter: **Wenfeng He**, Renmin University of China, China
Co-authors: Mei Xiaoling, Wei Zhong, Huanjun Zhu
A novel two-step methodology is developed to solve the multiperiod dynamic portfolio choice problem with high dimensional assets in the presence of return predictability conditional on a large number of state predictors. Specifically, in the first step, the new risk-premium projected-PCA (RP-PPCA) method is proposed to reduce the dimension of tradable assets. This method achieves dimension reduction (DR) by estimating latent factors with explanatory power in time series variation and expected return in high-dimension-low-sample-size data. In the second step, dynamic programming is used to solve the multiperiod portfolio choice problem. In each recursive step, an adjusted semiparametric model averaging (AMA) method is adopted to avoid the curse of dimensionality associated with a large set of state variables while remaining computationally efficient. Thus, the two-step approach is named DRAMA, which stands for a combination of a new dimension reduction method and an adjusted semiparametric model averaging method. Analytically, it is shown that the portfolios constructed by the DRAMA are approximately optimal under mild assumptions. Moreover, the numerical results based on empirical data from US stock markets show that the proposed portfolios have excellent out-of-sample performances.

EO103 Room 204 BAYESIAN MODELS AND METHODS FOR COMPLEX DATA Chair: Matias Quiroz

E0348: **Large skew-t copula models and asymmetric dependence in intraday equity returns**
Presenter: **Lin Deng**, University of Melbourne, Australia
Co-authors: Michael Stanley Smith, Worapree Ole Maneesoonthorn
Skew-t copula models are attractive for the modeling of financial data because they allow for asymmetric and extreme tail dependence. The copula implicit is shown in the skew-t distribution of Azzalini and Capitanio, allowing for a higher level of pairwise asymmetric dependence than two popular alternative skew-t copulas. Estimation of this copula in high dimensions is challenging, and a fast and accurate Bayesian variational inference (VI) approach is proposed. The method uses a conditionally Gaussian generative representation of the skew-t distribution to define an augmented posterior that can be approximated accurately. A fast stochastic gradient ascent algorithm is used to solve the variational optimization. The new methodology is used to estimate skew-t factor copula models for intraday returns from 2017 to 2021 on 93 U.S. equities. The copula captures substantial heterogeneity in asymmetric dependence over equity pairs, in addition to the variability in pairwise correlations. Intraday predictive densities are shown from the skew-t copula and are more accurate than from some other copula models, while portfolio selection strategies based on the estimated pairwise tail dependencies improve performance relative to the benchmark index.

E0349: **Natural gradient hybrid variational inference with application to deep mixed models**
Presenter: **Weiben Zhang**, University of Melbourne, Australia
Co-authors: Michael Stanley Smith, Worapree Ole Maneesoonthorn, Ruben Loaiza-Maya
Stochastic models with global parameters theta and latent variables z are common, and variational inference (VI) is popular for their estimation. A variational approximation (VA) is used that comprises a Gaussian with factor covariance matrix for the marginal of theta and the exact conditional posterior of $z|\theta$. Stochastic optimization for learning the VA requires the generation of z from its conditional posterior, while theta is updated using the natural gradient, producing a hybrid VI method. It is shown that this is a well-defined natural gradient optimization algorithm for the joint posterior of (z, theta). Fast-to-compute expressions for the Tikhonov-damped Fisher information matrix required to compute a stable natural gradient update are derived. The approach is used to estimate probabilistic Bayesian neural networks with random output layer coefficients to allow for heterogeneity. Simulations show that using the natural gradient is more efficient than using the ordinary gradient and that the approach is faster and more accurate than the two leading benchmark natural gradient VI methods. In a financial application, the accounting for industry-level heterogeneity is shown using the deep model, which improves the accuracy of probabilistic prediction of asset pricing models.

E0773: **A correlated pseudo-marginal approach to doubly intractable problems**
Presenter: **Matias Quiroz**, University of Technology Sydney, Australia
Co-authors: Robert Kohn, Scott Sisson, Yu Yang
Doubly intractable models are encountered in several fields, e.g. social networks, ecology, and epidemiology. Inference for such models requires the evaluation of a likelihood function, whose normalizing function depends on the model parameters and is typically computationally intractable. A signed pseudo-marginal Metropolis-Hastings (PMMH) algorithm is proposed with an unbiased block-Poisson estimator to sample from the posterior distribution of doubly intractable models. The advantage of the estimator over previous approaches is that its form is ideal for correlated pseudo-marginal methods, which are well known to increase sampling efficiency dramatically. Moreover, analytically derived heuristic guidelines are developed for optimally tuning the hyperparameters of the estimator.

E0823: **Bayesian models for locally stationary time series**
Presenter: **Mattias Villani**, Stockholm University, Sweden
Co-authors: Robert Kohn, Oskar Gustafsson
Some recent developments in Bayesian models and posterior sampling algorithms are presented for locally stationary processes for time series. Extensions to locally stationary seasonal models and spectral analysis are presented.

EO023 Room 209 RECENT ADVANCES IN COMPLEX DATA ANALYSIS Chair: Binyan Jiang

E0355: **Mixture conditional regression with ultrahigh dimensional text data for estimating extralegal factor effects**
Presenter: **Jiaxin Shi**, Peking University, China
Co-authors: Fang Wang, Yuan Gao, Xiaojun Song, Hansheng Wang

Testing judicial impartiality is a problem of fundamental importance in empirical legal studies, for which standard regression methods have been popularly used to estimate the effects of extralegal factors. However, those methods cannot handle control variables with ultrahigh dimensionality, such as those found in judgment documents recorded in text format. To solve this problem, a novel mixture conditional regression (MCR) approach is developed, assuming that the whole sample can be classified into a number of latent classes. Within each latent class, a standard linear regression model can be used to model the relationship between the response and a key feature vector, which is assumed to be of a fixed dimension. Meanwhile, ultrahigh dimensional control variables are then used to determine the latent class membership, where a naive Bayes-type model is used to describe the relationship. Hence, the dimension of control variables is allowed to be arbitrarily high. A novel expectation-maximization algorithm is developed for model estimation. Therefore, the key parameters of interest are estimated as efficiently as if the true class membership were known in advance. Simulation studies are presented to demonstrate the proposed MCR method. A real dataset of Chinese burglary offences is analyzed for illustration purposes.

E0438: **Supervised centrality via sparse network influence regression with an application to 2021 Henan floods social network**
Presenter: **Yingying Ma**, Beihang University, China

The social characteristics of players in a social network are closely associated with their network positions and relational importance. Identifying those influential players in a network is of great importance as it helps to understand how ties are formed and how information is propagated, and in turn, can guide the dissemination of new information. Motivated by a Sina Weibo social network on 2021 Henan Floods where response variables on each node are available, a new notion of supervised centrality is proposed, emphasizing the fact that the centrality of a player is task-specific. To estimate the supervised centrality and identify important players, a novel sparse spatial autoregression is developed by introducing individual heterogeneity to each user. To overcome the computational difficulties in fitting the model for large social networks, a forward-addition algorithm is further developed and shown to consistently identify a superset of the influential nodes. The method is applied to analyze three responses in Henan Floods data: the number of comments, reposts and likes, and obtain meaningful results. The simulation study further corroborates the developed theory.

E0936: **Random interval distillation for detecting multiple changes in dependent dynamic networks**
Presenter: **Weichi Wu**, Tsinghua University, China
Co-authors: Xinyuan Fan

The aim is to propose a new and generic approach for detecting multiple change points in general dependent data, termed random interval distillation (RID). By collecting random intervals with sufficient strength of signals and reassembling them into a sequence of informative short intervals, the new approach captures the shifts in signal characteristics across diverse dependent data forms, including locally stationary high-dimensional time series and dynamic networks with Markov formation. For practical applications, a clustering-based and data-driven procedure is introduced to determine the optimal threshold for signal strength, which is adaptable to a wide array of dependent data scenarios utilizing the connection between RID and clustering. The focus is on the application of RID to dependent dynamic networks with a secondary refinement tailored to it to enhance localization precision. Notably, for low-rank autoregressive networks, the methods achieve the minimax optimality as their independent counterparts. The effectiveness and usefulness of the methodology are examined via extensive simulation studies and a real data example, implementing it in the R-package rid.

E0940: **A two-way heterogeneity model for dynamic networks**
Presenter: **Binyan Jiang**, The Hong Kong Polytechnic University, Hong Kong

Analysis of networks that evolve dynamically requires the joint modelling of individual snapshots and time dynamics. The aim is to propose a new flexible two-way heterogeneity model towards this goal. The new model equips each node of the network with two heterogeneity parameters, one to characterize the propensity to form ties with other nodes statically and the other to differentiate the tendency to retain existing ties over time. With n observed networks, each having p nodes, a new asymptotic theory for the maximum likelihood estimation of $2p$ parameters is developed when $np \to \infty$. The global non-convexity of the negative log-likelihood function is overcome by virtue of its local convexity, and a novel method of moment estimator is proposed as the initial value for a simple algorithm that leads to the consistent local maximum likelihood estimator (MLE). To establish the upper bounds for the estimation error of the MLE, a new uniform deviation bound is derived, which is of independent interest. The theory of the model and its usefulness are further supported by extensive simulation and real data analysis.

E1065: **Semiparametric analysis of deep ordinal choice models**
Presenter: **Yiwei Fan**, Beijing Institute of Technology, China

Deep learning has achieved considerable success across various application domains, coupled with notable advancements in its theoretical foundations. Despite these strides, the exploration of deep learning in the context of the maximum rank correlation (MRC) estimator remains relatively limited. A smoothed MRC estimator is introduced for the ordinal choice model, which integrates a linear function for interpretation and a nonlinear function fitted using deep neural networks. A two-step algorithm is designed for estimation, which keeps the order relation among outputs without the parallelism assumption. Under regular conditions, statistical properties are established for the smoothed MRC estimator, including identification, convergence rate, and minimax-optimality, where the number of categories is allowed to increase with the sample size. Simulations and extensive real-world applications demonstrate the advantages of the proposed method in terms of classification accuracy and interpretability.

EO123 Room 210 APPLICATIONS OF SPATIAL ECONOMETRICS TO HIGH-DIMENSIONAL DATA Chair: Yasumasa Matsuda

E0388: **Bias-corrections for correlations and heteroskedasticities in large linear panel models with interactive effects**
Presenter: **Runyu Dai**, Tohoku University, Japan
Co-authors: Yasumasa Matsuda, Takashi Yamagata

An efficient iterative principal components (IPC) estimator of a large linear panel data model is considered with common factor type interactive effects. It is well-known that the original IPC estimator suffers from bias due to correlated and heteroskedastic idiosyncratic errors in cross-sectional and serial dimensions. The developed estimator corrects the bias by a residual sparse regression to correct correlations in both dimensions simultaneously, plus a conventional bias correction for heteroskedasticities. The asymptotic properties of the proposed estimator are rigorously established. Monte Carlo simulations show the approach works well in finite samples both in estimation and inference.

E0450: **Spatial factor models for surface time series**
Presenter: **Yasumasa Matsuda**, Tohoku University, Japan
Co-authors: Runyu Dai, Yi Wu

Surface time series is a kind of spatial panel data, i.e., panel data when a cross-sectional unit is spatially observed data. In surface time series

analysis, N, the spatial dimension, is sometimes very large, while T, the time length, is usually short. It follows that popular analysis by factor models has significant difficulties in managing this feature. The focus is on surface time series as a time series of functional data on space, for which functional principal component analysis (fPCA) is introduced to define factor models. fPCA works to manage the feature of large N and short T regarding the panel as a functional time series. Spatial factor models are applied to yearly income per capita in 1700 cities in Japan for 22 years from 2000 to 2021, and it demonstrates how fPCA works to solve the problem.

E0611: **Geographically weighted regression for compositional data**
Presenter: **Takahiro Yoshida**, The University of Tokyo, Japan
Geographically weighted regression (GWR) is a widely used spatial data analysis technique across various fields. Additionally, the extension for non-Gaussian distributed data has been progressing. However, studies on the extension for compositional data are limited. Spatial regression model developments for compositional data are crucial topics in compositional data analysis (CoDA) literature. Geostatistical compositional models, such as the compositional kriging model, are popular approaches because CoDA has been historically developed in geosciences in which a continuous spatial process can be assumed. Other study approaches employ conditional autoregressive models or simultaneous autoregressive (spatial econometric) models. Although spatial autocorrelated errors are considered in these models, models for compositional data with spatial heterogeneity or spatially varying relationships remain limited. The objective is to build a GWR model for compositional data to consider both spatial heterogeneity and the constant-sum constraint. The GWR model and log-ratio techniques of CoDA are accommodated, and then the GWR model in the simplex space. This model is applied to analyze an actual data set from a US census survey data. The interpretational usefulness of the results of spatially varying compositional semi-elasticities was empirically verified.

E0808: **Asset pricing with co-search interaction**
Presenter: **Stanley Iat-Meng Ko**, Tohoku University, Japan
Co-authors: Yasumasa Matsuda, Runyu Dai, Jieying Zhang
The effect of internet co-search activities of listed stocks on their returns in the stock market is studied. The internet traffic on the US Securities and Exchange Commission's (SEC) Electronic Data Gathering, Analysis, and Retrieval (EDGAR) website is explored, which holds all public US companies' information with hundreds of thousands of document views per day by users. Co-searched firms are identified, i.e. one firm is searched subsequently after another, and such information is incorporated into the conventional asset pricing model. The contributions are threefold. First, the micro-level behavioral information of individual stocks is introduced to the empirical asset pricing literature, whereas traditional asset pricing studies focus on aggregated portfolios. Second, with the identification of co-search peers, the co-search network is defined and constructed in the universe of trading stocks. Through the lens of the co-search network, the dependence of the virtual spatial stock return across the network is identified. Third, the traditional linear asset pricing model is extended using the spatial arbitrage pricing theory (S-APT).

EO158 Room 307 ADVANCES IN FUNCTIONAL DATA ANALYSIS AND MACHINE LEARNING FOR COMPLEX DATA Chair: Yuko Araki

E0822: **Comparison of the survival analysis with functional convex clustering and joint model for complex data**
Presenter: **Yuko Araki**, Tohoku University, Japan
Two methods are considered for cases where the pattern of change in covariates varies over time, and space affects survival time. One is a Cox proportional hazards model incorporating functional convex clustering, and the other is a joint model. The former facilitates capturing spatiotemporal variations and naturally incorporates the patterns of dynamics as covariates in survival time analysis. On the other hand, the joint model, by handling both longitudinal data and survival time simultaneously, is known to mitigate bias in survival analysis results due to temporal changes in covariates. The behavior of the two models is examined and compared through numerical experiments.

E0730: **Nonparametric function-on-scalar regression with deep learning**
Presenter: **Kazunori Takeshita**, Osaka university, Japan
Co-authors: Yoshikazu Terada
Nonlinear function-on-scalar (FOS) regression is considered using neural networks, where the predictor variables are scalar, and the response variables are functional data. Previous research proposed methods based on the approximation theory of neural networks. The advantage of the method is that it can be applied without imposing specific assumptions (such as additive models) on the true function, even when the dimensions of the predictor variables are relatively large. However, since the estimator is represented through basis functions, it lacks the adaptability that is known as an advantage of deep learning. To overcome this shortcoming, a new adaptive estimator is proposed using deep neural networks and the theoretical properties of the proposed method are shown. More precisely, the anisotropic Besov space is considered a model of true function. The anisotropic Besov space is characterized by direction-dependent smoothness and involves several function classes as special cases. The results indicate that it is possible to alleviate the curse of dimensionality when the true function has high anisotropic smoothness. To evaluate the performance of the proposed method, numerical experiments and real data analysis are conducted.

E0452: **Statistical analysis on in-context learning**
Presenter: **Masaaki Imaizumi**, The University of Tokyo, Japan
Deep learning and artificial intelligence technologies have made great progress, and the usage of foundation models has attracted strong attention to their general ability. Motivated by this fact, mathematical understanding is required to efficiently control and develop these technologies. A statistics-based analysis of a scheme called in-context learning is presented, which is a useful framework of meta-learning to describe foundation models. It is argued that in-context learning can efficiently learn the latent structure of the data, using the property of transformers used in the learning scheme can efficiently handle the distribution of observations.

E0505: **Prediction of trajectory for variable-domain functional data**
Presenter: **Hidetoshi Matsui**, Shiga University, Japan
Co-authors: Yoshikazu Terada
Functional data analysis (FDA) is one of the most useful methods for analyzing longitudinal data and has been widely used in various fields such as medicine and engineering. Standard FDA methods focus on functional data whose domains are identical for each individual. In contrast, the data is considered where the endpoints of functions differ for each individual. The problem of predicting the trajectory and endpoint of a function observed is approached only up to a specific time point under the condition that a set of other functions is completely observed. The idea of variable-domain functional data and dynamic prediction is applied to achieve this. The analysis of real data demonstrates the effectiveness of the proposed method.

E0449: **Nonparametric inference on Frechet mean and related population objects on manifolds**
Presenter: **Daisuke Kurisu**, The University of Tokyo, Japan
Co-authors: Taisuke Otsu
The focus is on the inference of the Frechet mean and related population objects on manifolds. The concept of nonparametric likelihood is developed for manifolds, and general inference methods are proposed by adapting the theory of empirical likelihood. In addition to the basic asymptotic properties, such as Wilks theorem of the empirical likelihood statistic, several generalizations of the proposed methodology are presented. Simula-

tion and real data examples illustrate the usefulness of the proposed methodology and its advantage against the conventional Wald test.

EO062 Room 313 NEW ADVANCES IN COMPLEX TIME SERIES AND SPATIAL LEARNING AND MODELLING Chair: Zudi Lu

E0220: **A varying-coefficient model of expected shortfall and its application to mixed-frequency data**
Presenter: **Jiangtao Wang**, Huazhong Normal University, China
The focus is to develop a nonparametric varying-coefficient approach for modelling the value-at-risk (VaR) and expected shortfall (ES) simultaneously since the ES is not elicitable but can be elicitable combined with VaR. Previous studies on conditional ES estimated only considered parametric model see-ups, which account for the stochastic dynamic of asset returns but ignore other exogenous economic variables and the investment situation. The approach overcomes this drawback and allows VaR and ES to be modelled directly in a flexible way using covariates that may be exogenous, especially sampled at different frequencies compared with the return series. A three-step procedure is developed based on the local linear smoothing technique for estimating the coefficient functions and establishing the consistency and asymptotic normality of the resultant estimator. To overcome the challenge associated with calculating the asymptotical variance, a random weight resampling approach is designed by perturbing the loss function directly. Simulation studies are presented to demonstrate the finite-sample performance of the proposed estimator. The favorable performance of the proposed method is further illustrated via an application for forecasting ES with mixed-frequency data.

E0496: **Spatial-temporal synthetic error model of causal analysis with application to policy causal effect evaluation**
Presenter: **Yan Zhang**, University of Southampton, United Kingdom
Co-authors: Zudi Lu
Causal analysis of spatial-temporal data is challenging owing to spatial-temporal interactions. The synthetic control method (SCM) is popular in estimating the causal effect of a given intervention on a single or a small number of units in a non-spatial panel data setting by weighted averaging of the control units to balance the outcomes and covariates of the treated unit. Inspired by the ideas of synthetic control method and spatial-temporal models, a spatial-temporal synthetic error model (STSEM) is proposed as a new framework of linear spatial-temporal causal inference model to infer the causal effect of some given intervention on the metric that is of interest for spatial-temporal data, with its synthetic weights determined by LASSO regression. Asymptotic properties of the proposed model are established, followed by which the significance of the causal effect can be tested. In addition, its performance is also compared in causal effect inference with the traditional SCM, the augmented SCM (ASCM), and a simplified STAR-PLR model in a simulation study and an empirical study, in which the causal effect of the Kansas tax cut on its GDP is demonstrated for inference.

E0596: **New asymptotics applied to functional coefficient regression and climate sensitivity analysis**
Presenter: **Qiying Wang**, University of Sydney, Australia
A general asymptotic theory is established for sample cross moments of nonstationary time series, allowing for long-range dependence and local unit roots. The theory provides a substantial extension of earlier results on nonparametric regression that include near-cointegrated nonparametric regression as well as spurious nonparametric regression. Many new models are covered by the limit theory, among which are functional-coefficient regressions in which both regressors and the functional covariate are nonstationary. Simulations show that finite sample performance matches well with the asymptotic theory and has broad relevance to applications while revealing how dual nonstationarity in regressors and covariates raises sensitivity to bandwidth choice and the impact of dimensionality in nonparametric regression. An empirical example is provided involving climate data regression to assess Earth's climate sensitivity to CO_2, where nonstationarity is a prominent feature of both the regressors and covariates in the model. This application is the first rigorous empirical analysis to assess the nonlinear impacts of CO_2 on Earth's climate.

E0650: **Nonlinear interpolation for irregularly observed spatial data: Learning from an additive kriging**
Presenter: **Zudi Lu**, University of Southampton, United Kingdom
Co-authors: Fumiya Akashi, Yan Sun, Dag Tjoestheim
Many applications have stimulated vast interest in theoretical and empirical research on spatial prediction. In fact, the need to obtain accurate predictions from observed data can be found in all scientific disciplines. Kriging, as a method of spatial prediction, was originally coined by a past study for optimal spatial linear prediction under minimum mean squared error after another study on mining grade evaluation. Since then, linear kriging and its extensions, such as generalised linear kriging, have been extensively developed in geostatistics. In general, linear kriging is optimal under Gaussian data assumption, but it often is not, as the Gaussianity for real data is widely violated. A nonlinear kriging is developed via an additive semiparametric structure to learn the nonlinear additive component functions. Theoretical consistency of the estimation is established under mild spatial mixing assumption on the data. Furthermore, the learned nonlinear structures of the component functions are parametrized in threshold (piecewise linear) functions, and hence, a nonlinear kriging is implemented by a co-kriging procedure. Both the simulation data and real data examples demonstrate that our learned nonlinear kriging can significantly outperform the traditional linear kriging for spatial prediction in terms of cross-validation error rates.

E0900: **Testing linearity of network interaction functions**
Presenter: **Francesca Rossi**, University of Verona, Italy
Co-authors: Abhimanyu Gupta, Jungyoon Lee
A computationally straightforward test is proposed and theoretically justified for the linearity of a network interaction function. Such functions commonly arise due to practitioner-imposed specifications or optimizing behavior. The test is nonparametric but based on the Lagrange Multiplier principle. This entails estimation only under the null hypothesis, which imposes a familiar, easy-to-estimate linear spatial autoregressive model. Monte Carlo simulations show excellent size control and power. An empirical study with Finnish data illustrates the test's practical usefulness, shedding light on debates in the public finance literature on the presence of tax competition among neighboring municipalities.

EO245 Room 405 ADVANCES ON SOME THEORETICAL AND APPLIED STATISTICS Chair: Hui Jiang

E1016: **Adjusted expected improvement for cumulative regret minimization in noisy Bayesian optimization**
Presenter: **Shouri Hu**, University of Electronic Science and Technology of China, China
The expected improvement (EI) is one of the most popular acquisition functions for Bayesian optimization (BO) and has demonstrated good empirical performances in many applications for minimizing simple regret. However, under the evaluation metric of cumulative regret, the performance of EI may not be competitive, and its existing theoretical regret upper bound still has room for improvement. To adapt the EI for better performance under cumulative regret, a novel quantity called the evaluation cost is introduced, which is compared against the acquisition function. With this, the expected improvement-cost (EIC) algorithm is developed. In each iteration of EIC, a new point with the largest acquisition function value is sampled, only if that value exceeds its evaluation cost. If none meets this criteria, the current best point is resampled. This evaluation cost quantifies the potential downside of sampling a point, which is important under the cumulative regret metric, as the objective function value in every iteration affects the performance measure. A high-probability regret upper bound of EIC is established in theory based on the maximum information gain, which is tighter than the bound of existing EI-based algorithms. It is comparable to the regret bound of popular BO algorithms such as Thompson

sampling (GP-TS) and upper confidence bound (GP-UCB). Experiments are further performed to illustrate the improvement of EIC over several popular BO algorithms.

E1027: **Change plane structural equation model**
Presenter: **Jingli Wang**, Nankai University, China
A multi-threshold change plane structural equation model is introduced to accommodate varying causal effects according to change planes. In the proposed model, the number of thresholds, their locations, and the coefficients of change planes are all unknown. This model serves as a non-trivial extension of the multi-threshold change point structural equation model. A group selection method is first used to detect the number of thresholds. Then, an iterative procedure is adopted to refine the estimation of threshold locations and other model parameters. The Wald ratio method is applied to estimate causal effects for identified subgroups. Furthermore, the consistency of estimated parameters and asymptotic distribution of the estimated causal effects are established. Finally, the performance of the proposed methodology is demonstrated by simulations and a real data example.

E1038: **Fixed-domain asymptotics for Gaussian random fields**
Presenter: **Saifei Sun**, City University of Hong Kong, Hong Kong
Co-authors: Wei-Liem Loh
The purpose is to consider the covariance parameter estimation for Gaussian random fields that are observed with measurement error and irregularly spaced design sites on a fixed and bounded domain. The Gaussian random fields are assumed to have smooth mean functions and isotropic covariance functions belonging to powered exponential, Matrn, or generalized Wendland class. Under fixed-domain asymptotics, consistent estimators are proposed for three microergodic parameters, namely the nugget, the smoothness parameter, and a parameter related to the coefficient of the principal irregular term of the covariance function. Upper bounds for the convergence rate of these estimators are also established.

E0924: **A distribution-free mixed-integer optimization approach to hierarchical modelling of clustered and longitudinal data**
Presenter: **Tom Chen**, Harvard Pilgrim Health Care and Harvard Medical School, United States
Co-authors: Madhav Sankaranarayanan, Intekhab Hossain
Recent advancements in mixed integer optimization (MIO) algorithms and hardware enhancements have led to significant speedups in resolving MIO problems. These strategies have been utilized for optimal subset selection, specifically for choosing k features out of p in linear regression given n observations. The method is broadened to facilitate cluster-aware regression, where selection aims to choose λ out of K clusters in a linear mixed effects (LMM) model with n_k observations for each cluster. Through comprehensive testing on a multitude of synthetic and real datasets, the method efficiently solves problems within minutes. Through numerical experiments, it is also shown that the MIO approach outperforms both Gaussian- and Laplace-distributed LMMs in terms of generating sparse solutions with high predictive power. Traditional LMMs typically assume that clustering effects are independent of individual features. However, an innovative algorithm is introduced that evaluates cluster effects for new data points, thereby increasing the robustness and precision of this model. The inferential and predictive efficacy of this approach is further illustrated through its application in student scoring and protein expression.

E1124: **Determining the Number of Common Functional Factors with Twice Cross-Validation**
Presenter: **Hui Jiang**, Huazhong University of Science and Technology, China
Co-authors: Lei Huang
The semiparametric factor model serves as a vital tool to describe the dependencepatterns in the data. It recognizes that the common features observed in the data areactually explained by functions of specific exogenous variables. Unlike traditional factor models, where the focus is on selecting the number of factors, our objective here isto identify the appropriate number of common functions, a crucial parameter in thismodel. In this paper, we develop a novel data-driven method to determine the number of functional factors using cross validation (CV). Our proposed method employsa two-step CV process that ensures the orthogonality of functional factors, which werefer to as Functional Twice Cross-Validation (FTCV). Extensive simulations demonstrate that FTCV accurately selects the number of common functions and outperformsexisting methods in most cases. Furthermore, by specifying market volatility as theexogenous force, we provide real data examples that illustrate the interpretability ofselected common functional factors.

EO167 Room 406 APPLIED PROBABILITY AND OPTIMISATION METHODS IN DATA SCIENCE Chair: Sarat Moka

E0352: **Stein pi-importance sampling**
Presenter: **Wilson Chen**, The University of Sydney, Australia
Co-authors: Congye Wang, Heishiro Kanagawa, Chris Oates
Stein discrepancies have emerged as a powerful tool for retrospective improvement of Markov chain Monte Carlo output. However, the question of how to design Markov chains that are well-suited to such post-processing has yet to be addressed. Stein importance sampling is studied, in which weights are assigned to the states visited by a Pi-invariant Markov chain to obtain a consistent approximation of P, the intended target. Surprisingly, the optimal choice of Pi is not identical to the target P; an explicit construction for Pi is therefore proposed based on a novel variational argument. Explicit conditions for convergence of Stein pi-importance sampling are established. For 70% of tasks in the PosteriorDB benchmark, a significant improvement over the analogous post-processing of P-invariant Markov chains is reported.

E1114: **Homogeneity and sparsity pursuit using robust adaptive fused lasso**
Presenter: **Le Chang**, the Australian National University, Australia
Fused lasso regression is a popular method for identifying homogeneous groups and sparsity patterns in regression coefficients based on either the presumed order or a more general graph structure of the covariates. However, the traditional fused lasso may yield misleading outcomes in the presence of outliers. We propose an extension of the fused lasso, namely, the robust adaptive fused lasso (RAFL), which pursues homogeneity and sparsity patterns in regression coefficients while accounting for potential outliers within the data. By utilizing Huber's loss or Tukey's biweight loss, the RAFL can resist outliers in the responses or in both the responses and the covariates. We also demonstrate that when the adaptive weights are properly chosen, our proposed RAFL achieves consistency in variable selection, consistency in grouping, and asymptotic normality. Furthermore, a novel optimization algorithm that employs the alternating direction method of multipliers, embedded with an accelerated proximal gradient algorithm, is developed to solve the RAFL efficiently. A simulation study shows that the RAFL offers substantial improvements in terms of both grouping accuracy and prediction accuracy compared with the fused lasso, particularly when dealing with contaminated data. Additionally, a real analysis of cookie data demonstrates the effectiveness of the RAFL.

E0237: **Markov Switching tensor regression**
Presenter: **Qing Wang**, Ca Foscari University, Italy
Co-authors: Roberto Casarin, Radu Craiu
A Markov Switching tensor regression model is proposed, which allows for model instability and accounts for multi-dimensional array data. Regarding model instability, the parameters are assumed to be time-varying and are driven by latent processes to address structural breaks in the data.

Regarding high dimensionality, the Soft PARAFAC strategy is followed to achieve dimensionality reduction while preserving the structural information between the covariates. Modified multi-way shrinkage prior is further imposed to address over-parametrization issues. An efficient MCMC algorithm that adopts random scan Gibbs within a back-fitting strategy is developed to achieve better scalability of the posterior approximation. The performance of the MCMC algorithm is demonstrated using synthetic datasets in simulation studies. Real-world applications test the proposed model against the benchmark Lasso regression, where the model delivers superior performance.

E0761: **Optimization methods for best subset selection problem in high-dimensional linear dimension reduction models**
Presenter: **Sarat Moka**, The University of New South Wales, Australia
Co-authors: Zdravko Botev, Benoit Liquet, Samuel Muller
Principal components analysis and partial least squares are two popular methods used for dimensionality reductions, and they have numerous applications in several fields, including economics. Both these methods build principal components, which are new variables that are combinations of all the original variables. A key drawback of these principal components is the difficulty of interpreting them when the number of variables is large. To overcome this difficulty, it is desirable to define principal components from the most relevant variables. However, selecting the most relevant variables is an NP-hard problem, particularly challenging in high-dimensional settings where the number of features can be far higher than the number of observations. The problem is stated as a best subset selection problem, and new efficient optimization methods have been developed to address this problem. Several empirical experiments are provided to illustrate the efficacy of the approaches.

EO063 Room 408 HIGH DIMENSIONAL AND COMPLEX DATA ANALYSIS WITH APPLICATIONS Chair: Su-Yun Huang

E0470: **EM estimation of the B-spline Copula with penalized log-likelihood function**
Presenter: **Xiaoling Dou**, Japan Womens University, Japan
Co-authors: Satoshi Kuriki, Gwo Dong Lin, Donald Richards
The B-spline copula function is defined by a linear combination of elements of the normalized B-spline basis. A modified EM algorithm is developed to maximize the penalized log-likelihood function, wherein the smoothly clipped absolute deviation (SCAD) penalty function is used for the penalization term. Simulation studies are conducted to demonstrate the stability of the proposed numerical procedure, show that penalization yields estimates with smaller mean-square errors when the true parameter matrix is sparse, and provide methods for determining tuning parameters and model selection. As an example is analyzed a data set consisting of birth and death rates from 237 countries, available at the website "Our World in Data", and the marginal density and distribution functions of those rates are estimated together with all parameters of our B-spline copula model.

E0575: **A genetic model for the analysis of quantitative traits in autotetraploid species**
Presenter: **Chen-Hung Kao**, Institute of Statistical Science, Taiwan
Many important crops and plants are polyploid species (with three or more complete chromosome sets). For example, potatoes, coffee, and alfalfa are autotetraploid plants, and some economically important aquaculture animals, such as Atlantic salmon and trout fish, are also autotetraploid. Statistical methods of QTL mapping have been well-established in diploid species. However, statistical methods are generally lacking or inadequate for polyploidy species, particularly autotetraploids, largely due to a lack of analytical methods that account for the complexities of autotetrasomic inheritance. While also considering the complexity of the autotetraploid genomes, a genotypic model is constructed that links the genotypic values of individuals with their genotypes (the coded variables) and can be used in the QTL mapping study. When applied to the autotetraploid species, the model can offer the advantage of consistency in model interpretation and statistical estimation as compared to the models of currently being employed in the analysis of quantitative traits.

E0463: **A geometric algorithm for contrastive principal component analysis in high dimension**
Presenter: **Shao-Hsuan Wang**, National Central University, Taiwan
Co-authors: Su-Yun Huang
Principal component analysis (PCA) has been widely used in exploratory data analysis. Contrastive PCA, a generalized method of PCA, is a new tool used to capture features of a target dataset relative to a background dataset while preserving the maximum amount of information contained in the data. With high dimensional data, contrastive PCA becomes impractical due to its high computational requirement of forming the contrastive covariance matrix and associated eigenvalue decomposition for extracting leading components. A geometric curvilinear-search method is proposed to solve this problem and provide a convergence analysis. The approach offers significant computational efficiencies. Specifically, it reduces the time complexity from $O(maxn, mp)$ to a more manageable $O(maxn, mpr)$, where n, m are the sample sizes of the target data and background data, respectively, p is the data dimension and r is the number of leading components.

E0391: **Bayesian multitask learning for medicine recommendation based on online patient reviews**
Presenter: **Yichen Cheng**, Georgia State University, United States
Co-authors: Yusen Xia, Xinlei Wang
A drug recommendation model is proposed that integrates information from both structured data (patient demographic information) and unstructured texts (patient reviews). It is based on multitask learning to predict review ratings of several satisfaction-related measures for a given medicine, where related tasks can be learned from each other for prediction. The learned models can then be applied to new patients for drug recommendation. This is fundamentally different from most recommender systems in e-commerce, which do not work well for new customers (referred to as the cold-start problem). To extract information from review texts, both topic modeling and sentiment analysis are employed. Variable selection is further incorporated into the model via Bayesian LASSO, which aims to filter out irrelevant features. To the best of knowledge, this is the first Bayesian multitask learning method for ordinal responses. Multitask learning is also applied to medicine recommendations for the first time.

E1123: **Generative adversarial models for extreme geospatial downscaling**
Presenter: **Guofeng Cao**, University of Colorado Boulder, United States
Addressing the challenges of climate change requires accurate and high-resolution mapping of climate variables. However, many existing datasets, such as the outputs of the state-of-the-art numerical climate models, are only available at very coarse resolutions due to model complexity and high computational demand. Deep learning methods, particularly generative adversarial networks (GANs), have proved effective in improving geospatial datasets. A conditional GAN-based extreme geospatial downscaling method is described. Compared to most existing approaches, the method can generate highly accurate datasets from very low-resolution inputs. More importantly, the method considers the uncertainty inherent to the downscaling process that tends to be ignored in existing methods. Given an input, the method can produce a multitude of plausible high-resolution samples. These samples allow for an exploration and inferences of model uncertainty and robustness. With a case study of gridded climate datasets, we showcase the performances of the framework in downscaling tasks with large scaling factors (up to 64) and highlight the advantages of the framework with a comparison with commonly used downscaling methods, including area-to-point kriging, deep image prior, enhanced super-resolution generative adversarial networks (ESRGAN), physics-informed resolution-enhancing GAN (PhIRE GAN), and an efficient diffusion model for remote sensing image super-resolution (EDiffSR).

EC266 Room 111 HIGH-DIMENSIONAL STATISTICS AND ECONOMETRICS Chair: Wanjie Wang

E0214: **On high-dimensional data analysis**
Presenter: **Wei-Cheng Hsiao**, Soochow University, Taiwan
Co-authors: Ching-Kang Ing
Big data, ubiquitous in various fields of natural and social sciences, often contains a large number of variables and features. This motivates the study of the model (variable or feature) selection problems in high-dimensional sparse models. A novel high-dimensional model selection procedure is introduced and demonstrates how variable selection consistency in high-dimensional interaction models is achieved. In addition, motivated by wafer data, a new high-dimensional model identification method is proposed and its selection consistency is obtained in situations where the location and dispersion components of the model obey sparsity conditions. Simulations and real data analysis are given to illustrate the finite sample performance of the proposed methods.

E0278: **High-dimensional convex nonparametric least squares with Lasso penalty**
Presenter: **Zhiqiang Liao**, Aalto University, Finland
The problem of variable selection is studied in convex nonparametric least squares. Whereas the Lasso is a popular technique for simultaneous estimation and variable selection, its performance is unknown in convex regression problems. The performance of the Lasso regularized convex nonparametric least squares estimator is investigated in a high-dimensional setting, and an alternative approach is proposed based on the unique structure of the subgradients. The proposed estimators perform favorably, while generally leading to sparser models relative to the other predictive models via the standard Lasso. Further, the estimators can be expressed as solutions to convex optimization problems and are amenable to modern optimization algorithms.

E0902: **Computational strategies for regression model selection in the high-dimensional case**
Presenter: **Marios Demosthenous**, Justus Liebig University of Giessen, Germany
Co-authors: Cristian Gatu, Erricos Kontoghiorghes
Computational strategies for finding the best-subset regression models are proposed. The case of high-dimensional (HD) data where the number of variables exceeds the number of observations is considered. Within this context, a theoretical combinatorial solution is proposed. It is based on a regression tree structure that generates all possible subset models. An efficient branch-and-bound algorithm that finds the best submodels without generating the entire tree is adapted to the HD case. Furthermore, the R package lmSubsets is employed in the HD case to identify the best submodel based on the AIC family selection criteria. Preliminary experimental results are presented and analyzed. The efficient extension of the lmSelect algorithm to HD is discussed.

E0907: **Quantile forward regression in high-dimensional distributional counterfactual analysis**
Presenter: **Hongqi Chen**, Hunan University, China
The focus is on introducing a novel quantile forward regression approach for constructing distributional artificial counterfactuals. In the context of counterfactual analysis, where the number of control units frequently surpasses the pre-treatment time dimension, the quantile forward regression provides an approach to mitigate this challenge. The methodology involves the step-wise selection of control units from a candidate set. The theoretical properties of quantile forward regression are established, encompassing a bound on its weak submodularity ratio and asymptotic convergence results, and its asymptotic efficacy is assessed. Through extensive Monte Carlo simulations, the superior finite sample performance of the quantile forward regression approach is showcased compared to the l_1-penalization approach. The evaluation focuses on counterfactual prediction accuracy and the selection of control units. Finally, the application of quantile forward regression is demonstrated in an empirical study, analyzing the impact of an anti-corruption campaign on luxury watch importation.

E0817: **Adaptive detection of change-points for high-dimensional covariance matrices**
Presenter: **Xiaoyi Wang**, Beijing Normal University at Zhuhai, China
A series of tests are built based on U-statistics to test the high-dimensional covariance matrix change points within the temporal independence assumption. The asymptotic distributions of the constructed U-statistics are derived under the null and local alternative hypotheses. Then, a family of maximum-type statistics is proposed, after which two test methods are developed based on the combination of the p-values of these maximum-type statistics. Some methods are also proposed to estimate the location of the change-point and obtain their corresponding convergence rates. In addition, three new adaptive estimations are built. Finally, the binary segmentation method is proposed to be combined with the three adaptive estimators to detect multiple change-points. The simulation study shows that the proposed test methods can maintain high power under alternatives with different sparsity levels. The proposed adaptive estimators perform well under different alternatives with both single and multiple change-points.

EC305 Room 207 APPLIED STATISTICS AND ECONOMETRICS Chair: William WL Wong

E0195: **Determinants of the retention intention of teachers: Evidence from a multilevel model using TALIS data**
Presenter: **Mike Smet**, KU Leuven, Belgium
Rising teacher turnover rates in various countries lead to teacher shortages in schools and reduced education quality. The aim is to investigate teacher retention determinants, aiding policy-makers in improving retention. It builds on literature highlighting factors such as gender, age, school type, job satisfaction, self-efficacy, discipline, school climate, stress, professional development, work conditions, compensation, and teacher-student relationships. Using a hierarchical linear model (HLM) to address the data's nested nature (teachers within schools, within regions), it examines teachers' intent to remain in the profession. Random intercepts at the school and region levels account for the nested structure. The dependent variable is the intention to continue working as a teacher (measured in years). Various predictors are included at both the teacher and school levels. Data is drawn from the OECD's TALIS 2018 survey, involving 48,730 teachers across 3,128 schools in 20 countries. Findings indicate crucial roles of individual factors like gender, age, education, contract type, motivation, professional development, self-efficacy, and workplace well-being in retention. Additionally, school factors like climate, special needs student ratio, and location are impactful. Conversely, school size, delinquency rate, and certain stress factors are not significant in predicting retention.

E0309: **A Bayesian approach for chronic hepatitis B prevalence estimation to improve the accuracy of economic evaluation**
Presenter: **William WL Wong**, University of Waterloo, Canada
Co-authors: Julien Smith-Roberge
Chronic hepatitis B (CHB) is usually a silent disease. The asymptomatic nature means that the disease often remains undiagnosed, leaving its prevalence highly uncertain. This generates significant uncertainty for the associated economic evaluations. The objective is to establish a mathematical framework for the estimation of CHB prevalence and the undiagnosed proportion. A state-transition model describing infection, disease progression and treatment response was mathematically formulated and developed. Model parameters were obtained from the published literature. The historical prevalence of CHB is estimated through a calibration process based on a Bayesian MCMC algorithm. The algorithm constructed posterior distributions of the historical prevalence of CHB by comparing the model-generated predictions of the annual numbers

of health events related to CHB against the observed numbers. The prevalence of CHB in Canada in 2018 was estimated to be 0.28%, and the percentage of undiagnosed among the total infected was 31.5%. The results are in line with a recently conducted seroprevalence survey. Prevalence estimates impact economic evaluation results on interventions with respect to CHB interventions. Considering the rapid development of interventions for CHB, updated prevalence estimates will become necessary. A platform is provided to estimate this information in a robust and efficient way.

E0262: **Empirical analysis of crude oil dynamics using affine vs. non-affine jump-diffusion models**
Presenter: **Katja Ignatieva**, University of New South Wales Sydney, Australia

The dynamics of the US oil ETF (USO) are examined through stochastic volatility (SV) models across three classes: SV with jumps in both returns and volatility (SVCJ), SV with jumps in returns only (SVJ), and pure SV without jumps. Eighteen models, including affine and non-affine variations, are evaluated using particle Markov chain Monte Carlo methods. The analysis employs the deviance information criterion (DIC), Bayes factors, probability plots, and deviation measures against the crude oil ETF volatility index (OVX) and realized volatility (RV) benchmarks. Findings highlight SVCJ models, especially SVCJ-PLY-0.5 and SVCJ-PLY-1.0, as most effective in capturing USO dynamics, outperforming standard SV models. The SVCJ-PLY-0.5 model is notably superior based on DIC and Bayes factors, closely aligning estimated volatility with OVX and RV. Statistical criteria favor jump models, with affine models SVJ-LIN-0.5 and SVCJ-LIN-0.5 showing particular promise for finance theory, ranking high among tested frameworks. The predictive accuracy of the evaluated models is underscored for forecasting future volatility trends.

E0231: **Sensitivity analysis with balancing weights estimators to address informative visit times in irregular longitudinal data**
Presenter: **Li Su**, University of Cambridge, United Kingdom
Co-authors: Sean Yiu

Irregular longitudinal data with informative visit times arise when patients' visits are partly driven by concurrent disease outcomes. To address the selection bias from such data, existing methods rely on unverifiable assumptions and haven't adequately accommodated informative visit times for marginal regression analyses. Based on novel balancing weights estimators, a new sensitivity analysis approach is proposed to address informative visit times. The balancing weights are obtained by balancing observed covariate distributions and including a selection function with specified sensitivity parameters to characterize the additional influence of the concurrent outcome on the visit process. A calibration procedure is proposed to anchor the magnitude of the sensitivity parameters to the amount of variation in the visit process that could be additionally explained by the concurrent outcome given the observed history. Simulations demonstrate that the balancing weights estimators outperform existing weighted estimators for robustness and efficiency. The proposed methods are applied to analyze data from a clinic-based cohort of psoriatic arthritis.

E1042: **Inference on tree-structured subgroups with subgroup size and subgroup effect relationship in clinical trials**
Presenter: **Yuanhui Luo**, The Hong Kong University of Science and Technology, Hong Kong
Co-authors: Xinzhou Guo

When multiple candidate subgroups are considered in clinical trials, it is often necessary to make statistical inferences on the subgroups simultaneously. Classical multiple testing procedures might not lead to an interpretable and efficient inference on the subgroups as they often fail to take the subgroup size and subgroup effect relationship into account. Built on the selective traversed accumulation rules (STAR), a data-adaptive and interactive multiple-testing procedure is proposed for subgroups, which can take subgroup size and subgroup effect relationship into account under a prespecified tree structure. The proposed method is easy to implement and can lead to a more interpretable and efficient inference on prespecified tree-structured subgroups. The merit of the proposed method is demonstrated by re-analyzing the panitumumab trial with the proposed method.

Friday 19.07.2024 08:15 - 09:55 Parallel Session K – EcoSta2024

EV284 Room 406 STATISTICAL AND FINANCIAL RESEARCH (VIRTUAL) Chair: Wenbo Wu

E1024: **IANOVA: Multi-sample means comparisons for imprecise interval data**
Presenter: **Yan Sun**, Utah State University, United States
Co-authors: Zac Rios, Brennan Bean
Interval data has become an increasingly popular tool for solving modern data problems. Intervals are now often used for dimension reduction, data aggregation, privacy censorship, and quantifying awareness of various uncertainties. Among many statistical methods being developed for interval data, the significance test is of particular importance due to its fundamental value both in theory and practice. The difficulty in developing such tests lies mainly in the fact that the concept of normality does not extend naturally to intervals, making the exact tests hard to formulate. As a result, most existing works have relied on bootstrap methods to approximate null distributions. This is not always feasible, considering limited sample sizes or other intrinsic data characteristics in practice. A novel test is proposed for comparing multi-sample means with interval data as a generalization to the classical ANOVA. Based on the random sets theory, the test statistic is constructed analogously to the F statistic for the classical ANOVA. The limiting null distribution is derived under usual assumptions and mild regularity conditions. Simulation studies with various data configurations validate the asymptotic results. Finally, a real interval data ANOVA analysis is presented that showcases the applicability of this new method.

E0925: **Comparative study on excess distribution estimators in iid settings**
Presenter: **Taku Moriyama**, Yokohama City University, Japan
The focus is on the excess distribution estimators in iid settings. There are two ways to estimate: the fit to the generalized Pareto distribution and the fully nonparametric estimation. The fitting estimator is justified by the approximation proven in the extreme value theory; however, the accuracy depends on how extremely large the target is. The nonparametric estimator does not need the approximation and has the advantage of wide applicability. Both theoretical and numerical comparative studies are conducted on excess distribution estimation. Asymptotic convergence rates of two estimators are obtained, and the mean integrated squared errors are numerically surveyed by simulation study. Further details are presented.

E1057: **Exploring the linkages between ESG ETFs and ESG indices: A weighting analysis**
Presenter: **Tsung-Han Ke**, National Chi Nan University, Taiwan
Co-authors: Hung-Chun Huang, Hsin-Yu Shih , Jing-Yuan Zhang, Hsuan Chu Huang
Environmental, social, and governance (ESG) investing has gained significant traction in recent years, driven by growing investor interest in sustainable and responsible investment strategies. ESG ETFs have emerged as a popular vehicle for ESG investing, offering investors diversified exposure to companies with strong ESG practices. However, the relationship between ESG ETFs and ESG indices remains an area of active research. The aim is to investigate the linkages between ESG ETFs and ESG indices, with a particular focus on the weighting of ESG factors. A quantitative approach is employed, utilizing data on ESG ETFs and ESG indices from various sources. The analysis will involve examining the correlation between ESG ETF returns and ESG index performance, as well as assessing the impact of different ESG factor weightings on ESG ETF performance. The relationship between ESG indices and global semiconductor innovation momentum is also investigated. The findings provide valuable insights into the potential influence of ESG considerations on corporate innovation strategies in the semiconductor industry.

EO149 Room 102 FURTHER DEVELOPMENTS IN FINANCIAL MODELLING (VIRTUAL) Chair: Rogemar Mamon

E1062: **Would a two-benchmark regime be better?**
Presenter: **Chengguo Weng**, University of Waterloo, Canada
After the manipulation scandal during the global financial crisis, LIBOR was phased out, and regulators around the world have since transitioned to risk-free reference rates (RFRs). In the United States, the secured overnight financing rate (SOFR), which is an RFR based on overnight repo transactions, has been designated as the sole replacement for LIBOR. Meanwhile, Europe and Japan have chosen to establish a two-benchmark regime consisting of an RFR and a LIBOR-like credit-sensitive reference rate (CSR). A two-agent model is considered with one representative firm and one representative bank in the economy, and a tractable model is used to compare the single RFR regime against the two-benchmark regime with an RFR and a CSR. It turns out that adding a new credit-sensitive benchmark in addition to the existing risk-free benchmark always improves total welfare but not necessarily for both the bank and the firm in some scenarios. The study also indicates that credit supply is higher and borrowing cost is lower in the two-benchmark system than would be the case for the single RFR system. Furthermore, RFR could be driven out of the market when the correlation between CSR and the banks' funding costs is strong enough. The impacts of fixed-rate lending and interest-rate swap trading are also discussed.

E1068: **The impact of intermediaries on insurance demand and pricing**
Presenter: **Yixing Zhao**, Guangdong University of Foreign Studies, China
The purpose is to study the impact of an independent insurance intermediary, who holds a fiduciary duty to an unsophisticated insurance buyer, on insurance demand and pricing. The intermediary adopts two remuneration systems, namely, a fee-for-advice system and a commission system. Insurance contracting between the buyer (represented by the intermediary) and the insurer is formulated as a Stackelberg insurance game. The analysis yields closed-form expressions for the buyer's equilibrium indemnity and the insurer's equilibrium premium loading. Subsequently, the intricate effects of fiduciary duty and remuneration arrangements are comprehensively explored on equilibrium strategies and stakeholder welfare, unraveling several critical economic implications. It is found that the phenomenon of over-insuring at high premiums, a robust empirical observation, is also attributed to the deterioration of fiduciary duty. Additionally, compared to direct contracting, the presence of the intermediary reduces the equilibrium indemnity but may increase or reduce the equilibrium premium loading. Findings also highlight the absence of a remuneration system that benefits all stakeholders, confirming the coexistence of both systems.

E1095: **Solution of fixed and free boundary problems for financial derivatives: An embedding approach**
Presenter: **Mariano Rodrigo**, University of Wollongong, Australia
In several recent articles, the embedding approach was used to analytically and numerically solve fixed and free boundary problems for single-layer and multilayer problems for the heat equation. The purpose is to examine how the embedding approach can be adapted for similar problems arising in the pricing of financial derivatives.

EO314 Room 103 RECENT ADVANCES IN GENOMICS AND METAGENOMICS DATA ANALYSIS Chair: Huijuan Zhou

E0221: **Multigroup analysis of compositions of microbiomes with covariate adjustments and repeated measures**
Presenter: **Huang Lin**, University of Maryland, United States

Microbiome differential abundance analysis for two-group comparisons is extensively documented in existing literature. However, many microbiome studies encompass more than two groups, often including ordered categories like disease stages, necessitating varied comparison approaches. Traditional pairwise comparisons fall short in terms of statistical power and control of false discovery rates. Furthermore, many studies involve repeated measures from the same participants, as seen in longitudinal microbiome research, yet there is a notable gap in the literature for effectively addressing these complex scenarios. To bridge this gap, ANCOM-BC2 is introduced, a general framework designed for multigroup analyses, accommodating covariate adjustments and repeated measures. ANCOM-BC2 has proven its efficacy in enhancing power and reducing false discovery rates when compared to competing methods in simulation studies. The methodology is also exemplified through two real-world datasets: one investigating the impact of aridity on soil microbiomes and the other examining the microbiome alterations in patients with inflammatory bowel disease post-surgical interventions.

E0223: **mPower: A real data-based power analysis tool for microbiome study design**
Presenter: **Jun Chen**, Mayo Clinic, United States

Power analysis is a critical step in designing a microbiome study. Previous power calculation tools mainly rely on parametric models, which underestimate the complexity of microbiome data and could produce overly optimistic power estimates. A simulation-based power analysis tool, mPower, is presented to facilitate realistic power calculation for microbiome studies. The tool uses a real data-based semi-parametric simulation framework to generate realistic microbiome data, upon which the power assessment is performed. Coupled with the recently developed differential analysis tool, LinDA, the power tool supports different study designs, including cross-sectional, case-control, and longitudinal studies, with or without confounders. It allows power analysis for both community-level and taxa-level testing. By using a database of large microbiome datasets from different environments, the users could perform power calculations based on the environment of interest. The application of the power analysis tool is showcased using several real examples.

E0600: **mbDecoda: A debiased approach to compositional data analysis for microbiome surveys**
Presenter: **Tao Wang**, Shanghai Jiao Tong University, China

Potentially pathogenic or probiotic microbes can be identified by comparing their abundance levels between healthy and diseased populations or, more broadly, by linking microbiome composition with clinical phenotypes or environmental factors. However, in microbiome studies, feature tables provide relative rather than absolute abundance of each feature in each sample. Moreover, microbiome abundance data are count-valued, often over-dispersed, and contain a substantial proportion of zeros. To carry out differential abundance analysis while addressing these challenges, mbDecoda is introduced, a model-based approach for debiased analysis of sparse compositions of microbiomes. mbDecoda employs a zero-inflated negative binomial model, linking mean abundance to the variable of interest through a log link function, and it accommodates the adjustment for confounding factors. An expectation-maximization algorithm is developed to efficiently obtain maximum likelihood estimates of model parameters. A minimum coverage interval approach is then proposed to rectify compositional bias, enabling accurate and reliable absolute abundance analysis. Through extensive simulation studies and analysis of real-world microbiome datasets, it is demonstrated that mbDecoda compares favorably to state-of-the-art methods in terms of effectiveness, robustness, and reproducibility.

E0715: **Differential inference for single-cell RNA-sequencing data**
Presenter: **Yingying Wei**, The Chinese University of Hong Kong, Hong Kong
Co-authors: Fangda Song, Kevin Y Yip, Yingying Wei

Single-cell RNA-seq (scRNA-seq) experiments are becoming more and more complicated with multiple treatment or biological conditions. However, guidelines on experimental designs and rigorous statistical methods for a comparative scRNA-seq study with data collected from multiple conditions are still lacking. Existing multi-stage approaches to identifying differential cell-type abundance and differentially expressed genes between conditions suffer from high error rates because multi-stage approaches ignore uncertainties in previous stages and propagate errors in earlier stages to later stages. DIFseq, a Bayesian hierarchical model, is introduced to rigorously quantify the condition effects on both cellular compositions and cell-type-specific gene expression levels for scRNA-seq data. DIFseq substantially outperforms state-of-the-art methods in terms of the accuracy of cell type clustering, differential abundance, and differential expression inference for both simulated and real data. Moreover, to the best of knowledge, the conditions are derived for a valid design for a comparative scRNA-seq study for the first time.

EO216 Room 104 NEW STREAMS IN STATISTICS FOR STOCHASTIC PROCESSES (VIRTUAL) Chair: Yuta Koike

E0568: **Two step estimations via the Dantzig selector for ergodic time series models**
Presenter: **Kou Fujimori**, Shinshu University, Japan
Co-authors: Koji Tsukuda

The estimation problems for conditionally heteroskedastic ergodic time series models are considered with high-dimensional and sparse parameters with some nuisance parameters. The asymptotic behavior of the Dantzig selector is first established based on the least squares method to estimate the unknown parameter of interest. Then, using the Dantzig selector and a consistent estimator for the nuisance parameter, the asymptotically normal estimator is constructed for the non-zero components of the parameter of interest. Applications to order selection problems for integer-valued autoregressive models of large order are presented.

E0576: **Quasi-maximum likelihood estimation and penalized estimation under non-standard conditions**
Presenter: **Junichiro Yoshida**, University of Tokyo, Graduate School of Mathematical Sciences, Japan
Co-authors: Nakahiro Yoshida

A general parametric estimation theory is suggested that allows the derivation of the limit distribution of estimators under the following two non-standard conditions: (i) The true parameter value may lie on the boundary of the parameter space, and (ii) even identifiability fails. For Singularity (i), the local form of the parameter space is sought around the true value lying on the boundary in the framework of the local asymptotic theory established by Ibragimov and Khasminskii. This approach can handle some complex examples that previous studies cannot under quasi-maximum likelihood estimation (in the ergodic or non-ergodic statistics, and with or without penalization). An example is penalized maximum likelihood estimation of variance components of random effects in linear mixed models. For Singularity (ii), penalized estimation is used to stabilize the asymptotic behavior of the estimator by forcing it to converge to the most parsimonious of all the true values. This estimator can show the oracle property even in singular models where other estimation methods for model selection, such as likelihood ratio tests, seem complicated. An example is a superposition of parametric proportional hazard models.

E0598: **Modeling lead-lag effects using bivariate Neyman-Scott processes with gamma kernels**
Presenter: **Takaaki Shiotani**, Graduate School of Mathematical Sciences, The University of Tokyo, Japan
The lead-lag effect refers to the correlation that occurs with a time difference between 2 time series data. It has been explored in financial engineering because it can be applied to statistical arbitrage or understanding market structures. Broadly, there are two approaches: one focusing on the correlation structure of prices and the other on the correlation structure of transaction/order times. The latter approach, which is thought to be robust against microstructure noise, is considered. An intuitively understandable model is proposed utilizing a multivariate Neyman-Scott point process model equipped with gamma kernels to model the lead-lag effect. Furthermore, an efficient computational method is developed for parameter estimation through quasi-likelihood. Numerical experiments are conducted to check the proposed estimator's performance. In addition, high-frequency Japanese stock data is analyzed to verify that the model effectively explains the correlation between pairs of stocks. Research provides a new point process model for lead-lag estimation that is highly interpretable and usable within a realistic computational timeframe. Additionally, the method can be applied to phenomena beyond financial data, such as social media and earthquake data.

E0637: **Predictive model selection for jump diffusion models**
Presenter: **Yuma Uehara**, Kansai University, Japan
A model selection problem is considered for jump-diffusion models based on high-frequency samples. The terminal time is supposed to diverge (ergodic setting), and the interest is to select drift and diffusion coefficients and jump distribution among candidates. An explicit AIC-type information criterion is proposed based on the threshold quasi-likelihood. Unlike the diffusion case, when the jump term is parametrized in some way, the stochastic flow approach cannot be directly used to get the transition density estimates, which is essential to evaluate the bias. To validate such an approximation, new transition density estimates are presented.

EO316 Room 105 ADVANCES IN COMPUTATIONAL METHODS FOR BAYESIAN STATISTICS Chair: Quan Zhou

E0855: **Asynchronous and distributed data augmentation for massive data settings**
Presenter: **Kshitij Khare**, University of Florida, United States
Data augmentation (DA) algorithms are slow in massive data settings due to multiple passes through the entire data. This problem is addressed by developing a DA extension that exploits asynchronous and distributed computing. The extended DA algorithm is called asynchronous and distributed (AD)DA, with the original DA as its parent. Any ADDA is indexed by a parameter r in (0,1) and starts by dividing the entire data into k disjoint subsets and storing them on k processes. Every iteration of ADDA augments only an r-fraction of the k data subsets with some positive probability and leaves the remaining (1-r)-fraction of the augmented data unchanged. The parameter draws are obtained using the r-fraction of new and (1-r)-fraction of old augmented data. It is shown that the ADDA Markov chain is Harris ergodic with the desired stationary distribution under mild conditions on the parent DA algorithm. ADDA is demonstrated to be significantly faster than its parent for many (k, r) choices in three representative models. The geometric ergodicity of the ADDA Markov chain is also established for all three models, which yields asymptotically valid standard errors for estimates of desired posterior quantities.

E0827: **Adaptivity of diffusion models to manifold structures**
Presenter: **Yun Yang**, University of Illinois Urbana-Champaign, United States
Co-authors: Rong Tang
Empirical studies have demonstrated the effectiveness of (score-based) diffusion models in generating high-dimensional data, such as texts and images, which typically exhibit a low-dimensional manifold nature. These empirical successes raise the theoretical question of whether score-based diffusion models can optimally adapt to low-dimensional manifold structures. While recent work has validated the minimax optimality of diffusion models when the target distribution admits a smooth density with respect to the Lebesgue measure of the ambient data space, these findings do not fully account for the ability of diffusion models to avoid the curse of dimensionality when estimating high-dimensional distributions. The aim is to consider two common classes of diffusion models: Langevin diffusion and forward-backwards diffusion. Both models can adapt to the intrinsic manifold structure by showing that the convergence rate of the inducing distribution estimator depends only on the intrinsic dimension of the data. Moreover, the considered estimator does not require knowing or explicitly estimating the manifold. It is also demonstrated that the forward-backwards diffusion can achieve the minimax optimal rate under the Wasserstein metric when the target distribution possesses a smooth density with respect to the volume measure of the low-dimensional manifold.

E0662: **Parallel and sequential coordinate ascent in variational inference**
Presenter: **Debdeep Pati**, Texas A&M University, United States
A surprising discordance is described between the sequential and parallel versions of coordinate ascent in variational inference. Focusing on the case of high dimensional linear regression, it is shown that the parallel version exhibits a lack of convergence under a general setting. This can be effectively remedied by using a sequential version of the algorithm under fairly relaxed assumptions. The techniques involve analyzing the spectral norm of the Jacobian of specific nonlinear dynamical systems.

E0547: **Geometric ergodicity of trans-dimensional Markov chain Monte Carlo algorithms**
Presenter: **Qian Qin**, University of Minnesota, United States
The convergence properties of trans-dimensional MCMC algorithms are studied(e.g., the reversible jump algorithm) when the total number of models is finite. It is shown that, for reversible and some non-reversible trans-dimensional Markov chains, under mild conditions, geometric convergence is guaranteed if the Markov chains associated with the within-model moves are geometrically ergodic. This result is proved in an L^2 framework using the technique of Markov chain decomposition. While the technique was previously developed for reversible chains, this is extended to the point that it can be applied to some commonly used non-reversible chains. Under geometric convergence, a central limit theorem holds for ergodic averages, even in the absence of Harris ergodicity. This allows for the construction of simultaneous confidence intervals for features of the target distribution.

EO087 Room 106 ADVANCES IN INFERENCE FOR HIGH-DIMENSIONAL AND CLUSTERED DATA (VIRTUAL) Chair: Chenlu Ke

E0630: **Adaptive nonparametric two-sample test for high-dimensional data**
Presenter: **Zi Ye**, Lehigh University, United States
An adaptive nonparametric two-sample test of high-dimensional data is proposed. No distributional assumption, moment condition, or parametric model is required to develop the tests and their theories.

E0821: **Estimating marginal association in clustered data with informative subgroups induced by a given covariate**
Presenter: **Samuel Anyaso-Samuel**, National Cancer Institute, United States
Informative cluster size (ICS) typically introduces bias in cluster-correlated data analyses. A complex form of informativeness where the number of observations corresponding to latent levels of a unit-level continuous covariate is studied within a cluster is associated with the response variable. This type of informativeness has not been explored in prior research. A novel test statistic designed to evaluate the effect of the covariate while

accounting for informativeness is presented. The covariate induces a continuum of latent subgroups within the clusters, and the test statistic is formulated by aggregating values from an established statistic that accounts for informative subgroup sizes when comparing group-specific marginal distributions. Through simulations, the test is compared with four traditional methods commonly employed in cluster-correlated data analyses. Only the test maintains the size across all data-generating scenarios with informativeness. The proposed method is illustrated to test for marginal associations in periodontal data with this distinctive form of informativeness.

E0828: **A Bayesian multiple testing procedure infused with historical data, with application in gene expression testing**
Presenter: **Ya Su**, Virginia Commonwealth University, United States
The process of identification of expressed genes between experimental groups is a difficult task due to the heteroscedasticity across a massive number of genes. This problem is addressed using a Bayesian framework, which facilitates the incorporation of prior knowledge obtained from different platforms or organisms. A new test statistic and sign-adjusted FDR that emphasizes information regarding the direction of the differentially expressed genes are proposed. The statistic is proven to achieve the highest count of true positives compared to all legitimate sign-adjusted false positive controlled methods. Simulation results provide numerical evidence. The approach is demonstrated through the analysis of two gene expression datasets.

EO322 Room 108 CUTTING-EDGE STATISTICAL METHODS FOR MODERN BIOMEDICAL PROBLEMS Chair: Yi Li

E0227: **Meta-analysis by integrating multiple observational studies with multivariate outcomes**
Presenter: **Yi Li**, University of Michigan, United States
Integrating multiple observational studies to make unconfounded causal or descriptive comparisons of group potential outcomes in a large natural population is challenging. Moreover, retrospective cohorts, being convenience samples, are usually unrepresentative of the natural population of interest and have groups with unbalanced covariates. A general covariate-balancing framework is proposed based on pseudo-populations that extend established weighting methods to the meta-analysis of multiple retrospective cohorts with multiple groups. Additionally, by maximizing the effective sample sizes of the cohorts, a FLEXible, Optimized, and Realistic (FLEXOR) weighting method is proposed, appropriate for integrative analyses. New weighted estimators are developed for unconfounded inferences on wide-ranging population-level features and estimands relevant to group comparisons of quantitative, categorical, or multivariate outcomes. The asymptotic properties of these estimators are examined, and accurate small-sample procedures are devised for quantifying estimation uncertainty. Through simulation studies and meta-analyses of TCGA datasets, the differential biomarker patterns of the two major breast cancer subtypes in the United States are discovered, and the versatility and reliability of the proposed weighting strategy are demonstrated, especially for the FLEXOR pseudo-population.

E0229: **Variable selection for interval-censored failure time data**
Presenter: **Jianguo Sun**, University of Missouri, United States
Interval-censored failure time data occur in many areas, including demographical studies, economic studies, medical studies and social sciences, and in different forms. The focus is on the variable selection for such data and present some recently developed tools.

E0247: **Post-estimation strategies in high-dimensional data analytics**
Presenter: **Ejaz Ahmed**, Brock, Canada
The rapid growth in the size and scope of data sets in a host of disciplines has created a need for innovative statistical strategies to understand such data. A variety of statistical and machine learning is needed to reveal the hidden data story. Complex big data analysis is a very challenging but rewarding research area as data sets include a larger number of features, data contamination, unstructured patterns, and so on. A host of models are now data-driven with a large number of predictors, namely high-dimensional data (HDD). For HDD analysis, many penalized methods were introduced for simultaneous variable selection and parameter estimation when the model is sparse. However, a model may have sparse signals as well as a number of predictors with weak signals. In this scenario, variable selection methods may not distinguish predictors from weak signals and sparse signals. For this reason, a high-dimensional shrinkage strategy is proposed to improve the prediction performance of a submodel. The proposed high-dimensional shrinkage strategy is demonstrated to perform uniformly better than the penalized and machine learning methods in many cases. The relative performance of the proposed HDSE strategy is appraised by both simulation studies and real data analysis. Some open research problems are discussed as well.

E0696: **Inference on potentially identified subgroups in clinical trials**
Presenter: **Xinzhou Guo**, The Hong Kong University of Science and Technology, Hong Kong
When subgroup analyses are conducted in clinical trials with moderate or high dimensional covariates, candidate subgroups often need to be identified from the data, and the potentially identified subgroups are evaluated in a replicable way. The classical statistical inference applied to the potentially identified subgroups, assuming the subgroups are the same as what is observed from the data, might suffer from bias issues when the regularity assumption that the boundaries of the subgroups are negligible is violated. A shift-based method is proposed to address the nonregularity bias issue and combine it with cross-fitting and subsampling to develop a de-biased inference procedure for potentially identified subgroups. The proposed method is model-free and asymptotically efficient whenever it is possible and can be viewed as an asymmetric smoothing approach. The merits of the proposed method are demonstrated by re-analyzing the ACTG 175 trial.

E1120: **Bootstrap cross-validation estimate**
Presenter: **Lu Tian**, Stanford University, United States
Cross-validation is a widely used technique for evaluating the performance of prediction models. It helps avoid the optimism bias in error estimates, which can be significant for models built using complex statistical learning algorithms. However, since the cross-validation estimate is a random value dependent on observed data, it is essential to accurately quantify the uncertainty associated with the estimate. This is especially important when comparing the performance of two models using cross-validation, as one must determine whether differences in error estimates are a result of chance fluctuations. Although various methods have been developed for making inferences on cross-validation estimates, they often have many limitations, such as stringent model assumptions. A fast bootstrap method is proposed that quickly estimates the standard error of the cross-validation estimate and produces valid confidence intervals for a population parameter measuring average model performance. Our method overcomes the computational challenge inherent in bootstrapping the cross-validation estimate by estimating the variance component within a random effects model. It is just as flexible as the cross-validation procedure itself. To showcase the effectiveness of our approach, we employ comprehensive simulations and real data analysis across three diverse applications.

EO050 Room 109 NEW ADVANCES IN BIOMEDICAL RESEARCH WITH APPLICATIONS TO HEALTH DATA Chair: Yichuan Zhao

E0316: **Transfer learning under the Cox model with interval-censored data**
Presenter: **Mengqi Xie**, Capital Normal University, China
Co-authors: Tao Hu, Jie Zhou
Transfer learning, focusing on information borrowing to address limited sample size issues, has gained increasing attention in recent years. The method aims to utilize data from other population groups as a complement to enhance risk factor discernment and failure time prediction among underrepresented subgroups. However, a literature gap exists in effective knowledge transfer from the source to the target for risk assessment with interval-censored data while accommodating population incomparability and privacy constraints. The objective is to bridge this gap by developing a transfer learning approach under the Cox proportional hazards model. The tuning-free Trans-Cox-MIC algorithm is introduced, enabling adaptable information sharing in regression coefficients and baseline hazards while ensuring computational efficiency. Extensive simulations showcase the method's accuracy, robustness, and efficiency. Application to the prostate cancer screening data demonstrates enhanced risk estimation precision and predictive performance in the African American population.

E0408: **A class of variable selection methods for (partial linear) varying coefficient EV models with longitudinal data**
Presenter: **Mingtao Zhao**, Anhui University of Fianance and Economics, China
A class of variable selection methods are proposed for (partial linear) varying coefficient EV models with longitudinal data based on bias-corrected method and quadratic inference functions. The proposed method is called the bias-corrected penalized quadratic inference functions method. Under some regularity conditions, some asymptotic properties of the proposed method can be established. Numerical results show that it is superior to other methods of the same type.

E0688: **Time-dependent ROC curve for multiple longitudinal biomarkers and its application in diagnosing cardiovascular events**
Presenter: **Lizhe Sun**, Shanxi University of Finance and Economics, China
Since it can help people detect the early signs of diseases, accurate diagnostic techniques based on biomarkers are crucial in biomedical research. A novel bivariate time-varying coefficients Logistic regression model for addressing the combination of longitudinal biomarkers is proposed. By using the B-spline method to estimate the proposed model, multiple longitudinal biomarkers can be effectively combined, and the accuracy of disease onset prediction is improved. It is shown that the proposed method is theoretically consistent. It exhibits superior performance compared to the existing method, as presented through numerical results. The proposed method is verified in a study on predicting the probability of onset of future cardiovascular events for type 2 diabetic patients by combining longitudinal biomarkers: HbA1c and LDL-C. The combination of longitudinal biomarkers has been demonstrated to significantly improve disease diagnosis compared to a combination of updated biomarkers.

EO223 Room 110 DESIGN AND SUBSAMPLING FOR MASSIVE DATA Chair: HaiYing Wang

E0160: **Subsampling strategies for heavily censored reliability big data**
Presenter: **Qingpei Hu**, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, China
While subsampling techniques have been extensively developed in the literature to downsize the data volume, there is a notable gap in addressing the unique challenge of handling extensive reliability data, in which a common situation is that a large proportion of data is censored. A subsampling method is proposed for reliability analysis in the presence of censoring data to estimate the parameters of lifetime distribution effectively and efficiently. The asymptotic properties of the subsampling-based estimators are given, and the optimal subsampling probabilities are derived by minimizing the criterion defined by the trace of the asymptotic covariance matrix. Efficient algorithms are proposed to implement the proposed subsampling methods to address the challenge of the optimal subsampling strategy depending on unknown parameter estimation from full data. Numerical studies and a real-world dataset are employed to illustrate the performance of the proposed methods. The results demonstrate the superior performance of the proposed methods compared to uniform subsampling and the effective alleviation of the computational burden compared with the computational time of full data analysis.

E0531: **Optimal subsampling for large-scale linear classification**
Presenter: **Jun Yu**, Beijing Institute of Technology, China
Subsampling is one of the popular methods to balance statistical efficiency and computational efficiency in the big data era. A new optimal subsampling framework is presented for linear classifiers based on a piecewise linear quadratic loss. The classifier not only aims to select an informative subset of the training sample to reduce data size but also embeds some summary statistics to maintain high accuracy. A novel view of hinge loss-based classifiers under the general subsampling framework will be presented with rigorous, proven statistical properties. Numerical results demonstrate that our classifiers outperform the existing methods in terms of estimation, computation, and prediction.

E0542: **Distributed logistic regression for massive data with rare events**
Presenter: **Xuetong Li**, Peking University, China
Co-authors: Xuening Zhu, Hansheng Wang
Large-scale rare event data are commonly encountered in practice. To tackle the massive rare events data, a novel distributed estimation method is proposed for logistic regression in a distributed system. For a distributed framework, the following two challenges are faced. The first challenge is how to distribute the data. In this regard, two different distribution strategies (i.e., the random strategy and the copy strategy) are investigated. The second challenge is how to select an appropriate type of log-likelihood function so that the best asymptotic efficiency can be achieved. Then, the under-sampled (US) and inverse probability weighted (IPW) types of objective functions are considered. The results suggest that the copy strategy, together with the IPW objective function, is the best solution for distributed logistic regression with rare events. The finite sample performance of the distributed methods is demonstrated by simulation studies and a real-world Swedish Traffic Sign dataset.

E0880: **Individual and interactive constrained online selection**
Presenter: **Changliang Zou**, Nankai University, China
Real-time decision-making gets more attention in the big data era. The focus is on the problem of sample selection in the online setting, where one encounters a possibly infinite sequence of individuals collected over time with covariate information available. The goal is to select samples of interest that are characterized by their unobserved responses until the user-specified stopping time. A new decision rule enables finding more preferable samples that meet practical requirements by simultaneously controlling two types of general constraints: individual and interactive constraints, which include the widely utilized false selection rate (FSR), cost limitations, and diversity of selected samples. The key elements of the approach involve quantifying the uncertainty of response predictions via predictive inference and addressing individual and interactive constraints in a sequential manner. Theoretical and numerical results demonstrate the effectiveness of the proposed method in controlling both individual and interactive constraints.

EO083 Room 111 RECENT ADVANCES IN STATISTICAL LEARNING IN GENETICS AND GENOMICS Chair: Tianying Wang

E0166: **Nonparametric composition-on-composition regression analysis for high dimensional microbiome data**
Presenter: **Xiang Zhan**, Peking University, China

High-dimensional compositional data are frequently encountered nowadays in scientific research of many disciplines, such as microbiome research. Regression analysis with compositional data being either responses or predictors has been well studied. However, when both responses and predictors are compositional, the inventory of analysis tools is surprisingly limited. Among the few existing methods, most of them rely on a log-ratio transformation to move compositional data analysis from simplex to real. Yet, a serious weakness of these methods is the failure to handle the substantial fraction of zeroes observed in microbiome data. To investigate associations between multiple high-dimensional microbial compositions, a nonparametric composition-on-composition (NCOC) regression analysis method is proposed, which does not require log-ratio transformations and hence can handle zeroes in the data. To account for high dimensionality, regression coefficients are estimated using a penalized estimation equation approach to improve its accuracy. Finally, statistical inference procedures are proposed to quantify uncertainty in the model predictions. The superior performance of NCOC and the validity and potential usefulness of the inference procedures are demonstrated through comprehensive numerical simulation studies, real data applications, and case studies.

E0375: **Ensemble methods for testing a global null**
Presenter: **Yaowu Liu**, Southwestern University of Finance and Economics, China

Testing a global null is a canonical problem in statistics and has a wide range of applications. In view of the fact that no uniformly most powerful test exists, prior and/or domain knowledge is commonly used to focus on a certain class of alternatives to improve the testing power. However, it is generally challenging to develop tests that are particularly powerful against a certain class of alternatives. Motivated by the success of ensemble learning methods for prediction or classification, an ensemble framework is proposed for testing that mimics the spirit of random forests to deal with the challenges. The ensemble testing framework aggregates a collection of weak base tests to form a final ensemble test that maintains strong and robust power. The framework for four problems is applied to global testing in different classes of alternatives arising from whole genome sequencing (WGS) association studies. Specific ensemble tests are proposed for each of these problems, and their theoretical optimality is established in terms of Bahadur efficiency. Extensive simulations and analysis of a real WGS dataset are conducted to demonstrate the type I error control and/or power gain of the proposed ensemble tests.

E0701: **An integrative multi-context Mendelian randomization method for identifying risk genes across human tissues**
Presenter: **Fan Yang**, Tsinghua University, China

Mendelian randomization (MR) provides valuable assessments of the causal effect of exposure on outcome, yet the application of conventional MR methods for mapping risk genes encounters new challenges. One of the issues is the limited availability of expression quantitative trait loci (eQTLs) as instrumental variables (IVs), hampering the estimation of sparse causal effects. Additionally, the often context/tissue-specific eQTL effects challenge the MR assumption of consistent IV effects across eQTL and GWAS data. To address these challenges, a multi-context multivariable integrative MR framework, mintMR, is proposed for mapping expression and molecular traits as joint exposures. It models the effects of molecular exposures across multiple tissues in each gene region while simultaneously estimating across multiple gene regions. A major innovation of mintMR involves employing multi-view learning methods to collectively model latent indicators of disease relevance across multiple tissues, molecular traits, and gene regions. The multi-view learning captures the major patterns of disease relevance and uses these patterns to update the estimated tissue relevance probabilities. mintMR is applied to evaluate the causal effects of gene expression and DNA methylation for 35 complex traits using multi-tissue QTLs as IVs. The proposed mintMR controls genome-wide inflation and offers new insights into disease mechanisms.

E0825: **Optimizing sample size for statistical learning with bulk transcriptomic sequencing: A learning curve approach**
Presenter: **Li-Xuan Qin**, Memorial Sloan Kettering Cancer Center, United States
Co-authors: Yunhui Qi, Xinyi Wang

Accurate sample classification using transcriptomics data is crucial for personalized medicine. The success of such endeavors depends on determining a suitable sample size and ensuring adequate statistical power without unnecessary resource allocation or ethical concerns. Current sample size calculation methods for sample classification rely on assumptions and algorithms that may not align with modern machine and deep learning techniques. The methodological gap is addressed by developing computational approaches to determine the required number of samples for accurate classification in transcriptomics studies using statistical learning. The approach establishes the power-versus-sample-size relationship by employing a data augmentation strategy followed by fitting a learning curve. Its performance is evaluated for both microRNA and RNA sequencing using data from the Cancer Genome Atlas, considering various data characteristics (such as sample size, marker filtering, and sequencing depth normalization) and algorithm configurations (including model selection, hyperparameter tuning, and offline augmentation), based on a range of evaluation metrics. Python and R code for implementation of the proposed approach is freely available on GitHub. The adoption of statistical learning in biomedical transcriptomics studies is expected to advance and accelerate their translation into clinically useful classifiers.

EO185 Room 212 STATISTICAL MODELING AND INFERENCE FOR LARGE-SCALE DATA ANALYSIS Chair: Zhaoxue Tong

E0287: **Two-way homogeneity pursuit for quantile network vector autoregression**
Presenter: **Ganggang Xu**, University of Miami, United States
Co-authors: Xuening Zhu, Wenyang Liu, Jianqing Fan

While the vector autoregression (VAR) model has received extensive attention for modeling complex time series, quantile VAR analysis remains relatively underexplored for high-dimensional time series data. To address this disparity, a two-way grouped network quantile (TGNQ) autoregression model is introduced for time series collected on large-scale networks, known for their significant heterogeneous and directional interactions among nodes. The proposed model simultaneously conducts node clustering and model estimation to strike a balance between complexity and interpretability. To account for the directional influence among network nodes, each network node is assigned two latent group memberships that can be consistently estimated using the proposed estimation procedure. The approach extends the homogeneity pursuit introduced in another study for VAR models, offering attractive asymptotic properties. Theoretical analysis demonstrates the consistency of membership and parameter estimators even with an over-specified number of groups. With the correct group specification, estimated parameters are proven to be asymptotically normal, enabling valid statistical inferences. Moreover, a quantile information criterion is proposed for consistently selecting the number of groups. Simulation studies show promising finite sample performance, and the methodology is applied to analyze connectedness and risk spillover effects among Chinese A-share stocks.

E0399: **Semi-supervised triply robust inductive transfer learning**
Presenter: **Mengyan Li**, Bentley University, United States
Co-authors: Tianxi Cai, Molei Liu

A semi-supervised triply robust inductive transfer learning (STRIFLE) approach is proposed, which integrates heterogeneous data from a label-rich source population and a label-scarce target population and utilizes a large amount of unlabeled data simultaneously to improve the learning

accuracy in the target population. Specifically, a high dimensional covariate shift setting is considered, and two nuisance models are employed, a density ratio model and an imputation model, to combine transfer learning and surrogate-assisted semi-supervised learning strategies organically and achieve triple robustness. Different from double robustness, even if both nuisance models are misspecified or the distribution of Y given the surrogates and covariates is not the same between the two populations when the shifted source population and the target population share enough similarities, the triply robust STRIFLE estimator can still partially utilize the source population, and it is guaranteed to be no worse than the target-only surrogate-assisted semi-supervised estimator with negligible errors. These desirable properties of the estimator are established theoretically and verified in finite samples via extensive simulation studies. The STRIFLE estimator trains a Type II diabetes polygenic risk prediction model for the African American target population by transferring knowledge from electronic health records linked to genomic data observed in a larger European source population.

E0386: **Robust estimation of the high-dimensional precision matrix**
Presenter: **Zhaoxue Tong**, Florida State University, United States
Co-authors: Runze Li

Estimating the precision matrix (inverse of the covariance matrix) in high-dimensional settings is crucial for various applications, such as Gaussian graphical models and linear discriminant analysis. A robust and computationally efficient estimator is proposed for high-dimensional heavy-tailed data. Building upon winsorized rank-based regression, the method offers robustness without sacrificing computational tractability. The statistical consistency of the estimator is established, with a focus on conditions required for the error variance estimator in winsorized rank-based regression. For sub-Gaussian data, the sample variance meets the criteria, while for heavy-tailed data, a robust variance estimator based on the median-of-means approach is proposed. Simulation studies and real data analysis show that the proposed method performs well compared with existing works.

EO197 Room 202 NEW ADVANCES IN NONPARAMETRIC LEARNING AND HIGH-DIMENSIONAL ANALYSIS (VIRTUAL) Chair: Wei Qian

E0938: **Kernelized epsilon-greedy algorithm for nonparametric bandits with covariates**
Presenter: **Sakshi Arya**, Case Western Reserve University, United States

Multi-armed bandit algorithms are popular for sequential decision-making in several practical applications, ranging from online advertisement recommendations to mobile health. The goal of such problems is to maximize cumulative reward over time for a set of choices/arms while considering covariate (or contextual) information. Epsilon-greedy is a popular heuristic for the multi-armed bandit problem; however, it is not one of the most studied theoretical algorithms for the presence of contextual information. The epsilon-greedy strategy is studied in nonparametric bandits, i.e. when no parametric form is assumed for the reward functions. The similarities between the covariates and expected rewards are assumed to be modeled as arbitrary linear functions of the contexts' images in a specific reproducing kernel Hilbert space (RKHS). A kernel epsilon-greedy algorithm is proposed, and its convergence rates are established for estimation and cumulative regret, with only a boundness assumption of the errors. The rates closely match the optimal rates for linear contextual bandits when restricted to a finite-dimensional RKHS. The results are then compared with existing algorithms like kernel upper confidence bounds (UCB) on synthetic data.

E1112: **High-dimensional learning for multi-sourced matrix data**
Presenter: **Chenglong Ye**, University of Kentucky, United States

The abundance of data sources has made it possible to learn user preferences for products from various user-product interactions. In contrast to existing literature that models the differences between the regression coefficient means, we explore the setting when covariates are absent or hard to access due to privacy concerns. We treat user-product preferences as a partially observed main matrix using the primary data and then introduce a matrix transfer learning algorithm that leverages low-rank matrix estimation techniques to facilitate the transfer. The proposed method utilizes one or multiple auxiliary datasets to help predict the unobserved values of the main matrix. It outperforms existing covariate-free methods in both synthetic and real data settings. Theoretically, we derive an upper bound for the prediction error of our proposed approach to explain the interplay of the difference between the main and auxiliary data and their sample sizes. We further model the heterogeneous treatment effect estimation from the panel data in the causal inference setting as a transfer learning task, and adapt our algorithm to impute such effects. We benchmark our performance to existing matrix-completion-based algorithms and show the benefits of using ours.

E0419: **Bayesian covariate-dependent latent space model with information adaptivity**
Presenter: **Peng Zhao**, University of Delaware, United States
Co-authors: Yabo Niu

Modern network data analysis often involves analyzing network structures alongside covariate features to gain deeper insights into underlying patterns. However, traditional statistical network models may not fully consider the integration of such rich node characteristics. To address this gap, a new Bayesian high-dimensional covariate-dependent latent space model is introduced. This framework links latent vectors representing network structures with low-rank approximations of high-dimensional covariate observations, capturing their mutual dependencies. To adaptively integrate dependencies, a shrinkage is used prior to the discrepancy between latent network vectors and low-rank covariate approximation vectors, accommodating both consistencies and inconsistencies between them. To achieve computation efficiency, a mean-field variational inference algorithm is developed to approximate the posterior distribution. The concentration rate of the posterior is established within a suitable parameter space, and it is demonstrated how the model facilitates adaptive information aggregation between networks and covariates. Extensive simulations and real-world data analyses confirm the effectiveness of our approach.

E1076: **A dynamic Bayesian network approach to interbank market**
Presenter: **Wei Qian**, University of Delaware, United States

Motivated by the importance of understanding the nature of interconnections in interbank market networks and how the interconnections respond to changes in market conditions, a Bayesian dynamic interbank network model is proposed with probit link that simultaneously incorporates three underlying mechanisms for interbank trading: dynamic activity indices for overall market confidence, bank-specific latent variables for individual banks' fitness as borrowers or lenders, and pairwise covariates characterizing past trading relationships. Correspondingly, a computationally efficient Gibbs sampling algorithm is developed for sampling model parameters from the posterior distribution, which can be technically interesting by handling constrained parameters for model identifiability with latent components. By applying the developed Bayesian modeling technique to the e-MID interbank data analysis, estimation is obtained for empirically useful latent parameters that show not only satisfactory trading link forecasting performance but also gain in-depth insights into economic and pricing information for the interbank market. New model-based proxies of network topology change and relationship lending are specifically proposed to investigate their impact on relevant economic variables and the price of liquidity. The extension of the proposed Bayesian model to an alternative logit link is also discussed for modeling flexibility.

EO172 Room 204 RECENT ADVANCES IN SINGLE CELL ANALYSIS Chair: Lin Hou

E0392: **Statistical inference of cell-type proportions estimated from bulk expression data**
Presenter: **Biao Cai**, City University of Hong Kong, United States
There is a growing interest in cell-type-specific analysis from bulk samples with a mixture of different cell types. A critical first step in such analyses is the accurate estimation of cell-type proportions in a bulk sample. Although many methods have been proposed recently, quantifying the uncertainties associated with the estimated cell-type proportions has not been well-studied. Lack of consideration of these uncertainties can lead to missed or false findings in downstream analyses. A flexible statistical deconvolution framework is introduced that allows a general and subject-specific covariance of bulk gene expressions. Under this framework, a de-correlated constrained least squares method is proposed called DECALS that estimates cell-type proportions as well as the sampling distribution of the estimates. Simulation studies demonstrate that DECALS can accurately quantify the uncertainties in the estimated proportions, whereas other methods fail. Applying DECALS to analyze bulk gene expression data of post-mortem brain samples from the ROSMAP and GTEx projects, it is shown that taking into account the uncertainties in the estimated cell-type proportions can lead to more accurate identifications of cell-type-specific differentially expressed genes and transcripts between different subject groups, such as between Alzheimer's disease patients and controls and between males and females.

E1020: **SDePER: A hybrid machine learning and regression method to deconvolve spatial barcoding-based transcriptomic data**
Presenter: **Xiting Yan**, Yale School of Medicine, United States
Co-authors: Yunqing Liu, Ningshan Li, Ji Qi, Gang Xu, Jiayi Zhao, Nating Wang, Xiayuan Huang, Wenhao Jiang, Aurelien Justet, Taylor Adams, Robert Homer, Amei Amei, Ivan Rosas, Naftali Kaminski, Zuoheng Wang
Spatial barcoding-based transcriptomic (ST) data require cell-type deconvolution for cellular-level downstream analysis. The aim is to present SDePER, a hybrid machine learning and regression method, to deconvolve ST data using reference single-cell RNA sequencing (scRNA-seq) data. SDePER uses a machine learning approach to remove the systematic difference between ST and scRNA-seq data (platform effects) explicitly and efficiently to ensure the linear relationship between ST data and cell type-specific expression profile. It also considers the sparsity of cell types per capture spot and across-spots spatial correlation in cell type compositions. Based on the estimated cell-type proportions, SDePER imputes cell-type compositions and gene expression at unmeasured locations in a tissue map with enhanced resolution. Applications to coarse-grained simulated data and four real datasets showed that SDePER achieved more accurate and robust results than existing methods, suggesting the importance of considering platform effects, sparsity and spatial correlation in cell-type deconvolution.

E1012: **Symmetric graph convolutional auto-encoder for scalable and accurate study of spatial transcriptomics**
Presenter: **Zhixiang Lin**, The Chinese University of Hong Kong, Hong Kong
Recent advances in spatial transcriptomics (ST) have enabled comprehensive profiling of gene expression with spatial information in the context of the tissue microenvironment. However, with the improvements in the resolution and scale of ST data, deciphering spatial domains precisely while ensuring efficiency and scalability is still challenging. The focus is on SGCAST, an efficient auto-encoder framework for identifying spatial domains. SGCAST adopts a symmetric graph convolutional auto-encoder to learn aggregated latent embeddings via integrating the gene expression similarity and the proximity of the spatial spots. This framework in SGCAST enables a mini-batch training strategy, making SGCAST memory-efficient and scalable to high-resolution spatial transcriptomic data with many spots. SGCAST improves the overall accuracy of spatial domain identification on benchmarking data. The performance of SGCAST is validated on ST datasets at various scales across multiple platforms. The superior capacity of SGCAST is illustrated on spatial transcriptomic data analysis.

E1115: **A general statistical framework for FDR control in feature screening of single-cell genomics**
Presenter: **Xinzhou Ge**, Oregon State University, United States
Large-scale feature screening is essential in high-throughput biological data analysis, particularly for identifying genes that exhibit differential expression across various conditions. The false discovery rate is the most widely used criterion to ensure the reliability of screened features. The most famous Benjamini-Hochberg procedure for FDR control requires valid high-resolution p-values, which are, however, often hardly achievable because of the reliance on reasonable distributional assumptions or large sample sizes. We propose a general statistical framework, Clipper, for large-scale feature screening with theoretical FDR control and without p-value requirement. Extensive numerical studies have verified that Clipper is a versatile and effective tool for correcting the FDR inflation crisis in multiple bioinformatics applications. Notably, it effectively resolves the "double dipping" issue prevalent in single-cell genomic analyses.

EO324 Room 207 RECENT ADVANCEMENTS IN THE DESIGN AND ANALYSIS OF RANDOMIZED EXPERIMENTS Chair: Fan Li

E0458: **Hierarchical Bayesian modeling of heterogeneous outcome variance in cluster randomized trials**
Presenter: **Guangyu Tong**, Yale University, United States
Heterogeneous outcome correlations across treatment arms and clusters have been increasingly acknowledged in cluster randomized trials with binary endpoints, where analytical methods have been developed to study such heterogeneity. However, cluster-specific outcome variances and correlations have yet to be studied for cluster-randomized trials with continuous outcomes. Models fitted in the Bayesian setting with hierarchical variance structure are proposed to quantify heterogeneous variances across clusters and explain it with cluster-level covariates when the outcome is continuous. The models can also be extended to analyzing heterogeneous variances in individually randomized group treatment trials, with arm-specific cluster-level covariates, or in partially nested designs. Simulation studies are carried out to validate the performance of the newly introduced models across different settings. The model is illustrated with the Kerala Diabetes Prevention Program study, in which heterogeneous variances and intraclass correlation coefficients are identified across clusters, and cluster-level characteristics are examined associated with such heterogeneity.

E0395: **A lasso approach to covariate selection and average treatment effect estimation for RCTs**
Presenter: **Peter Schochet**, Mathematica Inc., United States
Covariates are often used to improve power for estimating average treatment effects (ATEs) for randomized controlled trials (RCTs). Covariate pre-specification is often recommended as it maintains the Type 1 error rate in repeated sampling but is not required by major RCT registries and clearinghouses across fields. Thus, many studies identify predictive covariates once primary outcomes have been collected. These post hoc methods, however, can suffer from a lack of transparency and replicability. An approach from a recent study that develops Lasso machine learning methods for the post-hoc selection of covariates for RCTs that can address these issues is discussed. The approach involves pre-specifying a fully replicable process for selecting covariates. The focus is on two-stage estimators, where the first stage involves Lasso estimation, and the second stage involves adjusting regression-based ATE estimators for covariates using the first-stage Lasso results. The design-based approach is nonparametric, applies to continuous, binary, and discrete outcomes, pertains to clustered and non-clustered RCTs, and can be easily implemented using existing software. The l_1 consistency of the estimated Lasso coefficients, a finite population central limit theorem for the ATE estimators, and design-based variance estimation are discussed. Simulations suggest good statistical performance in real-world settings.

E0439: **Tyranny-of-the-minority regression adjustment in randomized experiments**
Presenter: **Hanzhong Liu**, Tsinghua University, China

Regression adjustment is widely used in the analysis of randomized experiments to improve the estimation efficiency of the treatment effect. A weighted regression adjustment method is termed tyranny-of-the-minority (ToM), wherein units in the minority group are given greater weights. It is demonstrated that ToM regression adjustment is more robust than the previous study's regression adjustment with treatment-covariate interactions, even though these two regression adjustment methods are asymptotically equivalent in completely randomized experiments. Moreover, ToM regression adjustment can be easily extended to stratified randomized experiments and completely randomized survey experiments. The design-based properties of the ToM regression-adjusted average treatment effect estimator are obtained under such designs. In particular, it is shown that the ToM regression-adjusted estimator improves the asymptotic estimation efficiency compared to the unadjusted estimator, even when the regression model is misspecified, and is optimal in the class of linearly adjusted estimators. The asymptotic properties of various heteroscedasticity-robust standard errors are also studied, and recommendations for practitioners are provided. Simulation studies and real data analysis demonstrate ToM regression adjustment's superiority over existing methods.

EO150 Room 209 COMPLEXITY IN BIOLOGICAL, NETWORK, AND GEOMETRIC DATA ANALYSIS (VIRTUAL) Chair: Yusha Liu

E0636: **Geometric shapes of the tree-induced partition**
Presenter: **Hengrui Luo**, Rice University, United States

The purpose is to start by explaining the geometric nature of these partitions, illustrating how decision trees essentially draw multidimensional boundaries to segregate them by partitioning the input space. Expanding on this, the setting of tensor-based inputs and outputs is moved into, showcasing the complex geometric shapes that emerge and the new challenges they bring to the tree-based models. To navigate these complexities, the innovative tensor-on-tensor tree regression approach is introduced, designed to adeptly manage this multidimensional geometric partitioning. In conclusion, a step is taken back to reflect on the fundamental geometric principles underpinning tree-induced partitions and ponder future research avenues in this fascinating intersection of geometry and tree-based models.

E0643: **Covariate-assisted Bayesian graph learning for heterogeneous data**
Presenter: **Yabo Niu**, University of Houston, United States
Co-authors: Yang Ni, Debdeep Pati, Bani Mallick

In traditional Gaussian graphical models, data homogeneity is routinely assumed with no extra variables affecting conditional independence. In modern genomic datasets, there is an abundance of auxiliary information, which often gets under-utilized in determining the joint dependency structure. A new Bayesian approach is presented to model undirected graphs underlying heterogeneous multivariate observations with additional assistance from covariates. Building on product partition models, a novel covariate-dependent Gaussian graphical model is proposed that allows graphs to vary with covariates so that observations whose covariates are similar share a similar undirected graph. To efficiently embed Gaussian graphical models into the proposed framework, both Gaussian likelihood and pseudo-likelihood functions are explored. Moreover, the proposed model induced by the prior has large support and is flexible to approximate any piece-wise constant conditional variance-covariance matrices. Furthermore, based on the theory of fractional likelihood, the rate of posterior contraction is minimax optimal, assuming the true density is a Gaussian mixture with a known number of components. The efficacy of the approach is demonstrated via numerical studies and analysis of protein networks for a breast cancer dataset assisted by genetic covariates.

E0690: **Limit results for distributed estimation of invariant subspaces in multiple networks inference and PCA**
Presenter: **Runbing Zheng**, Johns Hopkins University, United States
Co-authors: Minh Tang

The problem of distributed estimation of the leading singular vectors is studied for a collection of matrices with shared invariant subspaces. In particular, an algorithm is considered that first estimates the projection matrices corresponding to the leading singular vectors for each individual matrix, then computes the average of the projection matrices, and finally returns the leading eigenvectors of the sample averages. It is shown that the algorithm, when applied to (1) parameters estimation for a collection of independent edge random graphs with shared singular vectors but possibly heterogeneous edge probabilities or (2) distributed PCA for independent sub-Gaussian random vectors with spiked covariance structure, yields estimates whose row-wise fluctuations are normally distributed around the rows of the true singular vectors. Leveraging these results, a two-sample test is also considered for the null hypothesis that a pair of random graphs have the same edge probabilities, and a test statistic is presented whose limiting distribution converges to a central (resp. non-central) χ^2 under the null (resp. local alternative) hypothesis.

E0786: **A flexible model for correlated count data, with application to multi-condition differential gene expression analyses**
Presenter: **Yusha Liu**, University of North Carolina at Chapel Hill, United States

Detecting differences in gene expression is an important part of single-cell RNA sequencing experiments, and many statistical methods have been developed for this aim. Most differential expression analyses focus on comparing expression between two groups (e.g., treatment vs. control). However, there is increasing interest in multi-condition differential expression analyses in which expression is measured in many conditions, and the aim is to accurately detect and estimate expression differences in all conditions. It is shown that directly modeling single-cell RNA-seq counts in all conditions simultaneously while also inferring how expression differences are shared across conditions leads to greatly improved performance for detecting and estimating expression differences compared to existing methods. The potential of this new approach is illustrated by analyzing data from a single-cell experiment studying the effects of cytokine stimulation on gene expression.

EO056 Room 210 RECENT ADVANCES IN MODELING COMPLEX POPULATION DATA (VIRTUAL) Chair: Alfonso Landeros

E0219: **Multivariate varying coefficient spatiotemporal model**
Presenter: **Damla Senturk**, University of California Los Angeles, United States

As of 2020, 807,920 individuals in the U.S. had end-stage kidney disease (ESKD), with about 70% of patients on dialysis, a life-sustaining treatment. Dialysis patients experience high mortality rates, where frequent hospitalizations are a major contributor to morbidity and mortality. There is growing interest in identifying the risk factors for the correlated outcomes of hospitalization and mortality among dialysis patients across the U.S. Utilizing national data from the United States Renal Data System (USRDS), a novel multivariate varying coefficient spatiotemporal model is proposed to study the time dynamic effects of risk factors (e.g., urbanicity and area deprivation index) on the multivariate outcome of hospitalization and mortality rates, as a function of time on dialysis. While capturing time-varying effects of risk factors on the mean, the proposed model also incorporates spatiotemporal patterns of the residuals for efficient inference. Estimation is based on the fusion of functional principal component analysis and Markov Chain Monte Carlo techniques, following basis expansions of the varying coefficient functions and multivariate Karhunen-Loeve expansion of region-specific random deviations. Novel applications to the USRDS data highlight significant risk factors of hospitalizations and mortality as well as characterizing time periods on dialysis and spatial locations across the U.S. with elevated hospitalization and mortality risks.

E0222: **Spatiotemporal multilevel joint modeling of generalized longitudinal and survival outcomes in end-stage kidney disease**
Presenter: **Esra Kurum**, University of California, Riverside, United States

Individuals with end-stage kidney disease (ESKD) on dialysis experience high mortality and excessive burden of hospitalizations over time relative to comparable Medicare patient cohorts without kidney failure. A key interest in this population is to understand the time-dynamic effects of multilevel risk factors that contribute to the correlated outcomes of longitudinal hospitalization and mortality. Multilevel data from the United States Renal Data System (USRDS) is utilized, where repeated measurements/hospitalizations over time are nested in patients and patients are nested within (health service) regions across the U.S. A novel spatiotemporal multilevel joint model (STM-JM) is developed that accounts for the aforementioned hierarchical structure of the data while considering the spatiotemporal variations in both outcomes across regions. The proposed STM-JM includes time-varying effects of multilevel (patient- and region-level) risk factors on the correlated outcomes and incorporates spatial correlations across the spatial regions via a multivariate conditional autoregressive correlation structure. Efficient estimation and inference are performed via a Bayesian framework. An application of the proposed method to the USRDS data highlights significant time-varying effects of risk factors on hospitalization and mortality and identifies specific time periods on dialysis and spatial locations across the U.S. with elevated hospitalization and mortality risks.

E0238: **A Bayesian trivariate joint model of kidney disease progression, recurrent events, and terminal event in CKD**
Presenter: **Danh Nguyen**, University of California, Irvine, United States

Nearly 37 million adults in the U.S. have chronic kidney disease (CKD). The longitudinal trajectory of kidney function decline in patients with CKD is intricately related to the development of cardiovascular disease (CVD) and eventual "terminal" events (kidney failure and mortality). Understanding of the mechanism and risk factors underlying the three key outcome processes, (1) CKD progression, (2) CVD, and (3) subsequent terminal events in the CKD patient population, remains incomplete. Thus, a novel trivariate joint model is developed to study the risk factors associated with the interdependent outcomes of kidney function (as measured by longitudinal estimated glomerular filtration rate), recurrent cardiovascular events, and terminal events. Efficient estimation and inference are proposed within a Bayesian framework using Markov Chain Monte Carlo and Bayesian P-splines for hazard functions. The method is applied to study the aforementioned trivariate processes using data from the Chronic Renal Insufficiency Cohort Study, an ongoing prospective cohort study.

E0216: **Bayesian multivariate joint model of longitudinal, recurrent, and competing risk terminal events**
Presenter: **Qi Qian**, University of California, Los Angeles, United States

The longitudinal decline of kidney function is intricately related to hospitalizations due to cardiovascular disease (CVD) and eventual "terminal" kidney failure and mortality in patients with chronic kidney disease (CKD). To better understand the mechanism and risk factors underlying these interdependent processes, as well as to tailor decision making to the needs of individual patients, a novel Bayesian joint model is developed for the interdependent outcomes of kidney function, recurrent cardiovascular events, and competing-risk terminal events (kidney failure and death). The proposed joint modelling not only allows the study of the risk factors associated with each outcome but also facilitates the dynamic updating of cumulative incidence probabilities of each competing risk for future subjects based on their profiles of previous longitudinal measurements and recurrent events. Efficient and flexible estimation and prediction procedures are proposed within a Bayesian framework using Markov Chain Monte Carlo, and the predictive ability is assessed via the dynamic area under the receiver operating characteristic (ROC) curves (AUC) and expected Brier score (BS). The efficacy of the proposed joint model and prediction procedure is shown via extensive simulations. The proposed methodology is also applied to data from the Chronic Renal Insufficiency Cohort (CRIC) study.

EO118 Room 307 SEMI/NONPAR. METHODS FOR HIGHLY CORRELATED HIGH DIMENSIONAL DATA (VIRTUAL) Chair: Inyoung Kim

E0343: **Scalable Bayesian inference on high-dimensional multivariate linear regression**
Presenter: **Kyoungjae Lee**, Sungkyunkwan University, Korea, South
Co-authors: Xuan Cao

Jointly estimating the regression coefficient matrix and the error precision matrix in high-dimensional multivariate linear regression are considered. Scalable computation in this framework is often challenging for Bayesian methods, and thus, available approaches either adopted a generalized likelihood without guaranteeing the positive definiteness of the precision matrix or employed a maximization algorithm to target the posterior mode, which cannot handle uncertainty. Two Bayesian methods are proposed: an exact method and an approximate two-step method. An exact method is first proposed based on spike and slab priors for the coefficient matrix and DAG-Wishart prior to the error precision matrix. The complexity of the proposed algorithm is comparable to the state-of-the-art generalized likelihood-based Bayesian method. To further enhance scalability, a two-step approach is developed by ignoring the dependency structure among response variables. After estimating the coefficient matrix, the posterior of the error precision matrix is calculated based on the estimated errors. To justify the two-step approach, it is proven that (i) selection consistency and posterior convergence rates for the coefficient matrix and (ii) selection consistency for the directed acyclic graph (DAG) of errors. The practical performance of the proposed methods is demonstrated through synthetic and real data analysis.

E0743: **Semiparametric joint functional graphical regression**
Presenter: **Mengkun Chen**, Virginia Polytechnic Institute and State University, China
Co-authors: Inyoung Kim

A joint semiparametric functional graphical regression is proposed that combines a semiparametric high-dimensional multivariate functional regression with a functional graphical model for highly structured functional predictors. A functional graphical model has been used to study the dependence structures among functional variables. Various high-dimensional functional regressions have been developed to study the association between the response and functional predictors. For structured functional predictors, connected functional predictors act together to the response. Conversely, the dependence structure among the functional predictors can also be affected by the response in the functional regression. Therefore, a unified and integrated method is developed to simultaneously identify important functional predictors and build a network among these functional predictors. By jointly estimating the graphical model among the correlated functional predictors and linking it to the response through the regression model, the model estimation, variable selection, and prediction accuracy can be improved when dealing with the highly structured correlated predictors, and also recover the connectivity among them. The procedure is established under the hybrid framework, a mixture of Frequentist and Bayesian methods. The advantages of the proposed method are demonstrated through several simulation studies and finally implemented in the brain signal data.

E0780: **Informed weighted Dirichlet process mixture for functional clustering in highly correlated high-dimensional data**
Presenter: **Wenyu Gao**, University of North Carolina at Charlotte, United States
Co-authors: Inyoung Kim

Functional clustering in high-dimensional data poses challenges, especially in scenarios with unknown cluster counts. While nonparametric Bayesian methods such as the Dirichlet process mixture (DPM) model offer approaches, they often do not effectively leverage observational information. Conversely, the weighted Dirichlet process mixture (WDPM) model incorporates prior information via a weight function. However, its investigation remains limited, particularly in functional clustering. Informed weight functions are explored for WDPM in functional clustering,

addressing the gap in research by exploring covariates beyond Euclidean distances. The method is applied to fMRI data from autism spectrum disorder (ASD) patients, integrating spatial correlations and demographic information to enhance clustering accuracy.

E0872: **Functional nonconvex penalization kernel smoothing for high-dimensional additive regression**
Presenter: **Seyoung Park**, Yonsei University, Korea, South

Smooth backfitting has been shown to be a useful estimation technique for additive regression models in various scenarios. However, current studies have limitations because they are restricted to cases with a finite number of covariates or use only a penalized L1 method designed for high-dimensional settings. The iterative smooth backfitting algorithm, although simple and well-studied, tends to be very time-consuming, especially in high-dimensional settings. It has also been observed that one penalty can introduce significant estimation bias, whereas concave regularization can improve estimation. New kernel estimators and an efficient algorithm of nonconvex smooth backfitting for ultra-high dimensional additive models are presented. It is shown that the proposed nonconvex optimization has the oracle estimator as a unique stationary point. For the implementation of the proposed method, a composite gradient algorithm is designed and proven that the proposed iterative algorithm achieves a near-global optimum. The proposed algorithm allows parallel computation for updating the component functions in each iteration, significantly reducing the computational time compared to the existing iterative smooth backfitting algorithm.

EO026 Room 313 CONTEMPORARY APPROACHES TO ENVIRONMENTAL AND SPATIO-TEMPORAL STATISTICS Chair: Pulong Ma

E0244: **Clustering spatial functional data using a geographically weighted Dirichlet process**
Presenter: **Guanyu Hu**, The University of Texas Health Science Center at Houston, United States

A Bayesian nonparametric clustering approach is proposed to study the spatial heterogeneity effect for functional data observed at spatially correlated locations. A geographically weighted Chinese restaurant process equipped with a conditional autoregressive is considered prior to fully capturing the spatial correlation of function curves. To sample efficiently from the model, a prior called Quadratic Gamma is customized to ensure conjugacy. A Markov Chain Monte Carlo algorithm is designed to infer simultaneously the posterior distributions of the number of groups and the grouping configurations. The superior numerical performance of the proposed method over competing methods is demonstrated using simulated examples and various applications.

E0425: **Fast computer model calibration using annealed and transformed variational inference**
Presenter: **Won Chang**, University of Cincinnati, United States
Co-authors: Jaewoo Park

Computer models play a crucial role in numerous scientific and engineering domains. To ensure the accuracy of simulations, it is essential to properly calibrate the input parameters of these models through statistical inference. While Bayesian inference is the standard approach, employing Markov Chain Monte Carlo methods often encounters computational hurdles due to the costly evaluation of likelihood functions and slow mixing rates. Although variational inference (VI) can be a fast alternative to traditional Bayesian approaches, VI has limited applicability due to boundary issues and local optima problems. To address these challenges, flexible VI methods are proposed based on deep generative models that do not require parametric assumptions on the variational distribution. A surjective transformation is embedded in the framework to avoid posterior truncation at the boundary. Additionally, theoretical conditions are provided that guarantee the success of the algorithm. Furthermore, the temperature annealing scheme can prevent being trapped in local optima through a series of intermediate posteriors. The method is applied to infectious disease models and a geophysical model, illustrating that the proposed method can provide fast and accurate inference compared to its competitors.

E0860: **Physics-informed priors with application to boundary layer velocity**
Presenter: **Luca Menicali**, University of Notre Dame, United States
Co-authors: Stefano Castruccio, David Richter

One of the most popular recent areas of machine learning predicates the use of neural networks augmented by information about the underlying process in the form of partial differential equations (PDEs). These physics-informed neural networks are obtained by penalizing the inference with a PDE and have been cast as a minimization problem, which currently lacks a formal approach to quantifying uncertainty. A novel model-based framework is proposed, which regards the PDE as prior information of a deep Bayesian neural network. The prior is calibrated without data to resemble the PDE solution in the prior mean, while the degree of confidence in the PDE with respect to the data is expressed in terms of the prior variance. The information embedded in the PDE is then propagated to the posterior, yielding physics-informed forecasts with uncertainty quantification. The approach is applied to experimentally obtained turbulent boundary layer velocity in a wind tunnel using an appropriately simplified Navier-Stokes equation. The approach requires very few observations to produce physically consistent forecasts as opposed to non-physical forecasts stemming from non-informed priors, thereby allowing forecasting complex systems where some amount of data, as well as some contextual knowledge, is available.

E0892: **Modeling large-scale high-resolution spatial-temporal data with deep ESN and SPDE**
Presenter: **Kesen Wang**, King Abdullah University of Science and Technology, Saudi Arabia

In the past decades, clean and renewable energy has gained increasing attention due to a global effort to reduce carbon footprints. Wind, a primary form of renewable energy, has been heavily invested as a substitute and replenishment for the existing energy portfolio. However, wind possesses a highly spatially non-linear dynamic nature and varies rather capriciously in time, rendering wind modelling quite challenging. Given the complex dynamics of the wind, it is challenging for the existing statistical models to fully grasp the underlying spatial and temporal dependence structures. Hence, there is a need for a non-linear dynamic model of high temporal resolution, which precisely captures the spatial dependence and accurately describes the fast-evolving wind dynamics in time. A combined model based on machine learning and stochastic partial differential equation (SPDE) is proposed to address the complex wind dynamics in time and preserve dependence structure in space.

EO108 Room 405 RECENT ADVANCES IN HIGH DIMENSIONAL DATA ANALYSIS Chair: Chi Tim Ng

E0273: **Principal stratification with U-statistics under principal ignorability**
Presenter: **Xinyuan Chen**, Mississippi State University, United States
Co-authors: Fan Li

Principal stratification is a popular framework for causal inference in the presence of an intermediate outcome. While the principal average treatment effects have traditionally been the default target of inference, it may not be sufficient when the interest lies in the relative favorability of one potential outcome over the other within the principal stratum. Thus, the principal generalized causal effect estimands are introduced, which extend the principal average causal effects to accommodate nonlinear contrast functions. Under principal ignorability, the theoretical results are expanded in another study to a wider class of causal estimands in the presence of a binary intermediate variable. Specifically, identification formulas are developed, and the efficient influence functions of the generalized estimands are derived for principal stratification analyses. These efficient influence functions motivate a set of multiply robust estimators and lay the ground for obtaining efficient debiased machine learning estimators via cross-fitting based on U-statistics. The proposed methods are illustrated through simulations and the analysis of a data example.

E0498: **Change-point detection of time-varying Cox model**
Presenter: **Kaimeng Zhang**, Zhejiang University, China
Co-authors: Chi Tim Ng
A novel Cox model is presented that incorporates time-varying coefficients for time-dependent survival analysis. Transformable penalized likelihood estimation is employed to address the challenge of change point detection in the proposed model. The asymptotic properties of the local solutions are established, and numerical studies are conducted to demonstrate the effectiveness of the proposed approach. The efficacy of the method is illustrated through the analysis of data from the Rossi dataset. Overall, the research contributes to the advancement of time-dependent survival analysis by introducing a robust and efficient methodology with potential applications in various fields, including medicine and engineering.

E0713: **Admixture analysis of high-dimensional time series data**
Presenter: **Chi Tim Ng**, Hang Seng University of Hong Kong, Hong Kong
In the Big Data Era, the data are collected from many locations at different time points. For example, the air pollution data may consist of 15 indices measured hourly at around 1000 cities over five years. Policymakers in economics, environmental protection, and agriculture are interested in identifying a few locations that serve as proxies for the driving forces that affect all locations. The changes in the contribution proportions of these driving forces over time are useful indicators for both practitioners and academia. The goal of this proposed research is to study the novel statistical models and methods that can extract information from the multi-site multivariate time series data about the hidden driving forces that cannot be observed directly. This is solved in the proposed research by introducing the ideas of admixture analysis and latent Dirichlet distributions. These concepts have been employed by the researchers in the context of population genetics and text mining. This is the first research that extends these ideas to multi-site multivariate time series analysis. The admixture components in the novel model can then be used to describe the so-called hidden driving forces. With the extra time ingredient, the time of appearance and disappearance of a driving force is further investigated. This cannot be done directly with existing time series clustering methods and factor analysis methods.

EO198 Room 408 ADVANCES IN STATISTICAL NETWORK ANALYSIS (VIRTUAL) Chair: Keith Levin

E0259: **Node role discovery in networks: Approximating equitable partitions**
Presenter: **Michael Schaub**, RWTH Aachen University, Germany
Co-authors: Michael Scholkemper, Michael Schaub
Similar to community detection, partitioning the nodes of a network according to their structural roles aims to identify the fundamental building blocks of a network. The found partitions can be used, e.g., to simplify descriptions of the network connectivity, to derive reduced order models for dynamical processes unfolding on processes, or as ingredients for various graph mining tasks. A fresh look is offered at the problem of role extraction and its differences from community detection, and a definition of node roles related to graph-isomorphism tests, the Weisfeiler-Leman algorithm, and equitable partitions is presented. Two associated optimization problems (cost functions) are studied, grounded in ideas from graph isomorphism testing, and present theoretical guarantees associated with the solutions of these problems. Finally, the approach is validated via a novel "role-infused partition benchmark", a network model from which we can sample networks in which nodes are endowed with different roles in a stochastic way.

E0362: **Likelihood of weight loss or ACRONYM: Augmented degree corrected community reticulately organized network yielding model**

Presenter: **Benjamin Leinwand**, Stevens Institute of Technology, United States
Co-authors: Vince Lyzinski
Modeling networks can serve as a means of summarizing high-dimensional complex systems. Adapting an approach devised for dense, weighted networks, a new method is proposed for generating and estimating unweighted networks. This approach can describe a broader class of potential networks than existing models, including those where nodes in different "modules" connect to one another via various attachment mechanisms, inducing flexible and varied community structures. While unweighted edges provide less resolution than continuous weights, restricting to the binary case permits the use of likelihood-based estimation techniques, which can improve the estimation of nodal features. The extra flexibility may contribute to a different understanding of network-generating structures, particularly for networks with heterogeneous densities in different regions.

E1018: **Limits and potentials of graph neural networks: From spectral simplicity to polynomial bounds**
Presenter: **Arash Amini**, UCLA, United States
The interplay between complexity and efficiency is explored in graph neural networks (GNNs), drawing on theoretical and empirical insights. Empirical evidence shows how spectral GNNs for semi-supervised node classification benefit from simplicity. The need for complex GNN architectures is challenged, showing simpler methods can rival or exceed their performance. This motivates a theoretical deep dive into polynomial GNNs (Poly-GNNs), examining the effect of 'depth' (higher-order polynomial features) within a contextual stochastic block model. Contrary to expectations, it is discovered that increasing depth mainly modifies constants without altering the fundamental rates of separability. This finding emphasizes the profound effect of 'network noise' in deep GNNs and questions the prevailing belief that greater complexity necessarily improves discriminative power in GNNs.

E0158: **Nonparametric priors for graph matching**
Presenter: **Lizhen Lin**, The University of Notre Dame, United States
The graph matching problem seeks to establish correspondences between nodes in two graphs, when observing their connections. Achievable solutions to this statistical problem are often elusive due to computational complexity. Uncertainty quantification is even more challenging. We present a novel modeling idea for graph matching that takes inspiration from the extensive Bayesian nonparametric methodologies developed for random partition models. In particular, we combine a correlated Erdos-Renyi likelihood with a nonparametric prior on the space of random permutations obtained via an upgrade of the well-known Chinese restaurant process (CRP). We believe this to be the first model-based Bayesian nonparametric approach to the graph-matching problem. There are several possible generalizations of the proposed model: the CRP prior to random permutations can be replaced by more complex constructions, inducing structural information in the model. We propose model extensions introducing an interplay between a latent block connectivity structure and the cycle structure of the (random) permutation, realizing the matching. In such a framework, the matching is driven by the connectivity behaviour of the nodes: we can either allow matching just within the same connectivity block or, leveraging a conditional partially exchangeable scheme, induce higher prior probability on within-block matching without ruling out across-block matching. Posterior inference is carried out via MCMC algorithms.

EO321 Room 411 (Virtual sessions) ADVANCES IN FACTOR ANALYSIS Chair: Alexander Shkolnik

E0882: **Structural breaks and factor selection**

Presenter: **Siddhartha Chib**, Washington University in Saint Louis, United States

A new approach is developed to select factors, allowing the set to change at multiple unknown break dates. In a six-factor model since 1963, a marked shift towards parsimonious models has been documented in the last two decades. Before 2005, either five or six factors were selected, but just two were selected thereafter. This finding offers a simple implication for the factor zoo literature: ignoring breaks detects additional factors that are no longer relevant. Moreover, all omitted factors are priced by the selected factors in every regime. Finally, the selected factors outperform popular factor models as an investment strategy

E0906: **Deflated heteroPCA: Overcoming the curse of ill-conditioning in heteroskedastic PCA**

Presenter: **Yuxin Chen**, University of Pennsylvania, United States

The focus is on estimating the column subspace of a low-rank n1 x n2 matrix X from contaminated data. How to obtain optimal statistical accuracy while accommodating the widest range of signal-to-noise ratios (SNRs) becomes particularly challenging in the presence of heteroskedastic noise and unbalanced dimensionality (i.e., n2»n1). While the state-of-the-art algorithm emerges as a powerful solution for solving this problem, it suffers from "the curse of ill-conditioning," namely, its performance degrades as the condition number of X grows. In order to overcome this critical issue without compromising the range of allowable SNRs, a novel algorithm, called -, is proposed that achieves near-optimal and condition-number-free theoretical guarantees in terms of both L2 and fine-grained statistical accuracy. Further, an application of the algorithm and theory to two canonical examples, the factor model and tensor PCA, leads to remarkable improvement for each application.

E0846: **Convergence rate of the James-Stein principal component**

Presenter: **Youhong Lee**, University of California, Santa Barbara, United States

Co-authors: Alexander Shkolnik

The James-Stein estimator is presented for eigenvectors, building upon previous research that highlights its ability to reduce the distance to the true signal through the shrinkage constant and corrected eigenvector. Moving beyond the point estimates, the investigation focuses on the variability inherent to the James-Stein estimator. The sources of this variability are explored, and its impact is assessed on the estimator's components. The aim is to elucidate the asymptotic variance of the James-Stein estimator, particularly in relation to the signal-to-noise ratio and the shrinkage target. Moreover, the effect of James-Stein-type shrinkage is examined on this variability, differentiating between signal-noise and noise-noise correlations.

E1000: **The quadratic optimization bias of large covariance matrices**

Presenter: **Alexander Shkolnik**, University of California, Santa Barbara, United States

Co-authors: Hubeyb Gurdogan

The focus is on a notable puzzle involving the interactions between an optimization of a multivariate quadratic function and a plug-in estimator of a spiked covariance matrix. When the largest eigenvalues (i.e., the spikes) grow with the problem dimension, the optimized solutions inherit highly counter-intuitive properties out-of-sample. The plug-in estimator must be fine-tuned precisely or rendered virtually useless for a sufficiently large dimension. Central to the description is the quadratic optimization function, the roots of which determine this fine-tuning property. An estimator of this root is derived from a finite number of observations of a high-dimensional vector, and the consistency is proven within the high-dimensional limit. This estimator informs a low dimensional subspace correction of the sample covariance matrix when the dimension is large relative to the sample size.

Friday 19.07.2024 10:25 - 11:40 Parallel Session L – EcoSta2024

EO100 Room 102 RECENT DEVELOPMENTS IN ECONOMETRIC THEORY Chair: Cy Sin

E0779: **Specification analysis in quantile regression models with control functions**
Presenter: **Jen-Che Liao**, National Chengchi University, Taiwan
Co-authors: Xiaojun Song

Nonparametric consistent tests are proposed for correct specifications in a quantile regression model with a control function. Specifically, the tests include the specification test of a parametric conditional quantile function, the specification test of a parametric control function, and the test of exogeneity of explanatory variables. Using a transformation of quantile restrictions to mean restrictions under the null hypothesis, the proposed tests are constructed based on the mixed residuals that combine the null and alternative models. The asymptotic properties of the proposed tests are established. To improve inference accuracy, a simple wild bootstrap approach is suggested, with its theoretical justifications provided. A simulation study examines the finite-sample properties of the proposed tests. For empirical illustration, the proposed methods are applied to testing quantile Engel curves with endogenous total expenditure.

E0835: **Post-empirical Bayes regression**
Presenter: **Yu-Chang Chen**, University of California, San Diego, United States
Co-authors: Sheng-Kai Chang, Shuo-Chieh Huang

Empirical Bayes (EB) methods are widely utilized in economics for estimating individual and group-level fixed effects across diverse contexts, including teacher value-added, hospital qualities, and neighborhood effects. While estimates generated by EB are often incorporated into other statistical analyses like regression models, the econometric properties of post-EB regression have not been thoroughly investigated. This knowledge gap is addressed through two key contributions. First, a unified framework is introduced for two-step EB methods that apply to linear and non-linear models, offering insights into their frequentist properties and assessing their robustness against model misspecification. Second, a critical evaluation of commonly used two-step EB methods is undertaken in existing empirical research. The analysis demonstrates that naive implementations of post-EB regression can introduce a systematic bias, particularly in non-linear models

E0985: **On sandwich variance estimation: Bayesian versus frequentist**
Presenter: **Cy Sin**, National Tsing Hua University, Taiwan

It is well known the Eicker-Huber-White variances are not only heteroskedasticity-robust and nonlinearity-robust but also nonnormality-robust. Recently, some of the Eicker-Huber-White variances have been reviewed. Among other things, it is concluded: (a) Simulation studies suggest HC(4), a variant of robust variance estimator proposed in another study, does not over-reject or mildly under-rejects even in cases of non-normal distributions; (b) The original robust variance (denoted by HC(0)) and its variants considered by the prevalent statistical software (such as R and STATA), are all asymptotically equivalent. The focus is on a Bayesian approach, considering the balanced loss function (BLF) proposed in a recent study. Unlike the conventional inference loss function (ILF), this function balances estimation error and lack of fit. This function is, in turn, generalized upon the first proposed in another study, where the attention to normality-type likelihoods is confined. The Bayesian estimator of the variance-covariance matrix is asymptotically equivalent to the frequentist estimator. Non-normal likelihoods are covered. Simulation studies that compare the Bayesian estimator with the conventional estimators are performed.

EO117 Room 103 NEW DEVELOPMENTS IN THE FRONTIERS OF PRECISION MEDICINE AND DATA SCIENCE Chair: Shanghong Xie

E0345: **Fusing individualized treatment rules using auxiliary outcomes**
Presenter: **Donglin Zeng**, University of Michigan, United States

An individualized treatment rule (ITR) is a decision rule that recommends treatments for patients based on their covariates. In practice, the optimal ITR that maximizes its associated value function is also expected to cause little harm to other non-primary outcomes. Hence, one goal is to learn the ITR that not only maximizes the value function for the primary outcome but also approximates the optimal rule for the other auxiliary outcomes as closely as possible. A fusion penalty is proposed to encourage ITRs based on the primary outcome and auxiliary outcomes to yield similar recommendations. A surrogate loss function is then optimized using empirical data for estimation. The non-asymptotic properties are derived for the proposed method and show that the agreement rate between the estimated ITRs for primary and auxiliary outcomes converges faster to the true agreement rate than methods without using auxiliary outcomes. Finally, simulation studies and a real data example are used to demonstrate the finite-sample performance of the proposed method.

E0418: **Empirical Likelihood for fair classification**
Presenter: **Yichuan Zhao**, Georgia State University, United States
Co-authors: Pangpang Liu

Machine learning algorithms are commonly being deployed in decision-making systems that have a direct impact on human lives. However, if these algorithms are trained solely to minimize training/test errors, they may inadvertently discriminate against individuals based on their sensitive attributes, such as gender, race or age. Recently, algorithms that ensure fairness have been developed in the machine learning community. Fairness criteria are applied by these algorithms to measure fairness, but they often use the point estimate to assess fairness and fail to consider the uncertainty of the sample fairness criterion once the algorithms are deployed. It is suggested that fairness should be assessed by taking uncertainty into account. The covariance is used as a proxy for fairness and develops the confidence region of the covariance vector using empirical likelihood. The confidence region-based fairness constraints for classification take uncertainty into consideration during fairness assessment. The proposed confidence region can be used to test fairness and impose fairness constraints using the significance level as a tool to balance accuracy and fairness. Simulation studies show that the method exactly covers the target Type I error rate and effectively balances the trade-off between accuracy and fairness. Finally, data analysis is conducted to demonstrate the effectiveness of the method.

E0527: **Innovative trial designs in mobile and digital health using reinforcement learning**
Presenter: **Bibhas Chakraborty**, Duke-NUS Medical School, National University of Singapore, Singapore

Mobile health (mHealth) or, more broadly, digital health interventions (e.g., motivational text messages or nudges to promote healthy behaviors) are becoming increasingly common in tandem with advances in mobile and wearable sensor technologies. An innovative trial design arising in mHealth, namely, the micro-randomized trial (MRT), is discussed, which involves sequential, within-person randomization over many instances. The basic MRT design can be further improved to make it adaptive, thereby enabling it to learn from accumulated data as the trial progresses. This is appealing from an ethical perspective since adaptive learning tends to make better interventions available to the trial participants. Adaptive learning in such trial designs is often operationalized via reinforcement learning algorithms. Specifically, the role of a particular algorithm called Thompson sampling in designing adaptive MRTs is discussed. Theoretical and simulation results are shown to validate the proposed approach. mHealth clinical trials are discussed as case studies.

EO144 Room 104 STATISTICAL LEARNING IN NETWORK DATA Chair: Tianxi Li

E0190: **Distribution-free matrix prediction under arbitrary missing pattern**
Presenter: **Yuan Zhang**, The Ohio State University, United States
Co-authors: Meijia Shao
The purpose is to study the open problem of conformalized entry prediction in a row/column-exchangeable matrix. The matrix setting presents novel and unique challenges, but there exists little work on this interesting topic. The problem is meticulously defined, differentiating it from closely related problems and rigorously delineating the boundary between achievable and impossible goals. Two practical algorithms are then proposed. The first method provides a fast emulation of the full conformal prediction, while the second method leverages the technique of algorithmic stability for acceleration. Both methods are computationally efficient and can effectively safeguard coverage validity in the presence of arbitrary missing patterns. Further, the impact of missingness on prediction accuracy is quantified and fundamental limit results are established. Empirical evidence from synthetic and real-world datasets corroborates the superior performance of the proposed methods.

E0709: **A local perspective in general latent space network models**
Presenter: **Rachel Wang**, University of Sydney, Australia
Co-authors: Lijia Wang, Xin Tong, Xiao Han
The neighborhood effect in social networks significantly influences an individual's decision-making process, opinion formation, and various other personal dynamics. Thus, to understand the role of social networks in shaping individual behaviors and attitudes, it is important to begin with an understanding of an individual's localized viewpoint within the global network context. A general latent space network model and the problem of inferring the latent positions of the nodes are considered, utilizing only a partial information network centered on a given individual. Using a projected gradient descent algorithm, the convergence rate of the estimates is shown to depend on the neighborhood features of the node and a quantity is defined accordingly to measure the amount of bias in this individual's local view. Using simulated and real networks, particularly the co-sponsorship network in the US Congress, these local estimates of latent positions are compared with the global ones and show how the framework allows us to obtain a more nuanced understanding of local perspectives within social networks.

E0848: **Informative periphery detection and post-detection inference on weighted directed networks**
Presenter: **Wenqin Du**, University of Southern California, United States
Co-authors: Wen Zhou, Tianxi Li
In network analysis, introduced by peripheral or non-essential components, noises and biases can mask pivotal structures and hinder the efficacy of many network modeling and inference procedures. Recognizing this, the identification of the core-periphery (CP) structure has emerged as a crucial data pre-processing step. Many existing efforts either fail to account for the directionality or lack the theoretical justification of the identification procedure. Answers to three pressing questions are sought: (i) How can informative and non-informative structures in weighted directed networks be distinguished? (ii) What approach offers computational efficiency in discerning these components? (iii) Upon detecting CP structure, can uncertainty be quantified to evaluate the detection? The signal-plus-noise model is adopted, categorizing uniform relational patterns as non-informative, by which the sender and receiver peripheries are defined. Furthermore, instead of confining the core component to a specific structure, it is considered complementary to either the sender or receiver peripheries. Based on the definitions of the sender and receiver peripheries, spectral algorithms are proposed to identify the CP structure in directed weighted networks. The algorithm stands out with statistical guarantees, ensuring the identification of sender and receiver peripheries with overwhelming probability. Additionally, the methods scale effectively for expansive directed networks.

EO022 Room 105 MODERN TOPICS IN MACHINE LEARNING Chair: Yiming Ying

E0587: **Learning theory of spectral algorithms under covariate shift**
Presenter: **Zheng-Chu Guo**, Zhejiang University, China
In machine learning, it is commonly assumed that the training and test samples are drawn from the same underlying distribution. However, this assumption may not always hold true in practice. A scenario is delved into where the distribution of the input variables (also known as covariates) differs between the training and test phases. This situation is referred to as a covariate shift. To address the challenges posed by covariate shift, various techniques have been developed, such as importance weighting, domain adaptation, and reweighting methods. The focus is on the weighted spectral algorithm. Under mild conditions imposed on the weights, it is demonstrated that this algorithm achieves satisfactory convergence rates.

E0619: **Capacity dependent analysis for functional online learning algorithms**
Presenter: **Xin Guo**, The University of Queensland, Australia
Co-authors: Zheng-Chu Guo, Lei Shi
The purpose is to provide convergence analysis of online stochastic gradient descent algorithms for functional linear models. Adopting the characterizations of the slope function regularity, the kernel space capacity, and the capacity of the sampling process covariance operator, significant improvement in the convergence rates is achieved. Both prediction problems and estimation problems are studied, where the capacity assumption is shown to alleviate the saturation of the convergence rate as the regularity of the target function increases. It is shown that with a properly selected kernel, capacity assumptions can fully compensate for the regularity assumptions for prediction problems (but not for estimation problems). This demonstrates the significant difference between the prediction problems and the estimation problems in functional data analysis.

E0666: **Statistical learning theory for stochastic compositional gradient methods**
Presenter: **Yiming Ying**, University of Sydney, Australia
Many machine learning tasks can be framed as stochastic compositional optimization (SCO) problems, including reinforcement learning, AUC maximization, and meta-learning, where the objective function involves a nested composition associated with an expectation. While numerous studies have focused on the convergence behavior of SCO algorithms, little attention has been given to understanding their generalization, i.e., how these learning algorithms generalize from training to test examples. The stability and generalization analysis of stochastic compositional gradient descent algorithms is delved into within the framework of statistical learning theory. Specifically, the compositional uniform stability of two prominent stochastic compositional gradient descent algorithms is examined, namely SCGD and SCSC. In contrast to existing literature, the analysis yields dimension-independent bounds that underscore the critical dependence of the SCO algorithm's generalization on two key factors: the variance of the gradient of the inner function and the convergence of the moving average term to the true inner function. Moreover, the findings exemplify the intricate interplay between statistical and computational behaviors in stochastic optimization for machine learning.

EO133 Room 106 ADVANCES IN COMPLEX DATA ANALYSIS Chair: Daren Wang

E0704: **Changepoint estimation and inference in functional linear regression models**
Presenter: **Haotian Xu**, University of Warwick, United Kingdom
The focus is on identifying and estimating structural changes in the slope functions within functional linear regression models. Estimators are provided that are minimax optimal for this purpose, and the limiting distributions of these estimators are studied. The theoretical results are proved under general conditions, which allow for both covariates and noise sequences with heavy-tails and temporal dependence.

E1055: **Risk bounds for quantile additive trend filtering**
Presenter: **Zhi Zhang**, UCLA, United States
Co-authors: Kyle Ritscher, Oscar Hernan Madrid Padilla
Risk bounds for quantile additive trend filtering are investigated, a method gaining increasing significance in the realms of additive trend filtering and quantile regression. The constrained version of quantile trend filtering within additive models is investigated, considering fixed and growing input dimensions. In the fixed dimension case, an error rate that mirrors the non-quantile minimax rate is discovered for additive trend filtering, featuring the main term $n^{-2r/(2r+1)}$. In scenarios with a growing input dimension (d), quantile additive trend filtering introduces a polynomial factor of $d^{(2r+2)/(2r+1)}$. This aligns with the non-quantile variant, featuring a linear factor d, particularly pronounced for larger r values. Additionally, a practical algorithm is proposed for implementing quantile trend filtering within additive models using dimension-wise backfitting. Experiments are conducted with evenly spaced data points or data that samples from a uniform distribution in the interval $[0,1]^d$, applying distinct component functions and introducing noise from normal and heavy-tailed distributions. The findings confirm the estimator's convergence as n increases and its superiority, particularly in heavy-tailed distribution scenarios. These results deepen the understanding of additive trend filtering models in quantile settings, offering valuable insights for practical applications and future research.

E0426: **Do uncertain rewards still work: The lottery case in live streaming commerce**
Presenter: **Lin Qiu**, Southern University of Science and Technology, China
Co-authors: Dan Ding
Live streaming is now a booming e-commerce platform globally. Real-time comments play a vital role in engaging users during live streams, with two main types: product-related and non-product-related comments. Product-related comments clarify product details like size and price, aiding viewers' purchase decisions. Non-product comments include greetings, casual chats, and interactions between broadcasters and viewers. These comments are not directly related to products and are mainly used to interact with broadcasters. Lottery, as a type of popular uncertain reward, has been widely used in live streams to facilitate broadcaster-viewer interactions. The aim is to analyze the impact of two types of lotteries (physical goods lottery and social-oriented lottery) on broadcaster-viewer interactions through real-time comments and subsequent product sales in live streams. It is observed that an increase in physical goods lotteries results in a higher proportion of product-related comments compared to non-product-related ones, offering viewers more informative support for purchase decisions. Conversely, social-oriented lotteries generate more non-product-related comments than product-related ones, diluting informative product comments. This interference in purchase decision-making could negatively impact sales. The findings offer valuable insights for broadcasters and platform designers when formulating lottery strategies.

EO057 Room 108 TRUSTWORTHY AND EFFICIENT STATISTICAL LEARNING Chair: Yao Li

E0310: **Textual backdoor attack detection**
Presenter: **Yao Li**, University of North Carolina at Chapel Hill, United States
Co-authors: Xinglin Li, Xianwen He, Minhao Cheng
Backdoor attacks pose a stealthy threat to deep neural network-based classifiers. Such attacks introduce a backdoor into the model by contaminating part of the training data with carefully chosen triggers. The victim model then erroneously predicts inputs containing the same triggers as a certain class. In the field of natural language processing (NLP), the study on defenses against backdoor attacks is insufficient. To the best of knowledge, existing NLP defense methods primarily target special token-based triggers, leaving syntax-based triggers unaddressed. To fill this gap, a novel defense algorithm is proposed that effectively counters syntax-based as well as special token-based backdoor attacks. The algorithm replaces semantically meaningful words in sentences with entirely different ones but preserves the syntactic templates or special tokens and then compares the predicted labels before and after the substitution to determine whether a sentence contains triggers. Experimental results confirm the algorithm's performance against these two types of triggers, offering a comprehensive defense strategy for model integrity.

E0321: **High dimensional general linear hypotheses under a spiked covariance model**
Presenter: **Haoran Li**, Auburn University, United States
Co-authors: Debashis Paul, Jie Peng, Alexander Aue
The problem of testing linear hypotheses under a multivariate regression model is considered with a high-dimensional response and spiked noise covariance. The proposed family of tests consists of test statistics based on a weighted sum of projections of the data onto the estimated latent factor directions, with the weights acting as the regularization parameters. The asymptotic normality of the test statistics is established under the null hypothesis. The power characteristics of the tests is also established, and a data-driven choice of the regularization parameters is proposed under a family of local alternatives. The performance of the proposed tests is evaluated through a simulation study. Finally, the proposed tests are applied to the Human Connectome Project data to test for the presence of associations between volumetric measurements of the human brain and behavioral variables.

E0836: **Towards classification of covariance matrices via Bures-Wasserstein-based machine learning**
Presenter: **Jingyi Zheng**, Auburn University, United States
Co-authors: Yuyan Yi, Shu-Chin Lin, Michael Zirpoli
In the realm of machine learning, positive semi-definite (PSD) matrices emerge as crucial entities, especially in handling challenges posed by high-dimensional data. Three novel machine learning algorithms are introduced, tailored for the classification of PSD matrices on the manifold equipped with Bures-Wasserstein (BW) metric. Leveraging BW distance, barycenter estimation, and projection algorithms, the approach distinguishes itself from classical Euclidean methods by integrating the geometry of the Riemannian manifold where PSD matrices reside. In contrast to the prevalent Affine-Invariant (AI) Riemannian manifold analysis, the BW manifold analysis obviates the need for matrix regularization and significantly enhances computational efficiency. Additionally, a novel BW distance-based kernel function is proposed, which is further used in the kernel Support Vector Machine. Through comprehensive evaluations across multiple real datasets characterized by varying dimensions and matrix quantities, the findings underscore the exceptional performance of the proposed machine learning algorithms. The efficacy of BW metric-based machine learning methodologies in advancing the classification of PSD matrices is emphasized.

EO037 Room 109 MODERN STATISTICAL METHODS IN MACHINE LEARNING AND ECONOMICS Chair: Ruoqing Zhu

E0333: **Policy learning with continuous actions under unmeasured confounding**
Presenter: **Ruoqing Zhu**, University of Illinois at Urbana-Champaign, United States
Co-authors: Yuhan Li, Eugene Han, Wenzhuo Zhou, Zhengling Qi, Yifan Cui

In the field of reinforcement learning applied to personalized medicine, unmeasured confounding variables often hinder the optimization of treatment policies, particularly in offline settings. While most existing methods focus on off-policy evaluation (OPE), they are generally not directly suited for learning optimal policies. For example, common assumptions that the behavior policy depends solely on unobserved state variables can be practically violated in real-world medical scenarios. A novel identification framework is introduced to estimate policy values accurately. This is achieved by identifying a set of variables that are not involved in policy determination but can potentially affect the reward. By appropriately constructing bridge functions, an optimal policy is learned based on observed states, thereby enabling practical implementation. The framework additionally tackles the dose-finding problem in personalized medicine by considering a continuous action space. The asymptotic properties of the proposed estimators are also explored under suitable conditions. The method is applied to a study of romantic relationships from Germany.

E0832: **Estimation of average treatment effect for survival outcomes with continuous treatment in observational studies**
Presenter: **Qi Zheng**, University of Louisville, United States

In healthcare research, where extensive observational data such as claims data and electronic records are readily available, researchers often seek to investigate both the treatment effect and the pathway of that effect. While recent literature on causal effects in survival analyses primarily focuses on binary or multiple treatment settings, studies involving continuous treatment settings are rarely explored. The estimation of the average treatment effect (ATE) of continuous treatment is explored on time-to-event outcomes by addressing multiple confounding factors and considering censoring observations. The ATE is proposed to estimate using the accelerated failure time marginal structural model (AFT-MSM), incorporating the inverse probability of treatment weighting (IPTW) method along with censoring weights. The IPTW method is designed to mitigate the influence of confounding variables on treatment assignment while censoring weights and addressing potential biases arising from censored observations. Extensive simulation studies have demonstrated the effectiveness of the proposed method. The proposed methodology is applied to investigate the impact of blood lead levels on the time to death among older individuals in the United States, utilizing data from the NHANES III survey dataset.

E1072: **Byzantine-robust distributed learning under heterogeneity via convex hull search**
Presenter: **Zhao Chen**, Fudan University, China

In modern massive data modelling, distributed learning plays a critical role in enhancing scalability, efficiency and privacy protection. The heterogeneity and robustness of a distributed learning algorithm are key aspects related to the accuracy and reliability of learning results. Under the common framework of statistical learning, the convex hull search algorithm is proposed, which has four main advantages: fast convergence, high accuracy, adjustable robustness, and tuning friendliness. The corresponding convergence and asymptotic normality result for the CHS algorithm are established, which shows its adaptability to data heterogeneity. The algorithm for regression and clustering tasks is exemplified through synthetic data. Furthermore, real energy consumption data is implemented for Gaussian process regression hyperparameters optimization. Existing numerical results confirm the superiority and exhibit the wide applicability of the algorithm

EO021 Room 110 LARGE RANDOM MATRICES AND THEIR APPLICATIONS Chair: Zeng Li

E0170: **A leave-one-out approach to approximate message passing**
Presenter: **Xiaocong Xu**, Hong Kong University of Science and Technology, Hong Kong
Co-authors: Zhigang Bao, Qiyang Han

The aim is to present a non-asymptotic leave-one-out representation for approximate message passing (AMP) iterates, applicable to a broad class of Gaussian random matrix models with general variance profiles. In contrast to the typical AMP theory that describes the empirical distributions of the AMP iterate via a low-dimensional state evolution, the leave-one-out representation yields an intrinsically high-dimensional state evolution formula which provides non-asymptotic characterizations for the possibly heterogeneous, entrywise behaviour of the AMP iterate under the prescribed random matrix models. To exemplify some distinct features of our AMP theory in applications, the precise stochastic behaviour of the ridge estimator is analyzed for independent and non-identically distributed observations in the context of regularized linear estimation, whose covariates exhibit general variance profiles. It is found that its finite-sample distribution is characterized via a weighted ridge estimator in a heterogeneous Gaussian sequence model. Notably, in contrast to the i.i.d. sampling scenario, the effective noise and regularization are now full-dimensional vectors determined via a high-dimensional system of equations. The method of proof differs significantly from the master conditioning approach. It relies on an inductive method that sheds light on the intricate cancellation scheme for the trace function of certain random matrix recursions associated with the AMP.

E0171: **Robust estimation of number of factors in high dimensional factor modeling via Spearman's rank correlation matrix**
Presenter: **Zeng Li**, Southern University of Science and Technology, China

Determining the number of factors in high dimensional factor modelling is essential but challenging, especially when the data are heavy-tailed. A new estimator is introduced based on the spectral properties of Spearman's rank correlation matrix under the high dimensional setting, where both dimension and sample size tend to infinity proportionally. The estimator applies to scenarios where the common factors or idiosyncratic errors follow heavy-tailed distributions. It is proven that the proposed estimator is consistent under mild conditions. Numerical experiments also demonstrate the superiority of the estimator compared to existing methods, especially for the heavy-tailed case.

E0487: **Convergence rate to the Tracy-Widom laws for the largest eigenvalue of Wigner matrices**
Presenter: **Yuanyuan Xu**, AMSS, CAS, China

A quantitative Tracy-Widom law is discussed for the largest eigenvalue of Wigner matrices (with a variance profile). More precisely, it is proven that the fluctuations of the largest eigenvalue of a Wigner matrix of size N converge to its Tracy-Widom limit at a rate nearly $N^{-1/3}$, as N tends to infinity. The same result also holds for sample covariance matrices.

EO233 Room 111 ADVANCES ON BIOSTATISTICS Chair: Hua Shen

E0235: **A Bayesian approach to jointly modelling epidemic and behavioral dynamics**
Presenter: **Rob Deardon**, University of Calgary, Canada

One of the many difficulties in modelling epidemic spread is caused by behavioral change in the underlying population. This can be a major public health issue since, as seen during the COVID-19 pandemic, behavior in the population can change drastically as infection levels vary, both due to government mandates and personal decisions. Such changes in the underlying population result in major changes in the transmission dynamics of the disease, making the modelling challenging. However, these issues arise in agriculture and public health, as changes in farming practices are often observed as disease prevalence changes. A model formulation is proposed wherein time-varying transmission is captured by the level of alarm in the population and specified as a function of recent epidemic history. The alarm function itself can also vary dynamically, allowing for phenomena such as lockdown fatigue. The model is set in a data-augmented Bayesian framework as epidemic data are often only partially observed and can be utilized prior information to help with parameter identifiability. The identifiability of the population alarm is investigated across a wide range of scenarios, using both parametric functions and non-parametric Gaussian processes and splines. The benefits and utility of the proposed approach are illustrated through applications of COVID-19 and Ebola disease.

E0937: **An improved MC-SIMEX method**
Presenter: **Lili Yu**, Georgia Southern University, United States
Co-authors: Varadan Sevilimedu

The problem of misclassification in covariates is ubiquitous in medical data and often leads to biased estimates. The misclassification simulation extrapolation method (MC-SIMEX) is a popular approach to correct this bias. However, it utilizes an approximated extrapolation function derived from the simulated data, which not only reduces the reliability of the estimator but also increases computation time. The aim is to propose an improved MC-SIMEX estimator for generalized linear models, in which the closed-form exact extrapolation function is derived, thus addressing the challenges in the original MC-SIMEX method. Simulations demonstrate that the newly proposed method outperforms the original MC-SIMEX approach in terms of bias correction. Additionally, a real data example is provided to illustrate the effectiveness of the proposed method.

E1102: **A two-stage method for integrating probability and non-probability samples with misclassified covariates**
Presenter: **Hua Shen**, University of Calgary, Canada

Probability samples in clinical research often lack critical, accurately measured variables, while non-probability samples provide detailed data but are not representative and include measurement errors. The aim is to present a novel two-stage estimation technique to integrate these sample types, addressing misclassification in categorical covariates using a latent-variable framework without validation data. Through simulations, the method demonstrates superior accuracy over traditional approaches. Its effectiveness is validated on a real dataset to improve inference quality in studies involving misclassified covariates.

EO313 Room 212 MODERN MULTIVARIATE DATA: METHODS, MODELS, AND MORE Chair: Joshua Cape

E0308: **Variable target scalable particle filter**
Presenter: **Ning Ning**, Texas A&M University, United States

The challenge of tracking a variable number of interacting targets is addressed. The primary goal is to accurately detect targets entering and leaving the scene while maintaining a precise trajectory record for each target throughout their presence. Developing methods that effectively handle this complexity is vital for scenarios involving the continuous tracking of numerous interacting targets. To tackle this problem, the variable target scalable particle filter (VTSPF) is introduced within an online learning framework. VTSPF efficiently tracks multiple moving targets exhibiting complex interactions. Importantly, it demonstrates scalability in both spatial and temporal dimensions.

E0460: **Disentangled adversarial flow with ensemble learning for multi-source brain connectome analysis**
Presenter: **Meimei Liu**, Virginia Tech, United States
Co-authors: Yixin Chen, Zhengwu Zhang, Xin Xing

Understanding the brain's structural connectome and its role in cognitive functions has been advanced by diffusion magnetic resonance imaging. However, the heterogeneity across multiple neuroimaging studies, combined with limited labeled samples in specialized cohorts, poses significant challenges in developing accurate predictive models for cognitive abilities. A novel disentangled adversarial flow (DAF) model is introduced, leveraging large-scale datasets to enhance predictions in smaller neuroimaging studies. DAF generates domain-invariant brain connectome representations using a bidirectional architecture and a kernel-based measure to minimize domain label dependence. An ensemble DAF regression framework integrates multiple data sources, reducing information loss in multi-domain data. Validated on the adolescent brain cognitive development (ABCD) study, the human connectome project (HCP), and the Alzheimer's disease neuroimaging initiative (ADNI), DAF shows reduced discrepancies across domains and superior predictive accuracy with limited target domain data.

E0757: **On the validity of conformal prediction for network data under non-uniform sampling**
Presenter: **Robert Lunde**, Washington University in St Louis, United States

The properties of conformal prediction for network data are studied under various sampling mechanisms that commonly arise in practice but often result in a non-representative sample of nodes. These sampling mechanisms are interpreted as selection rules applied to a superpopulation, and the validity of conformal prediction conditional is studied based on an appropriate selection event. It is shown that the sampled subarray is exchangeable conditional on the selection event if the selection rule satisfies a permutation invariance property and a joint exchangeability condition holds for the superpopulation. The result implies the finite-sample validity of conformal prediction for certain selection events related to ego networks and snowball sampling. It is also shown that when data are sampled via a random walk on a graph, a variant of weighted conformal prediction yields asymptotically valid prediction sets for an independently selected node from the population.

EO205 Room 204 INTEGRATIVE APPROACHES IN BIOMEDICAL DATA ANALYSIS Chair: Xinlei Wang

E0209: **Mutual information for detecting multi-class biomarkers when integrating multiple omics studies**
Presenter: **Jian Zou**, Chongqing Medical University, China

Biomarker detection is crucial in biomedical research. While integrating multiple omics studies improves the statistical power and robustness of results, current methods struggle with multi-class scenarios (e.g., various disease subtypes or treatments). Mutual information concordance analysis (MICA), a statistical framework for identifying biomarkers with consistent multi-class expression patterns across multiple studies, is introduced. MICA employs information theory to test and detect these biomarkers globally, followed by a post hoc analysis to pinpoint studies with concordant patterns. Extensive simulations and two real applications demonstrate the superior performance of the proposed method.

E0212: **Bayesian model selection for high-dimensional data integration**
Presenter: **Yuehua Wu**, York University, Canada

Data integration problems are considered when correlated data are collected from multiple platforms. Each platform has linear relationships between the responses and a collection of predictors. The linear models are extended to include random errors from a much wider family of sub-Gaussian and sub-exponential distributions. The goal is to select important predictors across multiple platforms, where the number of predictors and the number of observations both increase to infinity. The marginal densities of the responses obtained from different platforms are combined to form a composite likelihood and propose a model selection criterion based on Bayesian composite posterior probabilities. Under some regularity conditions, it is proven that the model selection criterion is consistent with divergent true model sizes. A Monte Carlo Markov Chain algorithm is implemented for the model selection approach. Simulation results and a real data example are further presented.

E0870: **Analysis of disease prevalence: A clustering perspective that incorporates prior information**
Presenter: **Chenjin Ma**, Beijing University of Technology, China

The analysis of disease prevalence is of critical importance in biomedical research. The collective analysis of multiple diseases, significantly different from individual disease analysis, can provide valuable additional insights. A critical limitation of the existing analysis is that there is a lack of attention to prior information, which has been accumulated through many studies and can be valuable, especially when there are a large number of diseases, but the number of prevalence values for a specific disease is limited. The functional clustering analysis of disease prevalence trends is conducted. A novel approach based on the penalized fusion technique is developed to incorporate prior information mined from published articles. It is innovatively designed to take into account that such information may not be fully relevant or correct. Rigorous statistical and computational investigations are conducted. In the analysis of data from Taiwan NHIRD (National Health Insurance Research Database), new and interesting findings that differ from the existing ones are made.

EO319 Room 207 RECENT DEVELOPMENTS IN CAUSAL INFERENCE Chair: Wei Luo

E0303: **Inference of continuous treatment effects in large-scale observational data**
Presenter: **Shujie Ma**, University of California-Riverside, United States

Recent advances in technology have created numerous large-scale datasets in observational studies, which provide unprecedented opportunities for evaluating the effectiveness of various treatments. Under the condition of unconfounded treatment assignment, most existing methods rely on a parametric or a nonparametric modeling method for estimating the propensity score or the outcome regression functions. The parametric approach lacks robustness as it suffers from the model misspecification problem. Conventional nonparametric estimation methods suffer from the curse of dimensionality when the dimension of confounders is large. A new method is introduced for estimating and inferring continuous treatment effects in large-scale observational data. The nuisance function is estimated by artificial neural networks. The proposed method takes full advantage of the large sample size of large-scale data and provides effective protection against misspecification bias. The theoretical properties established are presented for the proposed estimator, and the method is illustrated through simulation studies and real data applications.

E0431: **Policy learning with distributional welfare**
Presenter: **Yifan Cui**, Zhejiang University, China
Co-authors: Sukjin Han

Optimal treatment allocation policies that target distributional welfare are explored, and an optimal policy is proposed that allocates the treatment based on the conditional quantile of individual treatment effects (QoTE). Depending on the choice of the quantile probability, this criterion can accommodate a policymaker who is either prudent or negligent. The challenge of identifying the QoTE lies in its requirement for knowledge of the joint distribution of the counterfactual outcomes, which is generally hard to recover even with experimental data. Therefore, minimax policies are introduced that are robust to model uncertainty. A range of identifying assumptions can be used to yield more informative policies. For both stochastic and deterministic policies, the asymptotic bound is established on the regret of implementing the proposed policies. In simulations and two empirical applications, optimal decisions based on the QoTE are compared with decisions based on other criteria. The framework can be generalized to any setting where welfare is defined as a functional part of the joint distribution of potential outcomes.

E0585: **Selection of mediators and dependence structure for high-dimensional causal mediation analysis**
Presenter: **Yeying Zhu**, University of Waterloo, Canada
Co-authors: Lijia Wang, Richard Cook

Causal mediation analysis examines the potential causal pathways between an exposure variable and outcome through intermediate variables with the goal of estimating direct and indirect effects. In practice, intermediate variables may be high-dimensional, in which case one may first aim to identify the true mediators among them. The dependence structure among mediators may then be studied with the goal of identifying a simple, sufficient structure. A two-stage penalized estimation procedure is proposed to meet these goals. The first stage involves selecting mediators by identifying non-zero indirect effects via a penalized regression. The second stage aims to simplify the correlation structure among selected mediators, enabling the estimation of individual, grouped or joint effects. Through the transformation of variables, the correlation selection problem can be reformulated as a standard LASSO problem. The two stages can be performed jointly or sequentially, and the performance of each implementation is studied through simulation studies. Finally, the proposed approach is applied to a psychiatry study in which the aim is to identify methylation loci that mediate the causal effect of childhood trauma on adult stress levels.

EO119 Room 209 ADVANCING STATISTICAL INFERENCE IN HIGH DIMENSIONAL AND COMPLEX DATA Chair: Xin Xing

E0550: **Sequential canonical variate regression for unified correlation analysis and outcome prediction of multi-omics data**
Presenter: **Chongliang Luo**, Washington University in St Louis, United States

When integrating multiple sets of data in biomedical research, the objectives are often two-fold: to learn the correlation structure between the sets and to optimize the prediction of clinical outcomes. The correlation structure can be achieved by canonical correlation analysis (CCA), and the extracted canonical variates (CVs) can be used to predict the outcome, hence the canonical variate regression (CVR). On the other hand, besides the shared joint structure across sets, set-specific structures for individual sets may also exist, which are predictive. This joint and individual structure motivates us to develop a sequential CVR (CVR-seq) method, which can further improve the prediction. The CVR-seq adaptively extracts layers of CVs that balance the correlation between sets and the prediction of the interested outcome. The CVR-seq method provides a flexible integration of multiple sets of data as well as interpretable outcome prediction. The usage of CVR-seq is demonstrated by simulation and the METABRIC study that integrates copy number alteration (CNA) and gene expression (GEXP) data to predict breast cancer risk-free survival.

E0624: **Model-X conditional knockoffs and conditional randomization tests using Gaussian graphical models**
Presenter: **Dongming Huang**, National University of Singapore, China
Co-authors: Lucas Janson

The model-X framework provides provable non-asymptotical error control on variable selection and conditional independence testing. It has no restrictions or assumptions on the dimensionality of the data or the conditional distribution of the response given the covariates. To relax the

requirement of the model-X framework that the distribution of the covariate samples is precisely known, it is proposed to construct knockoffs by conditioning on sufficient statistics when the distribution is known up to a parametric model with as many as Cnp parameters, where p is the dimension, n is the number of covariate samples (including unlabeled samples if available), and C is a constant. It is demonstrated how this idea can be implemented in Gaussian graphical models, and the new approach remains powerful under the weaker assumption. It is shown how such conditioning can be extended to constructing a conditional randomization test for testing conditional independence between the response and a subset of the covariates.

E0613: **GLSS of sample correlation matrix**
Presenter: **Xiao Han**, University of Science and Technology of China, China
A sample correlation matrix is an important random matrix used in high-dimensional data analysis. Motivated by this, spectral properties of large dimensional sample correlation matrices are studied by introducing generalized linear spectral statistics (GLSS). In particular, a joint CLT is established for this statistic related to various factors, which generalizes the result in a prior study. Meanwhile, the difference in GLSS is discovered between sample covariance matrices and sample correlation matrices. Extensive simulations verify the theoretical results.

EO072 Room 210 RECENT ADVANCES IN STATISTICAL METHODS FOR COMPLEX DATA ANALYSIS Chair: Zeya Wang

E1003: **Robust covariance matrix estimation for high-dimensional compositional data with application to sales data analysis**
Presenter: **Danning Li**, Northeast Normal University, China
Compositional data arises in a wide variety of research areas when some form of standardization and composition is necessary. Estimating covariance matrices is of fundamental importance for high-dimensional compositional data analysis. However, existing methods require the restrictive Gaussian or sub-Gaussian assumption, which may not hold in practice. The aim is to propose a robust composition-adjusted thresholding covariance procedure based on Huber-type M-estimation to estimate the sparse covariance structure of high-dimensional compositional data. A cross-validation procedure is introduced to choose the tuning parameters of the proposed method. Theoretically, by assuming a bounded fourth-moment condition, the rates of convergence and signal recovery property for the proposed method are obtained, and the theoretical guarantees are provided for the cross-validation procedure under the high-dimensional setting. Numerically, the effectiveness of the proposed method is demonstrated in simulation studies and also a real application to sales data analysis.

E1052: **Robust rank canonical correlation analysis for multivariate survival data**
Presenter: **Di He**, Nanjing University, China
Co-authors: Yong Zhou, Hui Zou
Canonical correlation analysis (CCA) is widely applied in statistical analysis of multivariate data to find associations between two sets of multidimensional variables. However, CCA often cannot be used directly for survival data or their monotone transformations, owing to right-censoring in the data. A new robust rank CCA (RRCCA) method is proposed based on Kendall's τ correlation and adjusts it to deal with multivariate survival data without requiring any model assumptions. Owing to the nature of rank correlation, the RRCCA is invariant against monotone transformations of the data. The estimation consistency of the RRCCA approach is established under weak conditions. Simulation studies demonstrate the superior performance of the RRCCA in terms of estimation accuracy and empirical power. Lastly, the proposed method is demonstrated by applying it to Stanford heart transplant data.

E1069: **Estimation of out-of-sample Sharpe ratio for high dimensional portfolio optimization**
Presenter: **Weichen Wang**, The University of Hong Kong, Hong Kong
Co-authors: Xuran Meng, Yuan Cao
Portfolio optimization aims at achieving a portfolio with good out-of-sample performance, typically measured by the out-of-sample Sharpe ratio. However, due to in-sample optimism, it is inappropriate to use the in-sample estimated covariance to evaluate the out-of-sample Sharpe, especially in high-dimensional settings. A novel method is proposed to estimate the out-of-sample Sharpe ratio using only in-sample data based on random matrix theory. Specifically, the classical framework of Markowitz mean-variance optimization is considered with a known mean vector under the high dimensional regime of p/n goes to $c > 0$, where p is the portfolio dimension and n is the number of samples or time points. Correcting the sample covariance is proposed by a regularization matrix, and a consistent Sharpe ratio estimator is provided. The new estimator works well under either of the three conditions: (1) bounded covariance spectrum, (2) arbitrary number of diverging spikes when $c < 1$, and (3) fixed number of diverging spikes when $c \geq 1$. The results can be extended to the global minimum variance portfolio and the construction of the out-of-sample efficient frontier. The effectiveness of the approach is demonstrated through numerical experiments and real data. Results highlight the potential of the new methodology as a powerful tool for portfolio managers to estimate the out-of-sample Sharpe and to decide the optimal parameter to construct their portfolios.

EO042 Room 307 STATISTICS FOR NON-EUCLIDEAN DATA Chair: Lujia Bai

E0748: **Inverse regression for spatially distributed functional data**
Presenter: **Suneel Babu Chatla**, University of Texas at El Paso, United States
Co-authors: Ruiqi Liu
Spatially distributed functional data are prevalent in many statistical applications. Given their complex and high-dimensional nature, functional data often require dimension-reduction methods to extract meaningful information. Inverse regression is one such approach that has become very popular in the past two decades. We study the inverse regression in the framework of functional data observed at irregularly positioned spatial sites. The functional predictor is the sum of a spatially dependent functional effect and a spatially independent functional nugget effect, while the relation between the scalar response and the functional predictor is modeled using the inverse regression framework. The domain-expanding infill (DEI) framework is discussed for spatial asymptotics, which is a mix of the traditional expanding domain and infill frameworks. The DEI framework overcomes the limitations of traditional spatial asymptotics in the existing literature. Under this unified framework, asymptotic theory is developed, and conditions that are necessary for the estimated eigen-directions to achieve optimal rates of convergence are identified. The asymptotic results include pointwise and L2 convergence rates.

E0404: **Testing for white noises in multivariate locally stationary functional time series**
Presenter: **Lujia Bai**, Tsinghua University, China
Co-authors: Weichi Wu, Holger Dette
Multivariate locally stationary functional time series provides a flexible framework for modelling complex data structures exhibiting both temporal and spatial dependencies while allowing for a time-varying data-generating mechanism. A specialized Portmanteau test tailored for assessing white noise assumptions is introduced for multivariate locally stationary functional time series without dimension reduction. The Gaussian approximation result is derived from the Gaussian approximation result for the kernel-weighted second-order functional time series, which is of independent interest. A simple bootstrap procedure is proposed to implement the test where the limiting distribution can be non-standard or even does not exist.

Through theoretical analysis and simulation studies, the efficacy and adaptability of the proposed portmanteau test are demonstrated in detecting departures from white noise assumptions in multivariate, locally stationary functional time series.

E0957: **Robust functional principal component analysis for non-Euclidean random objects**
Presenter: **Jiazhen Xu**, Australian National University, Australia
Co-authors: Andrew Wood, Tao Zou
Functional data analysis offers a diverse toolkit of statistical methods tailored for analyzing samples of real-valued random functions. Recently, samples of time-varying random objects, such as time-varying networks, have been increasingly encountered in modern data analysis. These data structures represent elements within general metric spaces that lack local or global linear structures, rendering traditional functional data analysis methods inapplicable. Moreover, the existing methodology for time-varying random objects does not work well in the presence of outlying objects. The aim is to propose a robust method for analysing time-varying random objects. The method employs pointwise Frechet medians and then constructs pointwise distance trajectories between the individual time courses and the sample Frechet medians. This representation effectively transforms time-varying objects into functional data. A novel, robust approach to functional principal component analysis based on a Winsorized U-statistic estimator of the covariance structure is introduced. The proposed robust analysis of these distance trajectories is able to identify key features of time-varying objects and is useful for downstream analysis. To illustrate the efficacy of the approach, numerical studies focusing on dynamic networks are conducted. The results indicate that the proposed method exhibits good all-round performance and surpasses the existing approach in terms of robustness.

EO178 Room 313 RECENT ADVANCES IN STOCHASTIC MODELING Chair: Shuyang Bai

E0396: **Average and conditional inward and outward spillovers of one unit's treatment under network interference**
Presenter: **Fei Fang**, Yale University, United States
Co-authors: Laura Forastiere, Edoardo Airoldi
In a connected social network, users may have varying levels of influence on others when they receive interventions. For example, giving an advertisement to a more influential person can have, on average, a greater impact on others' purchase decisions. Understanding and evaluating these effects can provide valuable insights for various applications, such as targeting strategies in marketing and behavioral interventions in public health. Under a partial interference assumption, influence effects are defined in two ways: i) the inward average spillover effect on a unit's outcome of a neighbors treatment, and ii) the outward average spillover of a unit's treatment on their neighbors outcomes. The comparison is investigated between the two causal effects in directed networks with different properties, including the conditions under which they are equivalent. Additionally, Horvitz-Thompson estimators are developed to assess both effects, on average and conditioning, on categorical covariates, and weighted least square estimators for these effects conditioning on continuous covariates. Design-based variance estimators are derived, and the consistency and asymptotic normality are established. Through simulations, the empirical performance of the proposed estimators is verified. Finally, the approach is employed to investigate the inward and outward average and conditional spillover effects of an information session on the adoption of weather insurance among rice farmers in China.

E0820: **Breuer-Major theorems for Hilbert space-valued random variables**
Presenter: **Marie Duker**, FAU Erlangen, Germany
Nonlinear functionals of Gaussian random variables, also called Gaussian subordinated variables, provide a flexible model in time series analysis. Under suitable conditions on the functionals and the correlation structure of the latent Gaussian variables, one can derive a central limit theorem (CLT). The concept of Gaussian subordination is studied in general Hilbert spaces, allowing the latent Gaussian variables and the functions to take values in suitable Hilbert spaces. To prove a CLT for Hilbert space-valued subordinated processes, techniques are employed from Malliavin calculus, a powerful tool in modern stochastic analysis. In a series of examples, the derived CLT is emphasized to recover limit theorems for a wide array of statistics relevant to functional data analysis and present novel applications to the theory of operator neural networks.

E1075: **On order selection for multivariate extremes via clustering**
Presenter: **He Tang**, University of Georgia, United States
Co-authors: Shuyang Bai, Shiyuan Deng
The estimation of multivariate extreme models with a discrete spectral measure is investigated using clustering techniques. The primary innovation involves devising a method for selecting the appropriate order that not only consistently identifies the true order in theory but also has a straightforward and easy implementation in practice. Specifically, an extra penalty term is introduced to the well-known simplified average silhouette width, which penalizes small cluster sizes and minimal dissimilarities between cluster centers. Consequently, a consistent method is provided for the order of a max-linear factor model, where a typical information-based approach is not viable due to the absence of likelihood.

EO121 Room 405 ANOTHER LOOK AT FINANCIAL ECONOMETRICS Chair: Shaoran Li

E0445: **Do peer characteristics explain returns: An aggregation approach**
Presenter: **Shuyi Ge**, University of Nankai, China
The investigation into firm characteristics that matter most has garnered substantial empirical interest. Yet, forecasting cross-sectional stock returns based on traits of other firms remains less explored. Leveraging economic connections among firms, two novel stock-level metrics are introduced: the peer index (PI) and the peer-deviation index (PDI). PI gauges peer relative strength and growth prospects, while PDI signifies a firm's position within its peer group. Evidence demonstrates that both metrics reliably and independently forecast future returns in samples excluding microcaps, supported by high t-ratios and alphas. PI's predictability arises from gradual responses to nuanced growth signals, while PDI identifies within-industry underperformers with overlooked strong fundamentals.

E0519: **Conformal prediction for interval outcomes**
Presenter: **Weiguang Liu**, UCL, United Kingdom
Interval censoring often occurs in economic data, leading to partial identification. Predicting outcomes that are interval censored is crucial in many practical scenarios. A technique is suggested to build prediction sets for such interval outcomes by leveraging a description of the precisely identified region of conditional cumulative distribution functions (CDF) combined with conformal inference. This approach ensures finite-sample coverage guarantees and achieves asymptotic efficiency.

E0793: **A tale of two news-implied linkages: Information structure, processing costs and cross-firm predictability**
Presenter: **Shaoran Li**, Peking University, China
News-implied linkages are decomposed into two types: leader-follower links (LF) and peer links (PE), based on people's reading and information-processing habits. It explores how the structure of information impacts processing costs and subsequently leads to market outcomes by examining momentum spillover effects via these distinct linkage types. The findings indicate that the information structure of leader-follower links is more readily comprehensible to investors than peer linkages. Empirical evidence of this is provided by demonstrating faster attention spillover from

leader to follower than among peer firms, using Baidu search data. Furthermore, it is documented that due to the lower information processing cost, information transmits through the leader-follower linkages more quickly, leading to a weaker momentum spillover effect compared to the more complex and less easily perceivable peer links.

EO071 Room 406 ADVANCES IN MIXTURE MODEL Chair: Shuchismita Sarkar

E0165: **Finite mixture of hidden Markov models for tensor-variate time series data**
Presenter: **Xuwen Zhu**, The University of Alabama, United States
Co-authors: Abdullah Asilkalkan, Shuchismita Sarkar
The need to model data with higher dimensions, such as a tensor-variate framework where each observation is considered a three-dimensional object, increases due to rapid improvements in computational power and data storage capabilities. A finite mixture of hidden Markov models for tensor-variate time series data is developed. Simulation studies demonstrate high classification accuracy for both cluster and regime IDs. To further validate the usefulness of the proposed model, it is applied to real-life data with promising results.

E0172: **On regime changes in text data using hidden Markov model of contaminated vMF distribution**
Presenter: **Shuchismita Sarkar**, Bowling Green State University, United States
Co-authors: Yingying Zhang, Xuwen Zhu, Yuanyuan Chen
A novel methodology is presented for analyzing temporal directional data with scatter and heavy tails. A hidden Markov model for contaminated von Mises-Fisher distribution is developed. The model is implemented using a backward elimination algorithm that provides additional flexibility for using it on contaminated as well as non-contaminated data. The method's utility for finding homogeneous time blocks (regimes) is demonstrated in several experimental settings and two real-life text data sets containing presidential addresses and corporate financial statements, respectively.

E0210: **On model-based clustering of multivariate categorical sequences**
Presenter: **Yingying Zhang**, Western Michigan Univesity, United States
Co-authors: Volodymyr Melnykov
Clustering algorithms designed for quantitative data have been explored extensively in the literature. However, many real-life data sets are categorical variables, including categorical sequences. The developed models for such data are designed for univariate categorical sequences. Oftentimes, data described by a single categorical sequence is not accurate enough. Observations expressed by multivariate categorical sequences can be utilized to properly reflect the dynamic nature. Currently, there is a lack of models developed for such type of data. Mixture models for multivariate categorical sequences are proposed, and the model is proven to classify observations successfully. At the same time, the proposed model also enjoys parsimonious properties compared with the traditional first-order Markov model. Both synthetic and real-life applications prove the superiority of the proposed method.

EO146 Room 408 BIOMEDICAL AND GENOMIC SCIENCES WITH PREDICTIVE AND INFERENTIAL MODELING Chair: Shan Yu

E0868: **Stability of random forests and random-forest powered prediction**
Presenter: **Yan Wang**, Wayne State University, United States
The stability property of random forests that hold even for its greedy version, which is practically implemented in popular packages, is first presented. It is then shown that, based on its stability, the random forest algorithm can be conveniently used to construct prediction intervals with guaranteed marginal coverage under mild conditions and without additional computation. Moreover, it turns out that the stability property can also be taken advantage of in the settings of active and semi-supervised learning using random forests, the guiding principle of which is bias correction rather than variance reduction, as in many existing algorithms. The general methodology presented here is anticipated to be applicable across a wide range of scenarios in medicine and engineering.

E1093: **Quantifying the global mediation effect for nonsparse high dimensional genomics mediators**
Presenter: **Chunlin Li**, Iowa State University, United States
Co-authors: Tianzhong Yang
While many existing epidemiological studies have examined associations between alcohol and cardiovascular outcomes, less has been done to explore causal biological pathways and mechanisms of the observed associations at the molecular level. To investigate this relationship, we propose a new causal measure to quantify the mediating role of molecular phenotypes, such as DNA methylation, in bridging alcohol intake and cardiovascular outcomes. The challenge of estimating this measure is two-fold. First, since alcohol consumption is associated with genome-wide changes at the molecular level, it is biologically plausible that many omics mediators with weak but collectively considerable effects are involved in the pathway; however, existing methods are plagued by inconsistency in the presence of non-sparse mediators. To address this issue, we develop a method to estimate the proposed measure in such situations consistently. Second, many epidemiological studies use case-control sampling, which introduces ascertainment bias in mediation analysis. To correct this bias, we propose a method of moment motivated by heritability estimation. Finally, a significant challenge in this research is the potential for residual confounding in observational studies, which can seriously compromise the validity of scientific findings. We will briefly discuss the approach to correct the confounding bias.

E0982: **BPED: A Bayesian basket design for pediatric trials with external data**
Presenter: **Yimei Li**, University of Pennsylvania, United States
The basket trial is a novel type of trial that evaluates one treatment in multiple indications (such as cancer types) simultaneously. One challenge of applying the basket trial design to pediatric studies is limited accrual, resulting in low statistical power. To address this issue, a Bayesian basket design for pediatric trials with external data (BPED) is proposed that performs dual information borrowing to improve the design efficiency: borrow information from the external data to the pediatric trial, and borrow information between the cancer types within the pediatric trial. BPED also accommodates potential heterogeneous treatment effects across cancer types by allowing each cancer type belonging to the sensitive or insensitive latent subgroups. The design adaptively updates the members of the subgroups based on the accumulated pediatric and external data to make go/no-go decisions for each cancer type. The simulation study shows that, compared to some existing designs, BPED yields higher power to detect the treatment effect for sensitive cancer types and maintains a desirable type I error rate for insensitive cancer types.

EC265 Room 202 STATISTICAL MODELS AND INFERENCE Chair: BaoLuo Sun

E0215: **Modelling and prediction of the wildfire data using a fractional Poisson process**
Presenter: **Sudeep Bapat**, Indian Institute of Technology Bombay, India
Modelling wildfire events has been studied in the literature using the Poisson process, which essentially assumes the independence of wildfire events. The fractional Poisson process is used to model the wildfire occurrences in California between June 2019 and April 2023 and predict the wildfire events that explain the underlying memory between these events. The method of moments and maximum likelihood estimate introduced approaches to estimate the parameters of the fractional Poisson process, which is an alternative to the method proposed in a prior study. The estimates of the fractional parameter are obtained as 0.8, proving that the wildfire events are dependent. The proposed model has reduced prediction error by 90% compared to the classical Poisson process model.

E0961: **Modeling multiple-criterion diagnoses by heterogeneous-instance logistic regression**
Presenter: **Chun-Hao Yang**, National Taiwan University, Taiwan
Mild cognitive impairment (MCI) is a prodromal stage of Alzheimer's disease (AD) that causes a significant burden in caregiving and medical costs. Clinically, the diagnosis of MCI is determined by the impairment statuses of five cognitive domains. If one of these cognitive domains is impaired, the patient is diagnosed with MCI, and if two out of the five domains are impaired, the patient is diagnosed with AD. This diagnostic procedure relates MCI/AD status modeling to multiple-instance learning, where each domain resembles an instance. However, traditional multiple-instance learning assumes common predictors among instances, but in this case, each domain is associated with different predictors. The aim is to generalize the multiple-instance logistic regression to accommodate the heterogeneity in predictors among different instances. The proposed model is dubbed heterogeneous-instance logistic regression. Two variants of the proposed model for the MCI and AD diagnoses are also derived. The proposed model is validated in terms of its estimation accuracy, latent status prediction, and robustness via extensive simulation studies. Finally, the national Alzheimer's coordinating center-uniform data set is analyzed using the proposed model, demonstrating its potential.

E1053: **On the invariance of the best linear unbiased estimators under the partitioned linear model**
Presenter: **Tatjana von Rosen**, Stockholm University, Sweden
A linear model with two sets of covariates is considered. Splitting covariates into two sets is often motivated by their nature, for example, patient-related and disease-related covariates. Depending on the nature of the covariates in the set, a partitioned fixed effects model or the mixed effects model can be formulated. The focus is on the estimation of fixed effects in one set of covariates being of primary interest. In particular, the necessary and sufficient conditions for the best linear unbiased estimator to be invariant with respect to the nature of the covariates in the other set will be derived. The proposed approach to establishing these conditions is based on solving a certain system of matrix equations. The consistency of the system of matrix equations is equivalent to the existence of the invariant BLUEs. Finally, the obtained conditions will be related to the existing ones often expressed using the linear spaces approach, for example, via column spaces of concern matrices.

EP001 Room 411 (Virtual sessions) POSTER SESSION Chair: Cristian Gatu

E0790: **Research on sleep staging based on hidden Markov model**
Presenter: **Bo Li**, Communication University of China, China
Co-authors: Xinghui Xiao, ying Wang
Sleep is an important physiological activity, and sleep staging can effectively evaluate and judge sleep structure. The purpose is to explore the implementation method of sleep staging, using 16 male subjects from the MIT-BIH multi-channel sleep database as the research object. By utilizing the temporal correlation characteristics of sleep state changes, information was extracted from electrocardiogram signals, respiratory rate, and body movement signals. A two-stage sleep staging experiment was conducted. In the first stage of sleep staging, the RR interval sequence of electrocardiogram signals was used as the observation variable, and 30 temporal, frequency domain and nonlinear features were extracted. According to the five classification results in the multi-channel sleep database, a continuous hidden Markov model was used to classify a complete sleep process. The average accuracy of the model on the test set was 64.73%. In the second stage of sleep staging, respiratory rate and body movement signals were introduced as observation variables and first-order and second-order high-order multivariate hidden Markov models were established for classification. The average recognition accuracy was 86.29%, and the high-order models had high consistency with the original annotations in distinguishing sleep and wakefulness, especially in correctly identifying sleep stages 1-3.

E0950: **Using Kendall Tau to assess the excess co-movement in time series**
Presenter: **Ying Zhang**, Acadia University, Canada
Co-authors: Will Sutherland
Kendall Tau is first introduced as an accordance measure between two stationary time series, and then two methods are discussed to estimate its variance and construct its confidence interval. Motivated by an economic phenomenon, the prices of commodities have a persistent tendency to move together; the proposal is to use the Kendall Tau confidence interval method to assess the excess co-movement between two-time series. To assess the excess dependence of two series, one needs to work with the residuals after modeling the trend and other effects, which provides the ideal conditions for our proposed confidence interval construction. The use of this method to investigate the excess co-movement of seven commodities from 1990-2020 is demonstrated.

E0959: **A data-driven new location recommendation system for sustained revenue growth in retail business**
Presenter: **Subin Jeong**, Chungnam National University, Korea, South
Co-authors: Minsu Park, Yong Hyun Um, Mingyu Go
The decision-making process regarding new location placement in retail franchising is a pivotal economic strategy where optimal placement influences revenue growth for businesses. Particularly in light of recent economic uncertainties and shifts in consumer behavior, these decisions have become increasingly complex. Consequently, the development of data-driven recommendation systems for new location placements has emerged as a crucial aspect in enhancing economic efficiency and strengthening market competitiveness within the retail sector. The aim is to construct a recommendation system incorporating customer characteristics using panel big data collected through digital platforms. To achieve this, a comprehensive model is proposed that simultaneously considers regional models based on local characteristics and entity-level models utilizing individual customer attributes and behavioral patterns. Given the nature of digital platform data, which encompasses longitudinal data on economic activities from the same entities over time, a longitudinal model reflecting intra-entity correlations is employed at the entity level. This is anticipated to contribute significantly to the formulation of strategies for new location selection and the enhancement of success rates within the retail business.

E1026: **Optimizing spatial cluster detection in small population: A comparative study of smoothing techniques**
Presenter: **Chi Hsiung Lien**, National Chengchi University, Taiwan
Co-authors: Yin Yee Leong
SatScan is a method for detecting spatial clusters that has been widely applied in many issues. When applied to small area populations, a small

sample size may result in unstable clustering detection using SatScan. The issue of insufficient sample size is addressed in small area populations by employing smoothing methods such as partial SMR and Whittaker-Henderson before conducting clustering detection with SatScan, aiming to enhance the stability of detection. Simulation results demonstrate that the use of smoothing methods indeed improves the stability of detection. The extent of improvement varies with different smoothing methods under different clustering scenarios, with Whittaker-Henderson consistently showing stable improvement across all scenarios.

E1040: **A study on graph neural network-based stock forecasting methods in stock market**
Presenter: **Mingyu Go**, Chungnam National University, Korea, South
Co-authors: Minsu Park

Predicting stock prices involves a complex process fraught with various uncertainties, encompassing a comprehensive consideration of diverse economic, corporate, and market factors. Therefore, by considering the intricate relationships among these factors, the accuracy of stock price predictions can be enhanced. A new approach is proposed to future stock price prediction by integrating traditional time series analysis with inter-stock network structures. To calculate inter-stock similarities, multidimensional analysis reflecting the characteristics of each stock was conducted, and inter-stock adjacency was defined using a graph generation algorithm. The constructed network and market indicators were effectively learned through the graph neural network (GNN), enabling the prediction of short-term future stock prices for each stock. As a result, the network-based approach has demonstrated a more precise reflection of market trends and proven superior predictive performance compared to traditional time series analysis. The importance of understanding and leveraging interplay is underscored among stocks in stock market analysis, and methodologies that can contribute to predicting future market values are proposed.

E1056: **Final project cost estimation in EVMS using LMM**
Presenter: **Yonghyun Um**, Chungnam National University, Korea, South
Co-authors: Min Koo Lee

The earned value management system (EVMS) is a management tool utilized by organizations to track costs and schedules periodically. This enables them to assess the current status and predict the future of their projects. While various existing methods exist for estimating final project costs, they often lack accuracy and efficiency due to limited data utilization. The aim is to propose a novel approach for estimating final project costs by leveraging data from completed projects to predict ongoing ones. Specifically, a linear model must be developed to increase the accuracy of the final project cost estimate and minimize the variance of the error. To achieve this, the dynamic time warping (DTW) algorithm is employed to cluster time-series data such as actual cost (AC) and planned value (PV). Subsequently, analysis was conducted using a linear mixed model (LMM). An LMM was constructed using a variable combining the clusters of AC and PV as a random effect. All covariates expected to affect the final project cost, which is the response variable, were added, and backward elimination was performed. Through a model created using data from completed projects, the differences is examined in estimated values from existing final project cost estimation methods.

E1118: **Deep neural networks on nonparametric regression for times series data**
Presenter: **Lu Wang**, Monash University, Australia

Deep neural networks (DNN) for nonparametric regression on time series data are studied. Under a general smooth condition and a suitable restriction on the structure of the regression function, a new DNN estimator is proposed based on sparsely connected multilayer feedforward neural networks with ReLU activation function. An asymptotic normality has been established for the proposed estimator. Empirical performance via simulations and a real-data application on asset returns prediction also validate our theoretical findings.

Friday 19.07.2024 13:10 - 14:25 Parallel Session N – EcoSta2024

EI005 Room 406 MULTIVARIATE MODELS AND THRESHOLDING STATISTICS **Chair: Lixing Zhu**

E0150: **Multivariate spatiotemporal models with low rank coefficient matrix**
Presenter: **Qingzhao Zhang**, Xiamen University, China
Multivariate spatiotemporal data arise frequently in practical applications, often involving complex dependencies across cross-sectional units, time points and multivariate variables. In the literature, few studies jointly model the dependence in three dimensions. A multivariate reduced-rank spatiotemporal model is proposed to model the cross-sectional, dynamic, and cross-variable dependence simultaneously. By imposing the low-rank assumption on the spatial influence matrix, the proposed model achieves substantial dimension reduction and has a nice interpretation, especially for financial data. Due to the innate endogeneity, the quasi-maximum likelihood estimator (QMLE) is proposed to estimate the unknown parameters. A ridge-type ratio estimator is also developed to determine the rank of the spatial influence matrix. The asymptotic distribution of the QMLE and the rank selection consistency of the ridge-type ratio estimator is established. The proposed methodology is further illustrated via extensive simulation studies and two applications to stock market and air pollution datasets.

E0151: **Gaussian approximation for thresholding statistics**
Presenter: **Yumou Qiu**, Peking University, China
Thresholding statistics that sum the thresholded standardized statistics over many components are more powerful than the sum-of-square type and maximum type test statistics in detecting sparse and weak signals for global hypotheses. However, the asymptotic distribution of the thresholding statistics has only been derived under the assumption of independent variables or certain conditions on the mixing dependence among variables. Gaussian approximation result is established for the thresholding statistics under general covariance structures and high-dimensionality. Due to the non-smoothness of the thresholding function, existing techniques to show Gaussian approximation results for the sum-of-square and maximum statistics can not be applied. A novel method has been developed to establish the Gaussian approximation results for thresholding statistics. Based on this result, a bootstrap procedure is constructed to approximate the distribution of the thresholding statistics under a high-dimensional setting. Simulation studies are conducted to show the utility of the proposed approach.

EO011 Room 102 HIGH-FREQUENCY ECONOMETRICS (VIRTUAL) **Chair: Donggyu Kim**

E0186: **Nonconvex high-dimensional time-varying coefficient estimation for noisy high-frequency observations**
Presenter: **Minseok Shin**, KAIST, Korea, South
Co-authors: Donggyu Kim
A novel high-dimensional time-varying coefficient estimator is proposed for noisy high-frequency observations. In high-frequency finance, it is often observed that noises dominate a signal of an underlying true process. Thus, usual regression procedures cannot be applied to analyze noisy high-frequency observations. To handle this issue, a smoothing method is first employed for the observed variables. However, the smoothed variables still contain non-negligible noises. To manage these non-negligible noises and the high dimensionality, a nonconvex penalized regression method is proposed for each local coefficient. This method produces consistent but biased local coefficient estimators. A debiasing scheme is proposed to estimate the integrated coefficients, and a debiased integrated coefficient estimator is obtained using debiased local coefficient estimators. Then, to further account for the sparsity structure of the coefficients, a thresholding scheme is applied to the debiased integrated coefficient estimator. This scheme is called the thresholded debiased nonconvex lasso (TEN-LASSO) estimator. Furthermore, the concentration properties of the TEN-LASSO estimator are established, and a nonconvex optimization algorithm is discussed.

E0804: **Property of financial adjacency matrix for detecting local groups**
Presenter: **Minseog Oh**, KAIST, Korea, South
Co-authors: Donggyu Kim
The property of the inverse sample covariance matrix of the assets' returns as the financial adjacency matrix is investigated. Within the framework of the multi-level factor model, employing the covariance matrix as an adjacency matrix is inadequate due to the predominant impact of common factors. To detect the local group structure more effectively, utilizing the inverse of the sample covariance matrix is suggested as an adjacency matrix to reduce the common factor effects and magnify the local factor effects. It is shown that the inverse of a covariance matrix has better properties for being an adjacency matrix to identify local group membership than the original input covariance matrix. The empirical study with the returns of the top 500 trading volume stocks demonstrates that using the inverse covariance matrix as a financial adjacency matrix helps detect local groups.

E0876: **Likelihood estimation in diffusion models with high frequency estimators of volatility**
Presenter: **Donggyu Kim**, KAIST, Korea, South
A parametric estimation procedure is introduced for unobserved stochastic volatility processes. Specifically, volatility processes are first estimated using high-frequency data, and an estimation procedure is developed based on the estimated volatility process. Then, the effect of using the volatility estimators is studied, and a bias adjustment scheme is developed, where the bias comes from the volatility estimator. Finally, its asymptotic distribution is established.

EO143 Room 103 RECENT ADVANCES IN STATISTICAL MACHINE LEARNING **Chair: Biao Cai**

E0711: **RankSEG: A consistent ranking-based framework for segmentation**
Presenter: **Ben Dai**, The Chinese University of Hong Kong, China
Co-authors: Chunlin Li
Segmentation has emerged as a fundamental field of computer vision and natural language processing, which assigns a label to every pixel/feature to extract regions of interest from an image/text. The Dice and IoU metrics are used to measure the degree of overlap between the ground truth and the predicted segmentation to evaluate the performance of segmentation. A theoretical foundation of segmentation is established with respect to the Dice/IoU metrics, including the Bayes rule and Dice-/IoU-calibration, analogous to classification-calibration or Fisher consistency in classification. The existing thresholding-based framework with most operating losses has been proven to be inconsistent with respect to the Dice/IoU metrics and thus may lead to a suboptimal solution. To address this pitfall, a novel consistent ranking-based framework, namely RankDice/RankIoU, is proposed, inspired by plug-in rules of the Bayes segmentation rule. Three numerical algorithms with GPU parallel execution are developed to implement the proposed framework in large-scale and high-dimensional segmentation. The numerical effectiveness of RankDice/mRankDice is demonstrated in various simulated examples and Fine-annotated CityScapes, Pascal VOC and Kvasir-SEG datasets with state-of-the-art deep learning architectures.

E0745: **Estimation strategies for treatment and spillover effects under network interference**
Presenter: **Yichen Qin**, University of Cincinnati, United States
A novel approach is introduced for treatment and spillover effects estimation in observational studies on social networks with arbitrary interference. The direct covariate balancing estimator is proposed, which is robust for modeling misspecification and avoids the extreme weights to gain finite sample efficiency. To the best of knowledge, this is the first attempt to adopt direct covariate balancing strategies in causal effect estimation under interference. The balancing estimator is further improved with the regrouping strategy to accommodate the limited sample sizes and vertex heterogeneity. Balancing the individual covariates as well as the network embeddings is also advocated to safeguard the complexity of the data-generating process. Both theoretical and numerical justifications are established. Through the analysis of a real social experiment, the proposed method reveals the heterogeneity of conditional treatment effects, which sheds some light on the complexity of networked experiments.

E0801: **Default risk propagation in a multilayer system**
Presenter: **Zhiwei Tong**, The University of Iowa, United States
Default risk propagation is investigated within a multilayer system with dependence between different layers of the financial system (interdependence) as well as within each layer (intradependence). In times of financial distress, losses spread across sectors of the economy, resulting in the impairment of the entire financial system and leading to systemically relevant consequences. Strong propagation of default risk is discovered across different layers. To show this insight, simplicity for a two-layer network of intermediaries is considered in the low-default environment in which each intermediary has a default probability p. Given a cluster of defaults, namely, a significantly large number of defaults, in one layer, the conditional probability of a cluster of defaults is investigated in the other layer. Under various general conditions on the intradependence and interdependence structures, it is obtained that this conditional probability is asymptotic to $c * p$ as p approaches 0 for some positive constant c, where c represents the propagation effect across the two layers.

EO219 Room 104 ECONOMETRICS AND ML FOR NETWORK FORMATION AND DYNAMICS Chair: Wenze Li

E0603: **Trade shock and formula instruments with firm-level network data**
Presenter: **Wenze Li**, Nanyang Technological University, Singapore
Co-authors: Liyu Dou, Ming Li, Wenjie Wang
Beginning in early 2018, the United States and China implemented a series of substantial tariff measures, escalating over subsequent years. Nonetheless, achieving consensus on the impact of the tariff conflict on the Chinese economy remains challenging. Thus, the purpose is to examine the effects of US-China trade shocks on China utilizing firm-level data and instrumental variables.

E0608: **Market access and dynamic network formation with firm-level data**
Presenter: **Haodan Liu**, Nanyang Technological University, Singapore
Co-authors: Liyu Dou, Ming Li, Wenjie Wang
The importance of market access for dynamic firm formation is examined. Using the universe of firm-to-firm transaction data in China, firms' ability to create supplier-buyer linkages is tested dynamically by opening several Chinese high-speed railway lines. Insights are offered into how connectivity enhancements shape economic networks.

E0700: **Estimating homophily and transitivity in dynamic networks: Evidence from Chinese registered company data**
Presenter: **Xinghao Yu**, The Chinese University of Hong Kong, Shenzhen, China
Co-authors: Liyu Dou, Ming Li, Wenjie Wang
Homophily and transitivity are two crucial characteristics influencing link formation and clustering in real-world networks. The former reflects the tendency of agents to connect with similar agents, while the latter property reflects the tendency of agents to form triadic relationships with higher utility. Estimation and inference on dynamic networks with these attributes have been widely considered in the literature while rarely used in empirical studies. A conditional stable neighborhood logit model is applied to an extensive and dynamic entrepreneurship network to examine the relevance of these identification conditions and theoretical results in the real-world network data, with the aim of investigating the formation of Chinese registered companies. The preliminary work reveals that both individuals' entrepreneurial spirit and the spillover utility from entrepreneurial success play a crucial role in forming entrepreneurial networks in China and that the latter has a more significant impact. More insights are selected to be discussed further.

EO122 Room 105 FINANCIAL MARKET DYNAMICS AND RISK ASSESSMENT INNOVATIONS Chair: Heng Xiong

E0302: **Distributionally robust insurance under the Wasserstein distance**
Presenter: **Wenjun Jiang**, University of Calgary, Canada
The optimal insurance contracting from the perspective of a decision maker (DM) who has an ambiguous understanding of the loss distribution is studied. The ambiguity set of loss distributions is represented as a pth-order Wasserstein ball, where p is an integer centered around a specific benchmark distribution. The DM selects the indemnity function that minimizes the worst-case risk within the risk-minimization framework, considering the constraints of the Wasserstein ball. Assuming that the DM is endowed with a convex distortion risk measure and that insurance pricing follows the expected-value premium principle, the explicit structures of both the indemnity function and the worst-case distribution are derived using a novel survival-function-based representation of the Wasserstein distance. A specific example is examined where the DM employs the GlueVaR, and numerical results are provided to demonstrate the sensitivity of the worst-case distribution concerning the model parameters.

E0886: **Automatic slicing of connectivity networks for enhanced analysis of volatility spillovers**
Presenter: **Heng Xiong**, Wuhan University, China
Effective portfolio management and hedging strategies hinge on accurately understanding the transmission of volatility across interconnected markets. A novel method is proposed that integrates nonparametric changepoint detection to automatically segment interconnectedness networks. Building upon this, the Diebold and Yilmaz (DY) framework is adapted to analyze these segments through return dynamics and EGARCH-filtered volatility. Unlike conventional methods, the sliced DY model significantly enhances the identification of shifts in market connectivity and minimizes divergence in conclusions across multiple measurement periods. To rigorously test this new approach, daily price data is employed from China's crude oil and agricultural futures markets. The empirical results robustly demonstrate that the proposed method successfully mitigates the issue of inconsistent findings across time metrics and enables a more nuanced examination of structural shifts.

E0974: **Introducing new rate factors and statistical learning to improve fire loss prediction accuracy**
Presenter: **Ah-ram Lee**, Ewha womans university, Korea, South
Under the current insurance practice, the premium is charged based on a simple calculation rule using a limited number of rating factors in South Korea. While varying depending on the specific type of insurance, the rate-making procedure is not generally based on statistical models. Particularly in the case of fire insurance, the risk premium is calculated from a simple classification based on building type and structure. Furthermore, such risk classification is solely based on the frequency, and the modelling of the severity part is ignored. The insurance risk is modelled separately

in terms of frequency and severity. The regression model and neural network approach are used to quantify the risk. Then, it is shown how to calculate the insurance premium based on these models.

EO141 Room 108 RECENT ADVANCES IN STATISTICAL LEARNING (VIRTUAL) Chair: BaoLuo Sun

E0681: **Using synthetic data to regularize maximum likelihood estimation**
Presenter: **Weihao Li**, National University of Singapore, China
Co-authors: Dongming Huang
To overcome challenges in fitting complex models with small samples, catalytic priors have recently been proposed to stabilize the inference by supplementing observed data with synthetic data generated from simpler models. Based on a catalytic prior, the maximum a posteriori (MAP) estimator is a regularized estimator that maximizes the weighted likelihood of the combined data. This estimator is straightforward to compute, and its numerical performance is superior or comparable to other likelihood-based estimators. Several theoretical aspects are studied regarding the MAP estimator in generalized linear models, with a particular focus on logistic regression. It is first proven that under mild conditions, the MAP estimator exists and is stable against the randomness in synthetic data. The consistency of the MAP estimator is then established when the dimension of covariates diverges slower than the sample size. Furthermore, the convex Gaussian min-max theorem is utilized to characterize the asymptotic behavior of the MAP estimator as the dimension grows linearly with the sample size. These theoretical results clarify the role of the tuning parameters in a catalytic prior and provide insights into practical applications. Numerical studies are provided to confirm the effective approximation of the asymptotic theory in finite samples and to illustrate adjusting inferences based on the theory.

E0768: **Covariate-robust clustering**
Presenter: **Yunjin Choi**, University of Seoul, Korea, South
Common clustering methods can be overly influenced by single covariates, leading to inconsistent results across different covariates if each covariate is analyzed separately. To address this issue, a novel clustering approach identifying clusters that are not dominated by a single covariate is proposed but rather captures the underlying structure across all covariates. This is achieved by minimizing the maximum error across covariates. In this approach, a covariate's error is the number of points assigned to clusters that deviate from their closest covariate-specific centers. The motivation is the customer segmentation based on reviews that consist of positive and negative comments. Existing methods, although employed on the whole comments, often yield clusters dominated by positive or negative comments. The approach addresses this by providing clustering that bridges the gap between results based solely on positive or negative reviews.

E0813: **On doubly robust estimation with nonignorable missing data using instrumental variables**
Presenter: **BaoLuo Sun**, National University of Singapore, Singapore
Co-authors: Wang Miao, Wickramarachchi Deshanee
Suppose the interest is in the mean of an outcome subject to nonignorable nonresponse. New semiparametric estimation methods are developed with instrumental variables, which affect nonresponse but not the outcome. The proposed estimators remain consistent and asymptotically normal even under partial model misspecifications for two variation-independent nuisance components. The performance of the proposed estimators is evaluated via a simulation study and applied to adjust for missing data induced by HIV testing refusal in evaluating HIV seroprevalence in Mochudi, Botswana, using interviewer experience as an instrumental variable.

EO138 Room 109 RECENT DEVELOPMENTS IN SURVIVAL ANALYSIS AND TRANSFER LEARNING Chair: Chun Yin Lee

E0353: **Semiparametric structural equation models with interval-censored data**
Presenter: **Shuwei Li**, Guangzhou University, China
Structural equation models offer a valuable tool for delineating the complicated interrelationships among multiple variables, including observed and latent variables. Over the last few decades, structural equation models have successfully analyzed complete and right-censored survival data, exemplified by wide applications in psychological, social, or genomic studies. However, the existing methodology for structural equation modeling is not concerned with interval-censored data, a type of coarse survival data arising typically from periodic examinations for the occurrence of asymptomatic disease. The aim is to fill this gap and provide a flexible semiparametric structural equation modeling framework. A general class of factor-augmented transformation models is proposed to model the interval-censored outcome of interest in the presence of latent risk factors. An expectation-maximization algorithm is subtly designed to conduct the nonparametric maximum likelihood estimation. Furthermore, the asymptotic properties of the proposed estimators are established by leveraging the empirical process theory. The numerical results obtained from extensive simulations and an application to the Alzheimer's disease data set demonstrate the proposed method's empirical performance and practical utility.

E0712: **Deep representation transfer learning for partially linear models**
Presenter: **Baihua He**, University of Science and Technology, China
Transfer learning has achieved many successes in practical applications. Some issues remain unexplored, especially the combination of parameter interpretability and model flexibility. A novel deep neural network-based transfer learning approach is presented within a partially linear model to enhance model performance using heterogeneous source domain data. The proposal addresses the challenges of high-dimensional and non-linear data, offering both high prediction accuracy and improved interpretability of primary parameters. Distinct from existing statistical methods, the framework facilitates positive transfer under certain domain heterogeneity, maintaining robustness against over-fitting through data augmentation. The consistency of representation learning, asymptotic normality of the primary parameters, and semi-parametric efficiency are established. The promising performance of the proposal is demonstrated through simulations and empirical validations on the house renting price study

E0645: **Survival analysis with a random change-point**
Presenter: **Chun Yin Lee**, Hong Kong Polytechnic University, Hong Kong
Co-authors: Kin Yau Wong
Contemporary works in change-point survival models mainly focus on an unknown universal change-point shared by the whole study population. However, in some situations, the change-point is plausibly individual-specific, such as when it corresponds to the telomere length or menopausal age. Also, maximum-likelihood-based inference for the fixed change-point parameter is notoriously complicated. The asymptotic distribution of the maximum likelihood estimator is non-standard, and computationally intensive bootstrap techniques are commonly used to retrieve its sampling distribution. The motivation is from a breast cancer study, where the disease-free survival time of the patients is postulated to be regulated by the menopausal age, which is unobserved. As menopausal age varies across patients, a fixed change-point survival model may be inadequate. Therefore, a novel proportional hazards model is proposed with a random change-point. A nonparametric maximum likelihood estimation approach is developed, and a stable EM algorithm is devised to compute the estimators. Because the model is regular, conventional likelihood theory is employed for inference based on the asymptotic normality of the Euclidean parameter estimates, and a profile-likelihood approach can consistently estimate the variance of the asymptotic distribution. A simulation study demonstrates the proposed methods' satisfactory finite-sample performance, yielding small bias and proper coverage probabilities.

EO130 Room 110 EXPERIMENT DESIGN AND RELIABILITY OPTIMIZATION Chair: Dianpeng Wang

E0311: **Adaptive grid designs for classifying monotonic binary simulations**
Presenter: **Tian Bai**, Beijing Insititute of Technology, China
Co-authors: Xu He, Dianpeng Wang
The motivation is the need for effective classification in ice-breaking dynamic simulations aimed at determining the conditions under which an underwater vehicle will break through the ice. This simulation is extremely time-consuming and yields deterministic, binary, and monotonic outcomes. Detecting the critical edge between the negative-outcome and positive-outcome regions with minimal simulation runs necessitates an efficient experimental design for selecting input values. Adaptive designs, which sequentially select input values based on obtained outcomes, outperform static designs significantly by eliminating redundant points without losing information. A new class of adaptive designs is proposed called adaptive grid designs. An adaptive grid is a sequence of grids with increasing resolution, such that lower-resolution grids are proper subsets of higher-resolution grids. By prioritizing simulation runs at lower resolution points and skipping redundant runs, adaptive grid designs require an order of magnitude fewer simulation runs to ensure a certain level of classification accuracy than the best possible static design and the same order of magnitude of runs as the best possible adaptive design. Numerical results across test functions, the road crash simulation, and the ice-breaking simulation validate the superiority of adaptive grid designs.

E0893: **Reliability analysis for systems with thermal balance control of internal components**
Presenter: **Dong Xu**, Beijing Institute of Technology, China
Co-authors: Yubin Tian, Dianpeng Wang
Temperature is a non-negligible source of failure in common electronic multi-component systems. To avoid localized high temperatures occurring in multiple power supply channels within a satellite power supply system, the thermal balance control method is investigated based on the minimum energy criterion, which could consider the local temperatures and the degree of dispersion of heat-generating components simultaneously. A reliability model for the complex system is proposed, and a Monte Carlo method is presented to evaluate the system reliability by mimicking the temperature evolution under a dynamic power supply demand modeling by a random process. Some numerical experiments are also made to demonstrate the superior performance of the novel method under different power supply demands. Meanwhile, with a given constraint of reliability, one method for optimizing the installation distance of components is proposed to minimize system volume.

E0647: **Numerical algorithm-aided approaches for analytically finding optimal designs**
Presenter: **Ping-Yang Chen**, National Taipei University, Taiwan
The traditional way in statistics to find optimal designs for regression models is an analytical approach. However, the mathematical technique often faces challenges with complex experimental setups or difficult-to-solve criteria, suggesting that developing flexible and effective algorithms to search for optimal designs is highly valuable. In particular, numerical results from an algorithm can be instrumental in deriving analytical descriptions of optimal designs. For instance, particle swarm optimization has been shown to be quite effective for finding optimal designs for hard design problems. How the output can be utilized is demonstrated to discover new analytic optimal designs, showcasing its utility in scenarios such as A-optimal designs for generalized linear models and the standardized maximin D-optimal designs for nonlinear inhibition models. These examples illustrate how optimization algorithms complement traditional analytical methods, enhancing the identification of optimal experimental designs.

EO032 Room 111 RECENT ADVANCES IN JOINT MODELLING OF MULTI-OUTCOME DATA Chair: Christiana Charalambous

E0734: **A joint model with finite-mixture structure for longitudinal and survival data**
Presenter: **Xuerui Yang**, University of Manchester, United Kingdom
Co-authors: Jianxin Pan, Christiana Charalambous
In longitudinal studies, heterogeneous time-to-event data is commonly collected together with longitudinal data. A finite-mixture Cox PH model is used to quantify the probabilities of clusters in survival data, and joint mean-covariance models are used to characterize within-subject patterns in the longitudinal outcomes. These two sub-models are linked via a shared parameter, but a third sub-model is also introduced to show how the covariance structures are connected. An MCMC algorithm for parameter estimation is proposed, and the models' performance is demonstrated via simulation studies. An application to AIDS data illustrates how the joint models may be used to capture the heterogeneity and longitudinal patterns.

E0981: **Maximum likelihood estimation for semiparametric regression models with interval-censored multistate data**
Presenter: **Yu Gu**, The University of Hong Kong, Hong Kong
Co-authors: Donglin Zeng, Gerardo Heiss, Danyu Lin
Interval-censored multistate data arise in many studies of chronic diseases, where the health status of a subject can be characterized by a finite number of disease states and the transition between any two states is only known to occur over a broad time interval. Time-dependent covariates are potentially related to multistate processes through semiparametric proportional intensity models with random effects. Nonparametric maximum likelihood estimation is studied under general interval censoring and develop a stable EM algorithm. It is shown that the resulting parameter estimators are consistent and that the finite-dimensional components are asymptotically normal with a covariance matrix that attains the semiparametric efficiency bound and can be consistently estimated through profile likelihood. In addition, it is demonstrated through extensive simulation studies that the proposed numerical and inferential procedures perform well in realistic settings. Finally, an application to a major epidemiologic cohort study is provided.

E1080: **Modelling biomarker variability in joint analysis of longitudinal and time-to-event data**
Presenter: **Christiana Charalambous**, University of Manchester, United Kingdom
Visit-to-visit variability of a biomarker has been recognized as an important driver in predicting related diseases. However, existing variability measures are criticized for being entangled with random variability caused by measurement error or for being unreliable. A new measure is proposed to describe the biological variability of a biomarker by capturing the fluctuation of individual trajectories behind longitudinal measurements. Given a mixed-effects model for longitudinal data with the mean function over time specified by cubic splines, the proposed variability measure can be mathematically expressed as a quadratic form of random effects. For the time-to-event data, a Cox model is used, which incorporates the defined variability and the current level of the underlying longitudinal trajectory as covariates. The proposed joint models are further extended to incorporate the weighted cumulative effects of both biomarker level and variability on the survival hazard. Simulation studies are conducted to reveal the advantages of the proposed methods over the two-stage method, as well as a simpler joint modelling approach that does not take biomarker variability into account. Finally, the models are applied to investigate the effect of systolic blood pressure variability on cardiovascular events in the Medical Research Council elderly trial.

EO065 Room 212 ADVANCES IN BAYESIAN MODELING AND COMPUTATION Chair: Cheng Li

E0268: **Robust Bayesian inference on Riemannian submanifold**
Presenter: **Rong Tang**, HKUST, Hong Kong
Non-Euclidean spaces routinely arise in modern statistical applications such as medical imaging, robotics, and computer vision, to name a few. While traditional Bayesian approaches are applicable to such settings by considering an ambient Euclidean space as the parameter space, the benefits of integrating manifold structure are demonstrated in the Bayesian framework, both theoretically and computationally. Moreover, existing Bayesian approaches that are designed specifically for manifold-valued parameters are primarily model-based and are typically subject to inaccurate uncertainty quantification under model misspecification. A robust model-free Bayesian inference is proposed for parameters defined on a Riemannian submanifold, which is shown to provide valid uncertainty quantification from a frequentist perspective. Computationally, a Markov chain Monte Carlo is proposed to sample from the posterior on the Riemannian submanifold, where the mixing time, in the large sample regime, is shown to depend only on the intrinsic dimension of the parameter space instead of the potentially much larger ambient dimension.

E0294: **Latent modularity in multi-view data**
Presenter: **Xuejun Yu**, National University of Singapore, Singapore
Co-authors: Andrea Cremaschi, Maria De Iorio, Ajay Jasra, Shu Qin Delicia Ooi, Xiu Ling Evelyn Loo, Navin Michael
Asthma, hypertension and obesity are three of the most common chronic diseases worldwide, with known presence of comorbid pathophysiological mechanisms. Despite studies indicating a complex coregulatory mechanism between these diseases exists, quantitative analyses in children are currently scarce. Furthermore, data collected from different sources are usually analyzed separately, neglecting shared information among subjects, underlining the need for a more comprehensive statistical approach. A novel Bayesian nonparametric model is developed for the joint analysis of biomarkers of different types related to obesity (longitudinal data), history of asthma (panel count data) and symptoms of hypertension (multistate process). Random partitions of subjects in each dataset are modelled independently conditionally on an underlying partition structure. The proposed strategy allows for sharing information among clustering structures within different datasets, thus providing more robust inference. Random partitions of different datasets are marginally dependent, with the level of dependence learnt from the data. The model involves mixed-type covariates, aiding the identification of risk factors affecting the evolution of diseases. A tailored MCMC algorithm is developed, which entails simpler computations than existing methods based on hierarchical random measures. An application from the Singaporean birth cohort GUSTO (Growing Up in Singapore Towards Healthy Outcomes) is presented.

E0435: **Semi-implicit variational inference with score matching**
Presenter: **Cheng Zhang**, Peking University, China
Semi-implicit variational inference (SIVI) greatly enriches the expressiveness of variational families by considering implicit variational distributions defined in a hierarchical manner. However, due to the intractable densities of semi-implicit distributions, typical SIVI approaches often use surrogate evidence lower bounds (ELBOs) or employ expensive inner-loop MCMC runs for unbiased ELBOs for training. New SIVI methods are introduced based on several alternative training objectives via score matching, which allows the leverage of the hierarchical structure of semi-implicit distributions to bypass the intractability of their densities. The basic score-matching framework for SIVI is started with a minimax formulation called SIVI-SM. How to further enhance the flexibility of semi-implicit distribution is then discussed by allowing multiple hierarchical layers, which can also be used to accelerate the diffusion model given the learned score networks. Lastly, KSIVI is introduced, a variant of SIVI-SM that eliminates the need for lower-level optimization through kernel tricks. An upper bound for the variance of the Monte Carlo gradient estimators of the KSD objective is derived, which allows establishing novel convergence guarantees of KSIVI.

EO148 Room 202 HIGH-DIMENSIONAL ROBUST STATISTICAL INFERENCE Chair: Cheng Wang

E0429: **Robust probabilistic principal component analysis with mixture of exponential power distributions**
Presenter: **Zhenghui Feng**, Harbin Institute of Technology, Shenzhen, China
Co-authors: Xinyi Wang, Xiao Chen, Heng Peng
The EP-MPPCA model is introduced, which serves as a flexible and robust alternative to conventional Gaussian-based mixtures of probabilistic principal component analysis (MPPCA) for high-dimensional data analysis. The EP-MPPCA model utilizes the exponential power distribution family, making it more adept at handling heterogeneous data distributions and outliers. Algorithms and estimation methods are provided for the EP-MPPCA model, and its performance is evaluated through simulations. In real data analysis, the EP-MPPCA model can be practically applied in two important applications: unsupervised clustering and image data reconstruction. Specifically, the EP-MPPCA model is shown to effectively handle outliers in high-dimensional image data, leading to improved reconstruction quality. Additionally, the model can achieve superior clustering results in an unsupervised manner for high-dimensional data.

E0430: **Dimension reduction for extreme regression via contour projection**
Presenter: **Jing Zeng**, University of Science and Technology of China, China
Co-authors: Liujun Chen
In the context of regression problems, a primary objective is to infer the extreme values of the response given a set of predictors. The high dimensionality and heavy-tailedness of predictors pose significant challenges, limiting the applicability of classical tools for inferring conditional extremes. The focus is on the central extreme subspace (CES), whose existence and uniqueness are guaranteed under fairly mild conditions. By projecting the data onto CES, the dimension of the predictors is reduced while all information for inferring conditional extremes is retained, which effectively addresses the high dimensionality issue. Then, the novel COPSE method is proposed to estimate CES, involving the projection of predictors onto an elliptical contour. Notably, COPES exhibits robustness against heavy-tailed data. The theoretical justification is provided for the consistency of COPES under mild assumptions. Overall, the proposal not only extends the toolkit for extreme regression but also broadens the scope of the dimension reduction techniques. The effectiveness of the proposal is demonstrated through extensive simulation studies and an application to Chinese stock market data.

E0984: **A portmanteau local feature discrimination approach to the classification with high-dimensional matrix-variate data**
Presenter: **Shan Luo**, Shanghai Jiao Tong University, China
Co-authors: Zehua Chen, Zengchao Xu
Matrix-variate data arise in many scientific fields, such as face recognition, medical imaging, etc. Matrix data contain important structure information, which can be ruined by vectorization. Methods incorporating the structure information into analysis have significant advantages over vectorization approaches. The focus is on the problem of two-class classification with high-dimensional matrix-variate data and the proposal of a novel portmanteau-local-feature discrimination (PLFD) method. The method first identifies local discrimination features of the matrix variate and then pools them together to construct a discrimination rule. The theoretical properties of the PLFD method are investigated and its asymptotic optimality is established. Extensive numerical studies are carried out, including simulation and real data analysis, to compare this method with other methods available in the literature, which demonstrate that the PLFD method has a great advantage over the other methods in terms of misclassification rate.

EO315 Room 204 RECENT ADVANCES IN BAYESIAN METHODOLOGY Chair: Ning Ning

E0203: **Bayesian biclustering and its application in education data analysis**
Presenter: **Weining Shen**, UC Irvine, United States
A novel nonparametric Bayesian IRT model is proposed that estimates clusters at the question level while simultaneously allowing for heterogeneity at the examinee level under each question cluster, characterized by the mixture of binomial distributions. The main contribution is threefold. First, the new model is presented and demonstrated to be identifiable under a set of conditions. Second, it is shown that the model can correctly identify question-level clusters asymptotically, and the parameters of interest that measure the proficiency of examinees in solving certain questions can be estimated at a root-n rate (up to a log term). Third, a tractable sampling algorithm is presented to obtain valid posterior samples from the proposed model. Compared to the existing methods, the model reveals the multi-dimensionality of the examinee's proficiency level in handling different types of questions parsimoniously by imposing a nested clustering structure. The proposed model is evaluated via a series of simulations and applied to an English proficiency assessment data set. This data analysis example nicely illustrates how the model can be used by test makers to distinguish different types of students and aid in the design of future tests.

E0249: **Sampling from high-dimensional, multimodal Bayesian posterior using adaptively-tuned, tempered Hamiltonian Monte Carlo**
Presenter: **Joonha Park**, University of Kansas, United States
Hamiltonian Monte Carlo (HMC) is widely used for sampling high-dimensional target distributions arising from Bayesian posterior. Although HMC scales favorably with increasing dimensions, it is very inefficient when the target distribution is strongly multimodal. Sampling highly multimodal target distributions is often tackled using tempering strategies, but the resulting algorithms are often difficult to tune in practice, especially in high dimensions. A method that combines the tempering strategy with Hamiltonian Monte Carlo is developed in a way that allows efficient sampling of high-dimensional, strongly multimodal distributions. The method consists in proposing a candidate for the next state of the Markov chain by solving the Hamiltonian equations of motion with time-varying mass. Compared to the simulated tempering method or the parallel tempering method, the method has a distinctive advantage in the case where target distribution changes at each iteration, such as in the Gibbs sampler. A careful tuning strategy is developed for the method and an adaptively-tuned, tempered Hamiltonian Monte Carlo (ATHMC) algorithm is proposed. The excellent sampling efficiency of ATHMC is demonstrated for high-dimensional, multimodal distributions using a mixture of Gaussians and a Bayesian posterior distribution for a sensor network self-localization problem.

E0807: **Uncertainty quantification in Bayesian reduced-rank sparse regressions**
Presenter: **Alexander Shestopaloff**, Queen Mary University of London, United Kingdom
Reduced-rank regression recognizes the possibility of a rank-deficient matrix of coefficients, which is particularly useful when the data is high-dimensional. A novel Bayesian model is proposed for estimating the rank of the coefficient matrix, which obviates the need for post-processing steps and allows for uncertainty quantification. The method employs a mixture prior to the regression coefficient matrix along with a global-local shrinkage prior to its low-rank decomposition. Then, the signal adaptive variable selector is relied onto perform sparsification and define two novel tools: the posterior inclusion probability uncertainty index and the relevance index. The validity of the method is assessed in a simulation study, and then its advantages and usefulness are shown in real-data applications on the chemical composition of tobacco and on the photometry of galaxies.

EO088 Room 207 STATISTICAL METHODS FOR CAUSAL INFERENCE AND POLICY LEARNING Chair: Yifan Cui

E0441: **Value enhancement of reinforcement learning via efficient and robust trust region optimization**
Presenter: **Fan Zhou**, Shanghai University of Finance and Ecnomics, China
Reinforcement learning (RL) is a powerful machine learning technique that enables an intelligent agent to learn an optimal policy that maximizes the cumulative rewards in sequential decision-making. Most of the methods in the existing literature are developed in online settings where the data are easy to collect or simulate. Motivated by high-stake domains such as mobile health studies with limited and pre-collected data, offline reinforcement learning methods are studied. To efficiently use these datasets for policy optimization, a novel value enhancement method is proposed to improve the performance of a given initial policy computed by existing state-of-the-art RL algorithms. Specifically, when the initial policy is not consistent, the method will output a policy whose value is no worse and often better than that of the initial policy. When the initial policy is consistent, under some mild conditions, the method will yield a policy whose value converges to the optimal one at a faster rate than the initial policy, achieving the desired value enhancement property. The proposed method is generally applicable to any parametrized policy that belongs to a certain pre-specified function class (e.g., deep neural networks). Extensive numerical studies are conducted to demonstrate the superior performance of the method.

E0451: **Off-policy evaluation in doubly inhomogeneous environments**
Presenter: **Zeyu Bian**, University of Miami, United States
Co-authors: Chengchun Shi, Zhengling Qi, Lan Wang
The aim is to study off-policy evaluation (OPE) under scenarios where two key reinforcement learning (RL) assumptions-temporal stationarity and individual homogeneity are both violated. To handle the "double inhomogeneities", a class of latent factor models is proposed for the reward and observation transition functions, under which we develop a general OPE framework that consists of both model-based and model-free approaches. The theoretical properties of the proposed value estimators are established, and empirically, it is shown that the approach outperforms competing methods that ignore either temporal nonstationarity or individual heterogeneity. Finally, the method is illustrated on a data set from the Medical Information Mart for Intensive Care.

E0617: **Rate-optimal online learning for dynamic assortment selection with positioning**
Presenter: **Yiyun Luo**, School of Statistics and Management, Shanghai University of Finance and Economics, China
In online retailing, the seller aims to offer an assortment of items with maximized expected revenue. A new online learning problem is introduced called dynamic assortment selection with positioning (DAP), which additionally investigates the positioning of items within the assortment. Specifically, customers make purchases based on the item's attractiveness as the product of the position effect and unknown preference parameter through a multinomial logit choice model. The objective is to maximize the revenue over a finite horizon. It is first demonstrated that any assortment-only algorithm that neglects position effects results in linear regrets. To address this gap, the truncated linear regression upper confidence bound (TLR-UCB) policy is proposed. TLR-UCB utilizes a novel geometric linear-bandit-type feedback structure to construct upper confidence bounds (UCB) for unknown preference parameters, accounting for both random and adaptive position effects. To ensure the validity of UCB construction, TLR-UCB adopts a truncation technique for conditional geometric responses before applying linear regression. In theory, a regret upper bound of $O(T^(1/2))$ is established for TLR-UCB, matching the derived regret lower bound for the DAP problem. Extensive experiments demonstrate the superior performance of TLR-UCB by incorporating the position effects into the dynamic assortment selection process.

EO182 Room 209 RECENT ADVANCES IN COMPLEX DATA ANALYSIS WITH HETEROGENEITY Chair: Ning Wang

E0612: **Detecting change points in low-rank tensors via tucker decomposition and one-dimensional series analysis**
Presenter: **Jiaqi Huang**, Beijing Normal University, China
Co-authors: Ning Wang, Lixing Zhu

The purpose is to address the challenge of detecting change points within tensor data characterized by a low-rank structure. It is proposed that these change points can manifest within the core tensors or their corresponding subspaces, as identified through Tucker decomposition. Initially, the approach targets change points associated with subspaces. A MOSUM-based criterion is introduced that captures the dimensionality of these subspaces, effectively simplifying the intricate, high-dimensional problem into a more manageable one-dimensional time series change point detection task. Subsequently, a change point detection algorithm is integrated with a dimensionality determination technique applied to the one-dimensional series, enabling the identification of all potential change points. Building upon this subspace analysis, change points are further pinpointed within core tensors using an adaptive ridge ratio statistic. The findings affirm the consistency of the estimated subspaces' foundational matrices derived from Tucker decomposition, as well as the consistency of the identified change points. Additionally, the methodology is extended to accommodate tensors featuring structural modes. Through numerical experiments, the robust performance of the method is demonstrated. A practical application is also included, using real-world data to illustrate the utility of the approach.

E0784: **Variants of high-dimensional EM algorithm for mixed linear regression**
Presenter: **Ning Wang**, Beijing Normal University, China

The expectation-maximization (EM) algorithm and its variants are widely used in statistics. In high-dimensional mixture linear regression, the model is assumed to be a finite mixture of linear regression, and the number of predictors is much larger than the sample size. The aim is to consider solving the high-dimensional problem in mixed linear regression. The regression coefficients are assumed to be sparse, and the sparsity pattern is the same for different mixtures. A group lasso penalized approach is presented, a modified group lasso penalized approach, and a subset selection-based approach. The non-asymptotic convergence results for the direct output of the algorithms will be provided. The proposed methods have encouraging performances in numerical studies.

E0812: **Generalized partially functional linear model based on multi-source data**
Presenter: **Xiaochen Zhang**, Beijing Normal University, China

Multi-source data may be presented in different forms (such as scalar data, functional data, network data, etc.). The integration of multi-source data analysis is of significant importance, and the aim is to propose a generalized partially functional linear model for integrating functional data with other forms of data (such as network structures, scalar data, etc.) from different source domains. To improve the estimation and prediction, the network cohesion is enforced using the Laplace quadratic penalty function. Simulation results and real data application demonstrate the satisfactory performance of the proposed methods.

EO174 Room 210 SEQUENTIAL HYPOTHESIS TESTING AND CHANGE-POINT DETECTION Chair: Liyan Xie

E0224: **Continual density ratio estimation for online time series with applications in change detection**
Presenter: **Yidong Ouyang**, The Chinese University of Hong Kong, Shenzhen, China
Co-authors: Liyan Xie

Density ratio estimation plays a crucial role in data-driven decision-making. A general setting with non-stationary data sequences is considered. Inspired by the telescoping density-ratio estimation (TRE) that can improve the estimation of ratios between two highly dissimilar densities, a continual density-ratio estimation (CDRE) framework is developed to track the density ratio over time. Sliding windows are constructed, and the density ratio is estimated between two consecutive windows, which will be used to update the density ratio estimate gradually. CDRE enjoys both computational and memory efficiency. Furthermore, a novel detection algorithm, called CDRE-CuSum, is also proposed to apply the CDRE outcomes for online change-point detection. The recursive structure of CDRE-CuSum statistics makes it efficient for online implementation. Empirically, the CDRE is demonstrated to perform well in tracking the density ratio for non-stationary time series and in detecting abrupt changes in data distributions.

E0225: **Online correlation change detection for high-dimensional data**
Presenter: **Jie Gao**, Chinese university of Hong Kong(Shenzhen), China
Co-authors: Liyan Xie, Zhaoyuan Li

The problem of change point detection in the correlation structure of streaming high-dimensional data is explored, with minimum assumptions posed on the underlying data distribution and correlation structure. Depending on the $/L_1/$ and $/L_\infty/$ norm of squared difference of vectorized pre-change and post-change correlation matrices, dense and sparse settings are considered, respectively. Both window-limited and Shewhart-type test statistics are proposed. A novel method for threshold selection is designed based on sign-flip permutation. In addition, two enhancement techniques, synthetic minority oversampling technique (SMOTE) and knockoff, are combined with window-limited test statistics to tackle the instability in detection due to small sample sizes. Theoretical evaluations of these proposed methods are conducted regarding average run length and detection delay. Numerical studies are conducted to examine the finite sample performances of the proposed methods. The methods are effective in most simulation cases, as the average detection delays are very close to the exact CUSUM. Moreover, a combined $/L_1/$ and $/L_\infty/$ norm approach is applied and has expected performance for transitions from sparse to dense settings. Two real datasets, El Nino event prediction and seismic event, are also analyzed to illustrate the proposed methods' efficacy in detecting fundamental changes with minimal delay.

E0767: **Sequential multiple testing: An overview of different setups**
Presenter: **Yiming Xing**, University of Illinois at Urbana-Champaign, China

The problem of simultaneously testing the marginal distributions of sequentially monitored data streams is considered. To solve this problem, the need is to specify, for each data stream, a time for making a decision, a time for stopping sampling, and a decision rule. Based on whether information could be shared among data streams, there are the decentralized setup and the centralized setup, and based on the relationship between the times of making a decision and the times of stopping sampling, the centralized setup could be further divided into the synchronous setup, the asynchronous-decision setup, the asynchronous-stopping setup, and other setups. An overview of all these setups is given, and one solution for the asynchronous-decision setup is elaborated. Specifically, a novel sequential multiple testing procedure is proposed, which minimizes the expected sample size in every data stream under every possible hypothesis configuration, asymptotically as certain global error metrics go to zero. This asymptotic optimality result is established under general parametric composite hypotheses, various error metrics, and weak distributional assumptions that allow for temporal dependence.

EO009 Room 313 ADVANCES IN TIME SERIES ANALYSIS Chair: Artem Prokhorov

E0754: **Phase spline-analysis for time series dynamics**
Presenter: **Lyudmila Gadasina**, Saint-Petersburg State University, Russia
Co-authors: Lyudmila Vyunenko, Ivan Labutkin

The analysis of the medium- and long-term time series dynamics was carried out using the phase shadow concept. The approach includes the following steps: determining a time interval, considered as a minimum unit within which the dynamics of the series can be interpreted as short-term volatility; time series smoothing by one-dimensional adaptive regression splines with the number of splines corresponding to the selected time intervals; constructing the phase shadow treated as the projection of the first derivative of the resulting smooth function y onto the plane (y, y'). Phase spline analysis allows one to identify and visualize cycles, bubbles, and structural shifts for a time series of economic data. The approach was tested on the NASDAQ index, and Bitcoin prices were taken over the period from November 2019 to January 2023, with an interval of 24 hours.

E0881: **On the modelling and prediction of high-dimensional functional time series**
Presenter: **Qin Fang**, the University of Sydney, Australia
Co-authors: Jinyuan Chang, Xinghao Qiao, Qiwei Yao

A two-step procedure is proposed to model and predict high-dimensional functional time series, where the number of function-valued time series p is large in relation to the length of time series n. In the first step, an eigenanalysis of a positive definite matrix is performed, which leads to a one-to-one linear transformation for the original high-dimensional functional time series, and the transformed curve series can be segmented into several groups such that any two subseries from any two different groups are uncorrelated both contemporaneously and serially. Consequently, in the second step, those groups are handled separately without the loss of information on the overall linear dynamic structure. The second step is devoted to establishing a finite-dimensional dynamical structure for each group's transformed functional time series. Furthermore, the finite-dimensional structure is represented by a vector time series. Modelling and forecasting for the original high-dimensional functional time series are realized via those for the vector time series in all the groups. The theoretical properties of the proposed methods are investigated, and the finite-sample performance is illustrated through both extensive simulation and two real datasets.

E1103: **Early warning systems for financial markets of emerging economies**
Presenter: **Artem Kraevskiy**, Sberbank, Russia
Co-authors: Artem Prokhorov, Evgeny Sokolovskiy

A new online early warning system (EWS) for what is known in machine learning is developed and applied as concept drift, in economics as a regime shift and in statistics as a change point. The system goes beyond the linearity assumed in many conventional methods and is robust to heavy tails and tail dependence in the data, making it particularly suitable for emerging markets. The key component is an effective change-point detection mechanism for conditional entropy of the data rather than for a particular indicator of interest. Combined with recent advances in machine learning methods for high-dimensional random forests, the mechanism is capable of finding significant shifts in information transfer between interdependent time series when traditional methods fail. The aim is to explore when this happens using simulations and illustrations are provided by applying the method to Uzbekistan's commodity and equity markets as well as to Russia's equity market in 2021-2023.

EO249 Room 408 ADVANCES IN MATHEMATICAL DATA SCIENCE Chair: Xin Guo

E0369: **Generalization and optimization of gradient methods for single-layer neural networks**
Presenter: **Yunwen Lei**, The University of Hong Kong, Hong Kong

Neural networks have achieved impressive performance in various applications. The generalization and optimization of shallow neural networks (SNNs) are discussed. Both gradient descent (GD) and stochastic gradient descent (SGD) are considered for training SNNs. It shows how the generalization and optimization should be balanced to obtain consistent error bounds under a relaxed overparameterization setting. The existing estimates are improved on the weak-convexity parameter of SNNs along the trajectories of the optimization process.

E0536: **Smoothed kth power expectile regression with MQ-type function**
Presenter: **Wenwu Gao**, Anhui University, China

Quantile regression, a powerful regression analysis method, estimates the conditional distribution of a response variable at different quantile levels. However, its non-differentiable loss function poses computational challenges. While asymmetric least squares regression simplifies the computation and asymmetric k-th power expectile regression offers higher asymptotic efficiency in certain cases, both methods suffer from non-differentiable loss functions. Convolution methods based on specific kernel functions have been employed to construct smooth approximations of the quantile regression loss. A novel and more intuitive approach for constructing smooth loss functions is proposed. Expressions for the smoothed versions of various loss functions are derived explicitly, including those for quantile regression $(k = 1)$, expectile regression, and k-th power expectile regression $(1 < k \leq 2)$. The transforms the original non-differentiable loss functions into infinitely differentiable ones, enabling faster gradient-based optimization techniques for solving quantile regression problems. Numerical simulations demonstrate that the proposed smooth loss functions maintain reliable accuracy while offering computational advantages. The contribution is not only the enrichment of the repertoire of non-smooth loss functions but also a provision of a new and efficient solution for regression analysis and related problems.

E0720: **Online outcome weighted learning with general loss functions**
Presenter: **Daohong Xiang**, Zhejiang Normal University, China
Co-authors: Aoli Yang, Jun Fan, Daohong Xiang

The pursuit of individualized treatment rules in precision medicine has generated significant interest due to its potential to optimize clinical outcomes for patients with diverse treatment responses. One approach that has gained attention is outcome-weighted learning. However, traditional offline learning algorithms, which process all available data at once, face limitations when applied to high-dimensional electronic health records data due to their sheer volume. Additionally, the dynamic nature of precision medicine requires that learning algorithms can effectively handle streaming data that arrives in a sequential manner. To overcome these challenges, a novel framework is presented that combines outcome-weighted learning with online gradient descent algorithms, aiming to enhance precision medicine practices. The framework provides a comprehensive analysis of the learning theory associated with online outcome-weighted learning algorithms, taking into account general classification loss functions. The convergence of these algorithms is established for the first time, providing explicit convergence rates while assuming polynomially decaying step sizes with (or without) a regularization term. Findings present a non-trivial extension of online classification to online outcome-weighted learning, contributing to the theoretical foundations of learning algorithms tailored for processing streaming input-output-reward type data.

EO061 Room 411 (Virtual sessions) STATISTICAL PROPERTIES OF EIGENSTRUCTURES IN HIGH DIMENSIONS Chair: Moritz Jirak

E0543: **Reviving pseudo-inverses: Asymptotic properties of large dimensional generalized inverses with applications**
Presenter: **Nestor Parolya**, Delft University of Technology, Netherlands
Co-authors: Taras Bodnar

The purpose is to establish a connection between modern RMT and high-dimensional statistics combined with machine learning, which is referred to as 'high-dimensional statistical learning'. Subsequently, the recent results concerning the regularized learning/estimation of large covariance matrices are presented using (Moore-Penrose) pseudoinverse and Tikhonov regularization combined with statistical shrinkage techniques. Findings contribute to constructing improved shrinkage estimators for the precision matrix, particularly in scenarios where the number of variables p is comparable to the sample size n, resulting in p/n converging to a constant $c > 1$ (singular sample covariance matrix). A real-data application in finance is concluded by, demonstrating the superiority of the proposed methods over benchmarks like nonlinear shrinkage and cross-validation techniques in machine learning.

E0920: **Tracy-Widom, Gaussian, and bootstrap: Approximations for leading eigenvalues in high-dimensional PCA**
Presenter: **Miles Lopes**, UC Davis, United States
Co-authors: Nina Doernemann

The leading eigenvalues of sample covariance matrices play a fundamental role in many aspects of high-dimensional statistics. Under certain conditions, when the data dimension and sample size diverge proportionally, these eigenvalues undergo a well-known phase transition: In the sub-critical regime, the eigenvalues have Tracy-Widom fluctuations of order $n^{-2/3}$, while in the supercritical regime, they have Gaussian fluctuations of order $n^{-1/2}$. However, the statistical problem of determining which regime underlies a given dataset has remained largely unresolved. The purpose is to develop a new testing framework and procedure to address this problem. In particular, it is demonstrated that the procedure has an asymptotically controlled level and that it is power-consistent for certain spiked alternatives. Also, this testing procedure enables the design of a new bootstrap method for approximating the distributions of functionals of the leading eigenvalues within the sub-critical regime, which is the first such method that is supported by theoretical guarantees.

E1078: **Principal component analysis and graph Laplacians in high dimensions**
Presenter: **Martin Wahl**, Bielefeld University, Germany

Given i.i.d. observations uniformly distributed on a closed manifold $M \subseteq R^p$, the spectral properties of the associated empirical graph Laplacian are studied based on a Gaussian kernel. The main results are non-asymptotic error bounds, showing that the eigenvalues and eigenspaces of the empirical graph Laplacian are close to the eigenvalues and eigenspaces of the Laplace-Beltrami operator of M. In the analysis, the empirical graph Laplacian is connected to kernel principal component analysis and considers the heat kernel of M as a reproducing kernel feature map. This leads to novel points of view and allows leveraging results for empirical covariance operators in infinite dimensions.

EC164 Room 307 FUNCTIONAL DATA ANALYSIS Chair: Pavel Krupskiy

E0326: **Two-dimensional functional mixed-effect model for repeatedly measured wearable device data**
Presenter: **Xinyue Li**, City University of Hong Kong, Hong Kong

With the rapid development of wearable device technologies, advanced accelerometers can record minute-by-minute physical activity for consecutive days. As the daily routine varies throughout the week, accelerometer data can be considered as repeatedly-measured functional observations, which smoothly vary over longitudinal visits with covariate-dependent mean and covariance functions. An innovative two-dimensional functional mixed-effect model (2DFMM) is proposed to characterize the "spatial" (longitudinal) and "temporal" (functional) structures, incorporating two-dimensional fixed effects for covariates of interest. A fast three-stage estimation procedure is also developed to provide accurate fixed-effect inference for model interpretability, and computational efficiency is improved when encountering large datasets. Extensive simulation studies are conducted to demonstrate the effectiveness of the proposed method in comparison with existing approaches. The proposed 2DFMM was further applied to Shanghai school adolescent accelerometer data, demonstrating the effectiveness and computational efficiency of 2DFMM in providing interpretable intraday and interday dynamic associations between physical activity and mental health assessments among Shanghai school adolescents, which further shed light on possible intervention strategies targeting daily physical activity patterns to improve school adolescent mental health.

E0930: **Modelling German renewable energy data**
Presenter: **Miao Yu**, Leibniz University Hannover, Germany

The purpose is to model hourly German onshore wind energy and solar power production playing a crucial role in the transmission towards greener energy systems and CO2 avoidance. Wind energy closely co-moves with wind speed, while solar power aligns with daily sunshine hours. Both wind and solar data exhibit irregular trends and significant variability. Interpreting energy production as continuous, functional data analysis allows for explaining seemingly irregular data structures. When modelling functional data, differential equations are used to both describe changes in variables over time and predict future trends. Estimating a suitable linear differential operator during model building is essential. Analyzing wind and solar data, it finds that a second-order operator best fits the data by minimizing the sum of squared residuals, which indicates that wind and solar production exhibit nonlinear dynamics with an accelerating trend. Summer data fits well with a single-coefficient second-order operator, showing minimal fluctuations. Winter data exhibits stronger fluctuations, approximating a sinusoidal curve, requiring a complex operator. Solar energy, less volatile than wind, fits well with the estimated second-order operator throughout the year. Relying only on one renewable source leads to green energy shortages due to different estimated cycles of wind and solar energy, potentially increasing CO2 intermittency.

E0943: **Factor modelling for matrix-variate functional time series in high dimensions**
Presenter: **Zihan Wang**, Tsinghua University, China
Co-authors: Dong Li, Xinghao Qiao

Nowadays, the analysis of interconnected systems is crucial across various fields, including transportation and social networks. To address this challenge, the aim is to introduce factor modelling for a new data type known as matrix-variate functional time series, which competes with existing factor modelling for tensor-time series by treating intraday observations as random functions instead of random vectors. Theoretical results on the consistency of the estimated quantities under mild conditions have been provided, and its finite-sample performances have been illustrated through extensive simulations under both fully and partially observed scenarios. Real data examples about dynamic transportation networks have been exercised to demonstrate the advantages of our proposed method in terms of flexibility, interpretability and forecasting performance compared to the tensor-based method.

Friday 19.07.2024 14:55 - 16:35 Parallel Session O – EcoSta2024

EO190 Room 102 VOLATILITY RISK AND ASSET PRICING **Chair: Xingzhi Yao**

E0169: **When MIDAS meets LASSO: The wisdom of low-frequency variables in forecasting value-at-risk and expected shortfall**
Presenter: **Yi Luo**, Xian Jiaotong-Liverpool University, China
Co-authors: Xiaohan Xue, Marwan Izzeldin
A new framework is proposed for the joint estimation and forecasting of value-at-risk (VaR) and expected shortfall (ES), integrating low-frequency variables. By maximizing the asymmetric Laplace (AL) likelihood function with an adaptive lasso penalty, the most informative variables are selected on a rolling window basis. In the empirical analysis, realized volatility, term spread, and housing starts serve as the strongest predictors of future tail risk. The out-of-sample backtesting results show that the method consistently outperforms other benchmark models and achieves the minimum loss in the joint forecasting of VaR and ES.

E0266: **Diffusive and jump risk premium in China: The role of trading mechanisms**
Presenter: **Shuyuan Qi**, Central University of Finance and Economics, China
Co-authors: Xiaoman Su
The purpose is to explore the diffusive and jump risk premiums present in the Chinese stock market, paying particular attention to the influence of trading mechanisms on risk premiums in the country. A three-step estimation method is introduced that effectively incorporates information from both physical and risk-neutral probability measures to estimate the risk premia. Notably, the Chinese stock market employs daily price limit rules and special treatment rules, which are specifically designed to uphold market stability and safeguard investors' interests. The intricacies of how these trading mechanisms shape the diffusive are delved into, and risk premiums are jumped in the Chinese stock market, highlighting their significant role in explaining the fluctuations of risk premia.

E0307: **A market-level tug of war: Investor heterogeneity and asset pricing**
Presenter: **Ran Tao**, University of Bristol, United Kingdom
Co-authors: Chardin Wese Simen, Lei Zhao
A daily tug-of-war between opposing investor clientele at the individual stock level has been documented in the asset pricing literature. A market-level tug of war is measured using the cross-sectional intensity of individual tug of war. The capital asset pricing model (CAPM) tends to perform better, and market betas are strongly and positively related to average returns on "quiet days" when the market-level tug of war is less intensive. It is further shown that the well-established findings of a robust risk-return trade-off on important information days (e.g., FOMC announcement days and influential firms earnings announcement days), and during pessimistic sentiment periods hold only when such days coincide with "quiet days". Overall, a novel explanation is provided for the CAPM's empirical failure and shows that investor disagreement has significant implications on asset pricing.

E0477: **Correlation risk premium and return predictability**
Presenter: **Xingzhi Yao**, Xián Jiaotong Liverpool University, China
Co-authors: Zhenxiong Li
The correlation risk premium, defined as the difference between the implied and realized correlation, is shown to be a robust predictor of long-term market returns. The presence of significant premium is documented across nine industrial sectors in the US market, indicating that investors are concerned about the high correlation and are willing to pay a premium in order to hedge market situations with sharp increases in correlation. In addition, it is shown that the industrial implied correlations are fractionally cointegrated, and the long-run component extracted by the co-fractional system carries a nontrivial degree of market return predictability. The empirical findings are supported by the simulation evidence.

EO136 Room 103 RECENT DEVELOPMENTS IN MULTIPLE TESTING **Chair: Bowen Gang**

E0590: **Tau-censored weighted Benjamini-Hochberg procedures under independence**
Presenter: **Huijuan Zhou**, Shanghai University of Finance and Economics, China
In the field of multiple-hypothesis testing, auxiliary information can be leveraged to enhance the efficiency of test procedures. A common way to make use of auxiliary information is by weighting p-values. However, when the weights are learned from data, controlling the finite-sample false discovery rate (FDR) becomes challenging, and most existing weighted procedures only guarantee FDR control in an asymptotic limit. In this article, two methods are introduced for constructing data-driven weights for tau-censored weighted Benjamini-Hochberg procedures under independence. They provide new insight into masking p-values to prevent overfitting in multiple tests. The first method utilizes a leave-one-out technique, where all but one of the p-values are used to learn a weight for each p-value. This technique masks the information of a p-value in its weight by calculating the infimum of the weight with respect to the p-value. The second method uses partial information from each p-value to construct weights and utilizes the conditional distributions of the null p-values to establish FDR control. Additionally, two methods are proposed for estimating the null proportion, and how to integrate null-proportion adaptivity into the proposed weights is demonstrated to improve power.

E0863: **ML-powered outlier detection: False discovery rate control and derandomization**
Presenter: **Yaniv Romano**, Technion—Israel Institute of Technology, Israel
The focus is on recent advancements in outlier (or out-of-distribution) detection, highlighting how conformal inference plays a pivotal role in creating outlier detection algorithms that control the false discovery rate. After outlining the advantages of using conformal p-values for this task, an inherent limitation of this approach is addressed: its randomized nature. Such randomness often leads to different outcomes when analyzing the same test data, complicating the interpretation of findings. To alleviate this issue, a principled solution is presented to make conformal inferences more stable by leveraging suitable conformal e-values instead of p-values to quantify statistical significance. The landscape of machine learning and multiple hypothesis testing is navigated to ensure that conclusions extracted from any complex outlier detection model are reliable, stable, and reproducible.

E1050: **Boosting e-BH via conditional calibration**
Presenter: **Zhimei Ren**, University of Pennsylvania, United States
Co-authors: Junu Lee
The e-BH procedure is an e-value-based multiple testing procedure that provably controls the false discovery rate (FDR) under any dependence structure between the e-values. Despite this appealing theoretical FDR control guarantee, the e-BH procedure often suffers from low power in practice. A general framework is proposed that boosts the power of e-BH without sacrificing its FDR control under arbitrary dependence. This is achieved by the technique of conditional calibration, where the e-values are taken as input and are calibrated to be a set of boosted e-values that are guaranteed to be no less and are often more powerful than the original ones. The general framework is explicitly instantiated in three classes of multiple testing problems: (1) testing under parametric models, (2) conditional independence testing under the model-X setting, and (3) model-free conformalized selection. Extensive numerical experiments show that the proposed method significantly improves the power of e-BH

while continuing to control the FDR.

EO104 Room 104 COMPLEX DATA: NETWORK, RANKING, AND SPATIAL PANEL DATA Chair: Wanjie Wang

E0300: **Consistent community detection in inter-layer dependent multi-layer networks**
Presenter: **Jingnan Zhang**, University of Science and Technology of China, China
Community detection in multi-layer networks, which aims at finding groups of nodes with similar connective patterns among all layers, has attracted tremendous interest in multi-layer network analysis. Most existing methods are extended from those for single-layer networks, which assume that different layers are independent. A novel community detection method is proposed in multi-layer networks with inter-layer dependence, which integrates the stochastic block model (SBM) and the Ising model. The SBM model models the community structure and the inter-layer dependence, which are incorporated via the Ising model. An efficient alternative updating algorithm is developed to tackle the resultant optimization task. Moreover, the asymptotic consistencies of the proposed method in terms of both parameter estimation and community detection are established, which are supported by extensive simulated examples and a real example of a multi-layer malaria parasite gene network.

E0416: **Homogeneity pursuit in ranking inferences based on pairwise comparison data**
Presenter: **Yuxin Tao**, Tsinghua University, China
Co-authors: Tracy Ke
The Bradley-Terry-Luce (BTL) model is one of the most celebrated models for ranking inferences based on pairwise comparison data, which associates individuals with latent preference scores and produces ranks. An important question that arises is the uncertainty quantification for ranks. It is natural to think that ranks for two individuals are not trustworthy if there is only a subtle difference in their preference scores. The homogeneity of scores in the BTL model is explored, which assumes that individuals cluster into groups with the same preference scores. The clustering algorithm in regression is introduced via data-driven segmentation (CARDS) penalty into the likelihood function, which can rigorously and automatically separate parameters and uncover group structure. Statistical properties of two versions of CARDS are analyzed. As a result, a faster convergence rate and sharper confidence intervals were achieved for the maximum likelihood estimation of preference scores, providing insight into the power of exploring low-dimensional structures in a high-dimensional setting. Real data examples are analyzed, including NBA basketball ranking and Netflix movie ranking, to highlight the method's improved prediction performance and interpretation ability.

E0815: **Temporal label recovery via manifold learning**
Presenter: **Wanjie Wang**, National University of Singapore, Singapore
Co-authors: Yuehaw Khoo, Xin Tong, Yuguan Wang
Often, dynamical data are collected without clear temporal labels. In order to recover the temporal labels from noisy data points, spectral algorithms are developed based on the graph Laplacian. The method does not require the similarity matrix to have certain monotone properties, commonly assumed in existing spectral seriation algorithms. The L-infinity error of the estimators is analyzed for the ordering, and the algorithms are applied to datasets from toy cryo-electron microscopy (cryo-EM) examples to demonstrate the efficacy of the methods.

E0490: **Multi-dimensional spatial panel data models with fixed effects: Formulation, estimation and inference**
Presenter: **Zhenlin Yang**, Singapore Management University, Singapore
The formulation is considered, estimation and inference for multi-dimensional (mD) spatial panel data (SPD) models with both observable and unobservable dimension-specific fixed effects (FEs), where the latter is of a growing dimension and may appear in the model additively or interactively of various orders. A general method of formulating the unobservable (mD) FEs so that the observable (mD) FEs can be identified is given. A general M-estimation method is proposed to estimate the common parameters, where the unbiased estimating equations are obtained by adjusting the concentrated quasi-score function with the unobserved FEs being concentrated out. The adjusted quasi-scores (AQS) remove the effects of estimating these incidental FE parameters and thereby lead to M-estimators that are consistent and asymptotically normal. The proposed methods allow (i) the identification and estimation of the effects of space or time-invariant covariates, (ii) the spatial weights and coefficients to vary with time, (iii) the error variance to vary in all dimensions, and (iv) the panels to be nested or unbalanced. Simple inference methods are introduced for each scenario studied. The asymptotic properties of these methods are studied, and finite sample performance is assessed using Monte Carlo simulations.

EO179 Room 105 RECENT ADVANCES IN INCOMPLETE DATA ANALYSIS Chair: Puying Zhao

E0588: **Modeling developmental trajectories with nonrandomly missing data**
Presenter: **Depeng Jiang**, University of Manitoba, Canada
Group-based trajectory analysis (GBTA) is commonly used for identifying distinctive developmental trajectories in longitudinal studies but may yield biased results when missing data is non-random. The aim is to compare conventional and extended GBTA in different missing data scenarios to understand their impact on trajectory identification. Simulation studies demonstrated that the extended GBTA outperformed the conventional GBTA in recovering trajectory estimates when classes were not well separated. In contrast, both methods were equally effective when classes were distinct. Applying these models to the Manitoba follow-up study data revealed four frailty trajectories, with age influencing trajectory membership. Marital status and living arrangements showed no significant associations. The extended GBTA's superiority is highlighted when handling non-random missing data, particularly in scenarios with unclear class distinctions. Valuable insights are provided for longitudinal studies with missing data, aiding researchers in understanding developmental heterogeneity and guiding future prevention strategies.

E0512: **Multiply robust estimation for two-part regression models with missing semi-continuous response**
Presenter: **Chunlin Wang**, Xiamen University, China
Co-authors: Qiyin Zheng
Two-part regression models are widely used for analyzing semicontinuous response data consisting of a mixture of excess zeros and skewed positive continuous data. The problem of missing semicontinuous response data is often encountered in many applications, and simply ignoring it may lead to erroneous conclusions. Multiple robust estimation procedures are proposed for the two-part regression parameters to allow for multiple candidate models for both the missing mechanism and imputation. The multiple robustness properties of the proposed estimators are established in the sense that they are consistent if any one of these multiple models is correctly specified. Other methods, including inverse probability weighting, imputation, and doubly robust estimators, have also been constructed and compared in simulation studies. The simulation results show the desirable finite-sample performance of the proposed multiply robust estimators under a variety of model settings. Real psychology data with the missing semicontinuous responses is analyzed for illustration.

E0526: **Equivalence assessment via the difference between two AUCs in a matched-pair design with nonignorable missing endpoints**
Presenter: **Yunqi Zhang**, Yunnan University, China
Co-authors: Weili Cheng, Puying Zhao
Equivalence assessment via various indices, such as relative risk, has been widely studied in a matched-pair design with discrete or continuous

endpoints over the past years. However, existing studies mainly focus on the fully observed or missing at random endpoints. Nonignorable missing endpoints are commonly encountered in a matched-pair design. To this end, several novel methods are proposed to assess the equivalence of two diagnostics via the difference between two correlated areas under ROC curves (AUCs) in a matched-pair design with nonignorable missing endpoints. An exponential tilting model is utilized to specify the nonignorable missing endpoint mechanism. Three nonparametric approaches and three semiparametric approaches are developed to estimate the difference between two correlated AUCs based on the kernel-regression imputation, inverse probability weighted (IPW), and augmented IPW methods. Under some regularity conditions, the consistency and asymptotic normality of the proposed estimators are shown. Simulation studies are conducted to study the performance of the proposed estimators. Empirical results show that the proposed methods outperform the complete-case method. An example from clinical studies is illustrated by the proposed methodologies.

E1014: **Robust estimation and testing for GARCH models via exponentially tilted empirical likelihood**
Presenter: **Yashuang Li**, Yunnan University, China
Co-authors: Puying Zhao, Niansheng Tang
The GARCH model has become one of the most powerful and widespread tools for dealing with time series heteroskedastic models. A commonly employed approach for inference on GARCH models is the quasi-maximum likelihood. However, unless the data are sampled regularly, the quasi-maximum likelihood estimator is inconsistent due to density misspecification or the presence of outliers. The main aim is to present a robust nonparametric likelihood analysis of GARCH models, including estimation of the coefficient parameters and model specification testing of the GARCH process. A set of identifying moment functions is specified by applying quantile regression models to the GARCH process. The moment restrictions allow the GARCH innovations to be generally distributed and are less sensitive to outliers. Exponentially tilted empirical likelihood (ETEL) is then explored to combine these quantile-related moment restrictions effectively. The ETEL framework allows for imposing over-identifying restrictions and offers implied probabilities for efficient and robust moment estimation and inference. Asymptotic properties of the resultant ETEL estimators and test statistics are investigated under mild conditions on the innovation distributions. The proposed strategies are illustrated and evaluated through numerical experiments on simulated and real datasets.

EO230 Room 106 RECENT ADVANCES IN FACTOR MODELLING AND LARGE-SCALE TIME SERIES ANALYSIS Chair: Yong He

E0306: **Modeling and learning on high-dimensional matrix-variate sequences**
Presenter: **Xu Zhang**, South China Normal University, China
Co-authors: Catherine Liu, Jianhua Guo, KC Yuen, Alan Welsh
A new matrix factor model named RaDFaM is proposed, which is strictly derived based on the general rank decomposition, and a structure of a high-dimensional vector factor model for each basis vector is assumed. RaDFaM contributes a novel class of low-rank latent structure that makes a tradeoff between signal intensity and dimension reduction from the perspective of tensor subspace. Based on the intrinsic separable covariance structure of RaDFaM, for a collection of matrix-valued observations, a new class of PCA variants is derived to estimate loading matrices and the latent factor matrices sequentially. The peak signal-to-noise ratio of RaDFaM has been proven to be superior in the category of PCA-type estimations. The asymptotic theory is also established, including the consistency, convergence rates, and asymptotic distributions for components in the signal part. Numerically, the performance of RaDFaM is demonstrated in applications such as matrix reconstruction, supervised learning, and clustering on uncorrelated and correlated data, respectively.

E0896: **CP factorization for tensor-variate time series**
Presenter: **Long Yu**, Shanghai University of Finance and Economics, China
Factor structure is used to model tensor-variate time series based on CP decomposition. The target is to identify the factor loadings up to sign change. The proposed procedure relies on eigen-analysis with a normalized and truncated auto-cross covariance. The estimator's accuracy is studied under general conditions, which allows sparse or dense loading vectors and strong or weak factors. It is also shown how to de-bias the estimator so that limiting representation is available, which will be useful in related inference problems. To further reduce the estimation error, an iterative algorithm is provided, the convergence is theoretically justified, and the accuracy of the iterative estimator is improved.

E0921: **A new non-parametric Kendall's tau for matrix-valued elliptical observations**
Presenter: **Yong He**, Shandong University, China
The aim is to propose a new non-parametric Kendall's tau for matrix-variate observations that are ubiquitous in areas such as finance and medical imaging, named row/column matrix Kendall's tau. For a random matrix following the matrix-variate elliptically contoured distribution, it is shown that the eigenspaces of the proposed row/column matrix Kendall's tau coincide with those of the row/column scatter matrix respectively, with the same descending order of the eigenvalues. To show the usefulness of the new non-parametric Kendall's tau, the focus is on a two-way dimension reduction model, namely the growing popular "matrix factor model" in the literature. Eigenvalue decomposition is performed on the generalized row/column matrix Kendall's tau to recover the loading spaces of the matrix factor model. Estimating the pair of factor numbers is also proposed by exploiting the eigenvalue ratios of the row/column matrix Kendall's tau. Theoretically, the convergence rates of the estimators are derived for loading spaces, factor scores and common components without any moment constraints on the idiosyncratic errors. Bahadur representation for the estimated loadings is also provided. The proposed methods can further be generalized to analyze high-order tensors. Thorough simulation studies are conducted to show a higher degree of robustness of the proposed estimators than the existing ones.

E0941: **Modelling matrix time series via a tensor CP-decomposition**
Presenter: **Jing He**, Southwestern University of Finance and Economics, China
Co-authors: Jinyuan Chang, Lin Yang, Qiwei Yao
The purpose is to model matrix time series based on a tensor canonical polyadic (CP)-decomposition. Instead of using an iterative algorithm, which is the standard practice for estimating CP-decompositions, a new one-pass estimation procedure is proposed based on a generalized eigenanalysis constructed from the serial dependence structure of the underlying process. To overcome the intricacy of solving a rank-reduced generalized eigenequation, a further refined approach is proposed, which projects it into a lower dimensional full-ranked eigenequation. This refined method can significantly improve the finite-sample performance. It is shown that all the component coefficient vectors in the CP-decomposition can be estimated consistently. The proposed model and the estimation method are also illustrated with both simulated and real data, showing effective dimension reduction in modelling and forecasting matrix time series.

EO253 Room 108 ADVANCE IN GENERALIZATION AND OPTIMIZATION OF MACHINE LEARNING ALGORITHMS Chair: Yunwen Lei

E0206: **Integral operator approach for spherical data fitting**
Presenter: **Shao-Bo Lin**, Xi'an Jiaotong University, China
For kernel interpolation of scattered data on spheres, it is well known that the attainable approximation error and the condition number of the underlying interpolation matrix cannot be made small simultaneously, referred to as the uncertainty phenomenon. An undesirable consequence is that kernel interpolation is susceptible to noisy data. The aim is to develop a novel integral operator approach for deterministic sampling and propose several remedies for the "uncertainty phenomenon". Based on the integral operator approach, it is proven that the popular spectral regularization, distributed learning and random sketching are feasible methods to circumvent the "uncertainty phenomenon". Numerical simulation results are also presented, showing that the mitigation methods are practical and robust in handling noisy data from challenging computing environments.

E0372: **On convergence of AdaGrad for stochastic optimization under relaxed assumptions**
Presenter: **Junhong Lin**, Zhejiang University, China
The convergence of AdaGrad with momentum is revisited (covering AdaGrad as a special case) on non-convex smooth optimization problems. A general noise model is considered where the noise magnitude is controlled by the function value gap together with the gradient magnitude. This model encompasses a broad range of noises including bounded noise, sub-Gaussian noise, affine variance noise and the expected smoothness, and it has been shown to be more realistic in many practical applications. The analysis yields a probabilistic convergence rate which, under the general noise, could reach $((\tilde{}(\infty/\tilde{\sqrt{T}}))O(1/\sqrt{T}))$. This rate does not rely on prior knowledge of problem parameters and could accelerate to $(\tilde{O}(1/T))$ where (T) denotes the total number of iterations when the noise parameters related to the function value gap and noise level are sufficiently small. The convergence rate thus matches the lower rate for stochastic first-order methods over non-convex smooth landscapes up to logarithm terms. A convergence bound for AdaGrad is further derived with momentum, considering the generalized smoothness where the local smoothness is controlled by a first-order function of the gradient norm.

E0411: **Generalization analysis of federated zeroth-order optimization**
Presenter: **Hong Chen**, Huazhong Agricultural University, China
The federated zeroth-order optimization (FedZO) algorithm enjoys the advantages of both zeroth-order optimization and federated learning and has shown exceptional performance on black-box attack and softmax regression tasks. However, there is little generalization analysis for FedZO, and its analysis on computing convergence rate is slower than the corresponding first-order optimization setting. The aim is to establish systematic theoretical assessments of FedZO by developing an analysis technique for on-average model stability. The first generalization error bound of FedZO is established under the Lipschitz continuity and smoothness conditions. Then, refined generalization and optimization bounds are provided by replacing bounded gradient with heavy-tailed gradient noise and utilizing the second-order Taylor expansion for gradient approximation. With the help of a new error decomposition strategy, the theoretical analysis is also extended to the asynchronous case. For FedZO, the fine-grained analysis fills the theoretical gap in the generalization guarantees and polishes the convergence characterization of the computing algorithm.

E0446: **Generalization analysis of gradient descent for shallow neural networks**
Presenter: **Puyu Wang**, Hong Kong Baptist University, Hong Kong
Recently, significant progress has been made in understanding the generalization of neural networks (NNs) trained by gradient descent (GD) using the algorithmic stability approach. However, most of the existing research has focused on one-hidden-layer NNs and has not addressed the impact of different network scaling. Network scaling corresponds to the normalization of the layers. The previous work is greatly extended by conducting a comprehensive stability and generalization analysis of GD for two-layer and three-layer NNs. For two-layer NNs, our results are established under general network scaling, relaxing previous conditions. In the case of three-layer NNs, our technical contribution lies in demonstrating its nearly co-coercive property by utilizing a novel induction strategy that thoroughly explores the effects of over-parameterization. As a direct application of our general findings, the excess risk rate of $O(1/\sqrt{n})$ is derived for GD in both two-layer and three-layer NNs. This sheds light on sufficient or necessary conditions for under-parameterized and over-parameterized NNs trained by GD to attain the desired risk rate of $O(1/\sqrt{n})$. Additionally, under a low-noise condition, a fast risk rate of $O(1/n)$ is obtained for GD in both two-layer and three-layer NNs.

EO236 Room 109 RECENT DEVELOPMENTS IN THEORY AND APPLICATIONS OF STATISTICAL LEARNING Chair: Jun Fan

E0795: **Generalization analysis of deep CNNs under maximum correntropy criterion**
Presenter: **Zhiying Fang**, Shenzhen Polytechnic University, China
Co-authors: Jun Fan, Yingqiao Zhang
Convolutional neural networks (CNNs) have gained immense popularity in recent years, finding their utility in diverse fields such as image recognition, natural language processing, and bioinformatics. Despite the remarkable progress made in deep learning theory, most studies on CNNs, especially in regression tasks, tend to rely heavily on the least squares loss function. However, there are situations where such learning algorithms may not suffice, particularly in the presence of heavy-tailed noises or outliers. This predicament emphasizes the necessity of exploring alternative loss functions that can handle such scenarios more effectively, thereby unleashing the true potential of CNNs. The generalization error of deep CNNs is investigated with the rectified linear unit (ReLU) activation function for robust regression problems within an information-theoretic learning framework. It is demonstrated that when the regression function exhibits an additive ridge structure, and the noise possesses a finite p-th moment, the empirical risk minimization scheme, generated by the maximum correntropy criterion and deep CNNs, achieves fast convergence rates. Notably, these rates align with the mini-max optimal convergence rates attained by a fully connected neural network model with the Huber loss function up to a logarithmic factor.

E0799: **Generalization analysis of deep ReLU networks for functional learning**
Presenter: **Linhao Song**, Central South University, China
Learning nonlinear functionals defined on $L^p([-1,1]^s)$ for $1 \leq p \leq \infty$ and $s \in \mathbb{N}$ is a significant learning task in broad applications. As a powerful tool of the nonparametric approach, neural networks that take functions as input were recently designed and employed to learn nonlinear functionals and achieved great success in practice. However, the underlying theoretical analysis of this approach lags heavily behind. For instance, the universal consistency and learning rates remain open. The least-squares regression problem is considered using functional neural networks with the rectified linear unit (ReLU) activation function. Within the framework of learning theory, it is shown that the learning algorithm is universally consistent. Besides, generalization error bounds are also established, showing a trade-off between the approximation ability and the capacity of the approximants measured by covering numbers. Based on this, the learning rates are further investigated under diverse assumptions on the target functional and the input function space.

E0803: **Covariance test for discretely observed functional data: When and how it works**
Presenter: **Yang Zhou**, Beijing Normal University, China
For covariance tests in functional data analysis, the existing methods are only developed for fully observed curves, while in reality, one observes such trajectories discretely with noise. To bridge this gap, a projection-based test statistic is constructed and allows the number of estimated

eigenfunctions potentially to grow with sample size, leading to a consistent nonparametric test with challenges arising from the concurrence of the diverging truncation and discretized observations. The pooling method and sample-splitting strategy are used to attain the test statistic and derive its asymptotic Chi-squared null distribution facilitated by advancing the perturbation bound of estimated principal components. The theoretic analysis reveals an interesting connection between the permissible truncation level, the sampling frequency and the sample size. It is shown that the asymptotic null distribution remains valid for different allowable ranges of truncation level, and when the sampling frequency reaches a certain magnitude of the sample size, it behaves as if the functions are fully observed. This investigation provides for the first time the theoretical justification and practical guidance on when and how the covariance test procedure works by allowing growing truncation levels for discretely observed functional data that range from "sparse" to "dense" paradigms.

E0781: **Regularized reduced-rank regression for structured output prediction with vector-valued functions**
Presenter: **Kun Cheng**, Beijing Jiaotong University, China
In predicting multiple response variables from the predictor variable, the reduced-rank regression (RRR) is an effectively linear method that implies that the matrix of regression coefficients is of lower rank. The focus is on exploiting RRR with a reproducing kernel Hilbert space (RKHS) approach. It models the multiple response variables as a linear combination of a few vectors with coefficients being (possibly nonlinear) functions of the predictor variable. In the underlying setting, coefficient functions are chosen from RKHS, while in RRR, they are restricted to linear functions. A set of solutions in RKHSs is characterized by the help of cross-covariance operators in RKHSs. Moreover, regularized estimators are constructed, and estimation errors are bounded by mild assumptions. A convergence rate for estimation errors is established. Simulations and real data analysis are provided to illustrate the efficiency of the proposed method.

EO099 Room 110 RECENT ADVANCES IN DESIGN AND MODELING FOR COMPLEX EXPERIMENTS Chair: Yaping Wang

E0250: **Stratification pattern enumerator and its applications**
Presenter: **Ye Tian**, Beijing University of Posts and Telecommunications, China
Space-filling designs are widely used in computer experiments. A minimum aberration-type space-filling criterion was recently proposed to rank and assess a family of space-filling designs, including orthogonal array-based Latin hypercubes and strong orthogonal arrays. However, it is difficult to apply the criterion in practice because it requires intensive computation to determine the space-filling pattern, which measures the stratification properties of designs on various subregions. A stratification pattern enumerator is proposed to characterise the stratification properties. The enumerator is easy to compute and can efficiently rank space-filling designs. It is shown that the stratification pattern enumerator is a linear combination of the space-filling pattern. Based on the connection, efficient algorithms are developed to calculate the space-filling pattern. In addition, a lower bound is established for the stratification pattern enumerator and present construction methods for designs that achieve the lower bound using multiplication tables over Galois fields. The constructed designs have good space-filling properties in low-dimensional projections and are robust under various criteria.

E0251: **Generalized linear orthogonal arrays and applications to strong orthogonal arrays**
Presenter: **Bochuan Jiang**, Beijing Jiaotong University, China
Orthogonal arrays are essential for designing experiments and have found extensive applications in the big data era. A new class of orthogonal arrays called generalized linear orthogonal arrays (GLOAs), is introduced, which encompasses linear orthogonal arrays, nonlinear orthogonal arrays from the Addelman-Kempthorne construction, and numerous other nonlinear orthogonal arrays as special cases. The most compelling feature of GLOAs is their highly deterministic projection properties, which facilitate the study of the complex relationships among columns of GLOAs. These properties are leveraged to explore the applications of GLOAs in computer experiments. Two methods for constructing strong orthogonal arrays (SOAs) and column-orthogonal SOAs of strength two plus are presented using GLOAs. These methods produce space-filling designs capable of accommodating a large number of factors, providing significant flexibility in terms of run sizes, and possessing appealing low-dimensional projection properties. Therefore, these designs are ideal for computer experiments.

E0578: **Nearly orthogonal Latin hypercube designs with multi-dimensional stratifications**
Presenter: **Hui Li**, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, China
Orthogonal Latin hypercube design, as a useful class of Latin hypercube design (LHD), plays a significant role in computer experiments. Two general methods are provided to construct nearly orthogonal LHDs. All column pairs of the generated designs can achieve stratifications on s -s grids when projected onto any two dimensions. Moreover, the designs generated by the second method can achieve stratifications on s-s-s grids in some three dimensions. The second method is further extended to construct nearly orthogonal designs, and the resulting designs enjoy good stratifications in two and three dimensions. Many new designs with good stratifications are constructed and tabulated.

E0631: **A new approach to optimal design under model uncertainty motivated by multi-armed bandits**
Presenter: **Zhengfu Liu**, Beijing Institute of Technology, China
Co-authors: Mingyao Ai, Holger Dette, Jun Yu
Optimal designs are usually model-dependent and likely to be sub-optimal if the postulated model is not correctly specified. In practice, it is common for a researcher to have a list of candidate models at hand, and a design that is efficient for both model discrimination and parameter estimation in the (unknown) true model has to be found. A reinforcement learning approach is used to achieve a balance between these competing goals in the design of experiments. A sequential algorithm is developed to provide a design that has asymptotically the same performance as an optimal design when the true model could be correctly specified in advance. A lower bound is established to quantify the relative efficiency between such a design and an optimal design for the true model in finite stages. Moreover, the resulting designs are also efficient for discriminating between the true model and other rival models from the candidate list. Some connections with other state-of-the-art algorithms for model discrimination and parameter estimation are discussed, and the methodology is illustrated by a small simulation study.

EO045 Room 111 FACTOR MODELS AND SEMIPARAMETRIC MODELS WITH APPLICATIONS Chair: Mengying You

E0975: **Joint dimension reduction and spatial clustering for spatial transcriptomics data analysis**
Presenter: **Wei Liu**, Sichuan University, China
Dimension reduction and spatial clustering are usually performed sequentially; however, the low-dimensional embeddings estimated in the dimension-reduction step may not be relevant to the class labels inferred in the clustering step. A dimension-reduction spatial-clustering (DR-SC) computation method has been developed to perform dimension reduction and (spatial) clustering simultaneously within a unified framework. Joint analysis by DR-SC produces accurate (spatial) clustering results and ensures the effective extraction of biologically informative low-dimensional features. DR-SC applies to spatial clustering in spatial transcriptomics, which characterizes the spatial organization of the tissue by segregating it into multiple tissue structures. With comprehensive simulations and real data applications, it is shown that DR-SC outperforms existing clustering and spatial clustering methods: it extracts more biologically relevant features than conventional dimension reduction methods, improves clustering performance, and offers improved trajectory inference and visualization for downstream trajectory inference analyses.

E0991: **Semiparametric mixed effect state space model with prior information for state-level COVID-19 prediction**
Presenter: **Mengying You**, Shanghai University of International Business and Economics, China

A state-space approach is introduced to model COVID-19 new case dynamics at the U.S. state level. The model incorporates prior information about curve shape through general smoothing splines. By utilizing a mixed-effects model, information between similarly shaped states is shared, and extra variability is allowed through random effect functions. An efficient state-space formulation of the model enables fast computation and excellent prediction of future observations. Compared to the cubic spline time series model using state-level COVID-19 data, the approach demonstrates a superior ability to extract common patterns and provide more reliable forecasts with smaller prediction intervals.

E0998: **Semiparametric maximum likelihood estimation with two-phase stratified case-control sampling**
Presenter: **Yaqi Cao**, Minzu University of China, China
Co-authors: Ying Yang, Jinbo Chen

In the two-phase stratified case-control sampling design, some covariates are available only for a subset of cases and controls, which are selected based on the outcome and fully collected covariates. The analysis often focuses on fitting a logistic regression model to describe the relationship between the outcome and all covariates. The interest also lies in characterizing the distribution of incomplete covariates conditional on fully observed ones in the underlying population, which is required for quantifying the predictive accuracy of the fitted model. It is desirable to include all subjects in the analysis to achieve consistency and efficiency of parameter estimation. A novel semiparametric maximum likelihood approach is proposed under rare disease assumption, where estimates are obtained through a novel reparametrized profile likelihood technique. The large sample theory is developed for the proposed estimator, showing through simulation that it has improved efficiency compared with the existing approach. The method is applied to the Breast Cancer Detection and Demonstration Project data, where one risk predictor, breast density, was measured only for a subset of study women.

E0992: **High-dimensional covariate-augmented overdispersed Poisson factor model**
Presenter: **Qingzhi Zhong**, Jinan University, China

The current Poisson factor models often assume that the factors are unknown, which overlooks the explanatory potential of certain observable covariates. The focus is on high dimensional settings, where the number of the count response variables and/or covariates can diverge as the sample size increases. A covariate-augmented overdispersed Poisson factor model is proposed to jointly perform a high-dimensional Poisson factor analysis and estimate a large coefficient matrix for overdispersed count data. A group of identifiability conditions are provided to guarantee computational identifiability theoretically. The interdependence of both response variables and covariates is incorporated by imposing a low-rank constraint on the large coefficient matrix. A novel variational estimation scheme that combines Laplace and Taylor approximations is proposed to address the computational challenges posed by nonlinearity, two high-dimensional latent matrices, and the low-rank constraint. A criterion based on a singular value ratio is also developed to determine the number of factors and the rank of the coefficient matrix. Comprehensive simulation studies demonstrate that the proposed method outperforms the state-of-the-art methods in estimation accuracy and computational efficiency. An application to the CITE-seq dataset demonstrates the practical merit of the method.

EO162 Room 212 EXTREME VALUE STATISTICS IN TIME AND SPACE Chair: Gilles Stupfler

E0862: **Bayesian analysis of peaks over threshold**
Presenter: **Stefano Rizzelli**, University of Padova, Italy
Co-authors: Simone Padoan, Clement Dombry

The peaks over threshold (POT) method is one of the most popular approaches to statistical analysis of univariate extremes. Even if Bayesian methods in this context have found increasing use in applications, no asymptotic guarantee has been established. Contraction rates and asymptotic normality of posterior distributions for Bayesian inference are established using POTs, providing contraction rates for predictive distributions and also allowing (random) covariates to be included in the analysis. Simulations show that the properties established in the large sample theoretical study can be observed in practice for finite samples of moderate size.

E0618: **ANOVEX: Analysis of variability for heavy-tailed extremes**
Presenter: **Antoine Usseglio-Carleve**, Avignon Universita, France
Co-authors: Stephane Girard, Thomas Opitz

Analysis of variance (ANOVA) is commonly employed to assess differences in the means of independent samples. However, it is unsuitable for evaluating differences in tail behaviour, especially when means do not exist or empirical estimation of moments is inconsistent due to heavy-tailed distributions. An ANOVA-like decomposition is proposed to analyse tail variability, allowing for flexible representation of heavy tails through a set of user-defined extreme quantiles, possibly located outside the range of observations. Building on the assumption of regular variation, a test is introduced for significant tail differences among multiple independent samples and its asymptotic distribution is derived. The theoretical behavior is investigated by the test statistics for the case of two samples, each following a Pareto distribution, and explore strategies for setting hyperparameters in the test procedure. To demonstrate the finite-sample performance, simulations highlight generally reliable test behavior for a wide range of situations. The test is applied to identify clusters of financial stock indices with similar extreme log returns and to detect temporal changes in daily precipitation extremes at rain gauges in Germany.

E0798: **Local stationarity in the extremes**
Presenter: **Manon Felix**, University of Geneva, Switzerland
Co-authors: Davide La Vecchia, Gilles Stupfler

The purpose is to present a novel approach that bridges extreme theory and non-stationary theory, aiming to characterize the concept of E-local stationarity. Unlike conventional literature on locally stationary processes, the proximity between a non-stationary process and its stationary approximation cannot be defined solely in terms of the Lp norm due to the potential non-existence of the first moment. Consequently, the methodology relies on pre-asymptotic measures. By leveraging extreme theory applicable to strictly stationary and strongly mixing (potentially multivariate) processes, we develop a localized extremogram in time. Conditions are established to ensure the asymptotic normality of the empirical extremogram, which is weighted by a kernel and centered by a pre-asymptotic version. Furthermore, the properties of the frequency domain analog of the extremogram are derived within the framework.

E0581: **Analysis of variability in extremes with application in change point detection**
Presenter: **Chen Yan**, INRAE/Inria, France

ANOVA is a widely used statistical method for comparing means across several groups. However, analyzing tail behavior rather than the mean can be more insightful in certain contexts. ANOVEX (ANalysis Of Variability in EXtremes) is introduced, a novel approach designed to compare extreme behaviors across J-1 groups. ANOVEX evaluates extreme quantiles within each group, utilizing methods like the Weissman estimator, and assesses the variances of extreme log-quantiles both within and between groups. Under the hypothesis of identical extreme behavior across groups, the variance ratio is shown to asymptotically follow a chi-square distribution with J-1 degrees of freedom. Additionally, ANOVEX's utility extends to change point detection and estimation in contexts divergent from traditional mean and variance shifts, particularly focusing on detecting and

estimating changes in distributions' tails. This aspect broadens ANOVEX's applicability, offering a robust tool for analyzing extremes in various data-intensive fields.

EO312 Room 202 RECENT ADVANCES IN STATISTICAL LEARNING Chair: Xiaohang Wang

E0330: **Transporting randomized trial results to estimate counterfactual survival functions in the target populations**
Presenter: **Zhiqiang Cao**, Shenzhen Technology University, China

When the distributions of treatment effect modifiers differ between a randomized trial and an external target population, the sample average treatment effect in the trial may be substantially different from the target population average treatment. Despite the increasingly rich literature on transportability, little attention has been devoted to methods for transporting trial results to estimate counterfactual survival functions in target populations when the primary outcome is time to event and subject to right censoring. Doubly robust estimators are studied to estimate counterfactual survival functions and the target average survival treatment effect in the target population, and their respective approximate variance estimators are provided. The focus is on a common scenario where the target population information is observed only through a complex survey and elucidates how the survey weights can be incorporated into each estimator considered. Simulation studies are conducted to examine the finite-sample performances of the proposed estimators in terms of bias, efficiency and coverage under both correct and incorrect model specifications. Finally, the proposed method is applied to assess the transportability of the results in the action to control cardiovascular risk in diabetes blood pressure (ACCORD BP) trial to all adults with diabetes in the United States.

E0382: **Feasibility probability of random linear programming**
Presenter: **Wei Zhang**, South China University of Technology, China

The linear feasibility problem, that is, $Ax = b, x > 0$ is basic in linear programming, statistical physics and biological mathematics. When the system has a solution, i.e., the system is feasible. It is noticed that when the elements of matrix A and vector b are stochastically generated according to some probability distribution, there is a transition between feasibility and infeasibility. Especially when the mean value and variance of the distribution vary or when the number of rows and columns of A varies, the fraction of the feasible instances will change accordingly. An analytical method is given to derive the feasibility probability of the system to tell how many random instances have a solution.

E0501: **A comparative study on federated learning in supply chain forecasting: Pareto optimality on accuracy and efficiency**
Presenter: **Jingjin Wu**, BNU-HKBU United International College, China
Co-authors: Hang Qi, Jieping Luo

Federated learning (FL) is an emerging learning mechanism that can achieve accuracy, efficiency, and privacy concurrently, which could be particularly useful in scenarios where large volumes of sensitive data are involved, such as supply chain management and forecasting. A dataset composed of sales records of multiple products across five different regions globally is considered, and a comparative study of centralized learning, distributed learning, and FL focuses on the accuracy of predicting future demand for certain products and the required volume of data for transmission by the multi-layer perception (MLP) model. The results show that MLP with FL has the fastest convergence rate; that is, it can achieve the highest accuracy in predicting the future demands of certain products compared to the other two learning approaches, given the same number of training rounds. In addition, the performances of FL approaches are compared with different combinations of market data sets across regions, and the Pareto optimality is examined in terms of accuracy and transmission efficiency under different scenarios. The potential transformative impact of FL is demonstrated on global supply chain management in various aspects.

E0694: **Application of artificial intelligence for medical big data**
Presenter: **Liyilei Su**, Shenzhen Technology University, China
Co-authors: Bingding Huang

In recent years, with the development of deep learning and other technologies, artificial intelligence (AI) has achieved significant results in a number of fields, such as image recognition, natural language processing, and decision support. In medical image big data, automatic diagnosis of medical images is currently one of the most worthwhile research directions; with the help of complex data classification and pattern recognition technology, the disease diagnosis system can quickly and accurately identify the possible disease risks from the massive amount of medical data. AI can be applied to all aspects, from diagnosis to treatment, such as image acquisition and processing, assisted reporting, follow-up planning, data storage, data mining, and so on. Compared with traditional manual analysis methods, machine learning-based disease diagnosis systems can not only dramatically improve the efficiency of disease diagnosis but also significantly improve the accuracy of diagnostic results. Due to its accuracy and wide range of applications, AI can have a huge impact on medical science, assisting doctors in the diagnosis and treatment of patients and changing the traditional medical mode, making automated and intelligent medical treatment possible, speeding up the diagnosis time and reducing subjective judgement based on the medical big data.

EO132 Room 204 RECENT RESULTS IN COMPUTATIONAL STATISTICS AND FINANCIAL TIME SERIES Chair: Minh-Ngoc Tran

E0267: **Data scaling effect of deep learning in financial time series forecasting**
Presenter: **Chen Liu**, The University of Sydney, Australia
Co-authors: Minh-Ngoc Tran, Chao Wang, Richard Gerlach, Robert Kohn

For many years, researchers have been exploring the use of deep learning in the forecasting of financial time series. However, they have continued to rely on the conventional econometric approach for model optimization, optimizing the deep learning models on individual assets. The stock volatility forecast is used as an example to illustrate global training - optimizes the deep learning model across a wide range of stocks - is both necessary and beneficial for any academic or industry practitioners who is interested in employing deep learning to forecast financial time series. Furthermore, a pre-trained foundation model for volatility forecast is introduced, capable of making accurate zero-shot forecasts for any stocks.

E1019: **Hierarchical spatial copula models for large spatial data**
Presenter: **David Gunawan**, University of Wollongong, Australia
Co-authors: Alan Pearse, Noel Cressie

A fully Bayesian, hierarchical spatial statistical model that incorporates copulas in the process model is presented. These models have several novel features compared to other spatial copula models: The spatial copula models explicitly account for measurement error and allow the noisy and incomplete spatial data to be distinguished from the underlying process; spatial change-of-support is addressed by modelling the spatial process at the level of basic areal units (BAUs) that discretise a spatial domain into a finite number of fine resolution areas; and spatial copula models are developed for large spatial datasets using ideas from fixed rank Kriging. Additionally, it is shown that, when building spatial copula models on point-level spatial support, some choices of copula may lead to violations of Kolmogorov consistency and thus fail to define a valid spatial stochastic process. The models are constructed to avoid such pitfalls for all choices of copula function. It is confirmed that full Bayesian inference on these models is feasible and yields accurate and valid inferences via a simulation study. An illustrative application to remotely sensed atmospheric methane is also presented.

E1105: **Calibrated generalized Bayesian inference**
Presenter: **Robert Kohn**, University of New South Wales, Australia
Co-authors: David Frazier, Christopher Drovandi
The aim is to provide a simple and general solution to the fundamental open problem of inaccurate uncertainty quantification of Bayesian inference in misspecified or approximate models and of generalized Bayesian posteriors more generally. While existing solutions are based on explicit Gaussian posterior approximations or computationally onerous post-processing procedures, correct uncertainty quantification is demonstrated as achievable by substituting the usual posterior with an alternative posterior that conveys the same information. This solution applies to both likelihood-based and loss-based posteriors, and the reliable uncertainty quantification of this approach is formally demonstrated. The new approach is demonstrated through a range of examples, including generalized linear models and doubly intractable models.

E1091: **Natural gradient variational Bayes without Fisher matrix analytic calculation and its inversion**
Presenter: **Minh-Ngoc Tran**, University of Sydney, Australia
Co-authors: Antoine Godichon-Baggioni, Duy Nguyen, Minh-Ngoc Tran
The purpose is to introduce a method for efficiently approximating the inverse of the Fisher information matrix, a crucial step in achieving effective variational Bayes inference. A notable aspect of the approach is the avoidance of analytically computing the Fisher information matrix and its explicit inversion. Instead, an iterative procedure is introduced to generate a sequence of matrices that converge to the inverse of Fisher information. The natural gradient variational Bayes algorithm without analytic expression of the Fisher matrix and its inversion is provably convergent. It achieves a convergence rate of order O(logs/s), with s the number of iterations. A central limit theorem for the iterates is also obtained. The implementation of the method does not require the storage of large matrices, and it achieves a linear complexity in the number of variational parameters. The algorithm exhibits versatility, making it applicable across a diverse array of variational Bayes domains, including Gaussian approximation and normalizing flow Variational Bayes. A range of numerical examples is offered to demonstrate the efficiency and reliability of the proposed variational Bayes method.

EO239 Room 207 CAUSAL INFERENCE AND MACHINE LEARNING FOR SURVIVAL ANALYSIS Chair: Wenbo Wu

E0653: **Meta-learners to analyze treatment heterogeneity in survival data**
Presenter: **Ying Ding**, University of Pittsburgh, United States
An important aspect of precision medicine focuses on characterizing diverse responses to treatment due to unique patient characteristics, also known as heterogeneous treatment effects (HTE), and identifying beneficial subgroups with enhanced treatment effects. Estimating HTE with right-censored data in observational studies remains challenging. A meta-learner-based procedure is proposed with pseudo-outcomes for analyzing HTE in survival data, which includes a pseudo-outcome-based meta-learner framework for estimating HTE, a variable importance metric for identifying predictive variables to HTE, and a data-adaptive procedure to select subgroups with enhanced treatment effects. The proposed procedure is applied to analyze subgroup treatment heterogeneity of a written asthma action plan (WAAP) on time-to-ED (Emergency Department) return due to asthma exacerbation, using a large EHR dataset with visit records expanded from pre- to post-COVID-19 pandemic. Vulnerable subgroups of patients are identified as having poorer asthma outcomes but enhanced benefits from WAAP and characterized patient profiles. The research offers valuable insights for healthcare policymakers and providers in advocating influenza vaccination and strategic WAAP distribution to particularly vulnerable groups during a disruptive public health event, ultimately enhancing pediatric asthma control.

E0422: **Bayesian semiparametric model for sequential treatment decisions with informative timing**
Presenter: **Arman Oganisian**, Brown University, United States
A Bayesian semiparametric model is developed for the impact of dynamic treatment rules on survival among patients diagnosed with pediatric acute myeloid leukemia (AML). The data are from a phase III clinical trial in which patients move through a sequence of four treatment courses. At each course, they undergo treatment that may or may not include anthracyclines (ACT). While ACT is known to be effective at treating AML, it is also cardiotoxic and can lead to early death for some patients. The task is to estimate the potential survival probability under hypothetical dynamic ACT treatment strategies, but there are several impediments. First, since ACT is not randomized, its effect on survival is confounded over time. Second, subjects initiate the next course depending on when they recover from the previous course, making timing potentially informative of subsequent treatment and survival. Third, patients may die or drop out before ever completing the full treatment sequence. A generative Bayesian semiparametric model is developed based on Gamma process priors to address these complexities. At each treatment course, the model captures subjects' transition to subsequent treatment or death in continuous time. G-computation is used to compute a posterior over potential survival probability that is adjusted for time-varying confounding. Using the approach, the efficacy of hypothetical treatment rules that dynamically modify ACT is estimated based on evolving cardiac function.

E0327: **A flexible Bayesian g-formula for causal survival analyses with time-dependent confounding**
Presenter: **Fan Li**, Yale University, United States
Co-authors: Xinyuan Chen, Liangyuan Hu
In longitudinal observational studies with a time-to-event outcome, a common objective in causal analysis is to estimate the causal survival curve under hypothetical intervention scenarios within the study cohort. The g-formula is a particularly useful tool for this analysis. To enhance the traditional parametric g-formula approach, a more adaptable Bayesian g-formula estimator is developed. It incorporates Bayesian additive regression trees in the modeling of the time-evolving generative components, aiming to mitigate bias due to model misspecification. Specifically, a more general class of g-formulas is introduced for discrete survival data. These formulas can incorporate the longitudinal balancing scores, which serve as an effective method for dimension reduction and are vital when dealing with an expanding array of time-varying confounders. The minimum sufficient formulation of these longitudinal balancing scores is linked to the nature of treatment regimes, whether static or dynamic. For each type of treatment regime, posterior sampling algorithms are provided, which are grounded in the Bayesian additive regression trees framework. Simulation studies are conducted to illustrate the empirical performance of the proposed Bayesian g-formula estimators and compare them with existing parametric estimators. The practical utility of the methods is further demonstrated in real-world scenarios using data from the Yale New Haven Health System's electronic health records.

E0752: **Debiased lasso for stratified Cox models with application to the national kidney transplant data**
Presenter: **Lu Xia**, Michigan State University, United States
Co-authors: Bin Nan, Yi Li
The scientific registry of transplant recipients (SRTR) system has become a rich resource for understanding the complex mechanisms of graft failure after kidney transplant, a crucial step for allocating organs effectively and implementing appropriate care. As transplant centers that treated patients might strongly confound graft failures, Cox models stratified by centers can eliminate their confounding effects. Also, since recipient age is a proven nonmodifiable risk factor, a common practice is to fit models separately by recipient age groups. The moderate sample sizes, relative to the number of covariates, in some age groups, may lead to biased maximum stratified partial likelihood estimates and unreliable confidence intervals, even when samples still outnumber covariates. To draw reliable inferences on a comprehensive list of risk factors measured from both donors and recipients in SRTR, a debiased lasso approach is proposed via quadratic programming for fitting stratified Cox models. Asymptotic

properties are established, and the method is verified via simulations to produce consistent estimates and confidence intervals with nominal coverage probabilities. By accounting for nearly 100 confounders in SRTR, results from the proposed method may inform the refinement of donor-recipient matching criteria for stakeholders.

EO237 Room 210 NON-PARAMETRIC STATISTICAL METHODS FOR COMPLEX BIOMEDICAL DATA Chair: Haochang Shou

E0191: **Improving the reproducibility of brain imaging feature selection with weighted regularization**
Presenter: **Fengqing Zhang**, Drexel University, United States

Advances in brain imaging techniques and machine learning (ML) models allow researchers to combine many brain imaging features to aid the diagnosis of psychiatric disorders jointly. Increasingly, the reproducibility of ML results has drawn great attention. Studies examining the reproducibility problem in brain imaging have largely focused on prediction accuracy. However, achieving high prediction accuracy and discovering relevant features are not necessarily the same. An important yet under-investigated problem is the reproducibility of feature selection in brain imaging studies. A new metric is proposed to quantify the reproducibility of neuroimaging feature selection via bootstrapping. The reproducibility index (R-index) is estimated for each feature as the reciprocal coefficient of variation of absolute mean difference across a larger number of bootstrap samples. The R-index in regularized classification models is then integrated as penalty weight. Reproducible features with a larger R-index are assigned smaller penalty weights and, thus, are more likely to be selected by the proposed models. The performance of the proposed models is evaluated using both simulated and multimodal neuroimaging data.

E0639: **Principal component analysis in Bayes spaces for sparsely sampled density functions**
Presenter: **Sonja Greven**, Humboldt University of Berlin, Germany
Co-authors: Lisa Steyer

A novel approach to functional principal component analysis in Bayes spaces is presented in the setting where densities are the object of analysis, but only a few individual samples from each density are observed. The observed data is used directly to account for all sources of uncertainty instead of relying on the prior estimation of underlying densities in a two-step approach, which can be inaccurate if small or heterogeneous numbers of samples per density are available. The approach is based on Bayes spaces, which extend the Aitchison geometry for compositional data to density functions to account for the constrained nature of densities. The isometric isomorphism is exploited between the Bayes space and the $L2$ subspace $L2_0$ with integration-to-zero constraint through the centered log-ratio transformation. As only discrete draws from each density are observed, the underlying functional densities are treated as latent variables within a maximum likelihood framework, and a Monte Carlo expectation maximization algorithm is employed for model estimation. Resulting estimates are useful for exploratory analyses of density data, for dimension reduction in subsequent analyses, and for improved preprocessing of sparsely sampled density data compared to existing methods. The proposed method is applied to analyze the distribution of maximum daily temperatures in Berlin during the summer months for the last 70 years and distributions of rental prices in Munich districts.

E0719: **Integration of longitudinal physical activity data from multiple sources**
Presenter: **Jingru Zhang**, Fudan University, China
Co-authors: Haochang Shou, Hongzhe Li

As various devices have been developed to collect physical activity data, a critical problem is how to integrate datasets across different conditions to better understand the characterization of physical activity. The key to this problem is to remove site effects while maintaining common features. However, since wearable sensor devices are deployed to record physical activity minute-by-minute continuously over multiple days, the longitudinal time-dependent structure makes the integration challenging. A new method is proposed to integrate longitudinal physical activity datasets, which model the shared information by common eigenvalues and eigenfunctions while allowing for site-specific scale and rotation. The proposed method is applied to NHANES datasets with different types of wearable sensors. The results demonstrate the method's superiority in removing site effects while preserving biological signals compared to existing approaches. A framework is developed for integrating longitudinal time-dependency datasets and provides insights into the analysis of physical activity data.

EO077 Room 307 RECENT PROGRESS ON FUNCTIONAL DATA ANALYSIS Chair: Weichi Wu

E0939: **Communication-efficient distributed functional regression**
Presenter: **Heng Lian**, City university of Hong kong, Hong Kong

The distributed estimation is considered via the communication-efficient surrogate loss approach of prior studies for functional linear regression in a unified framework, including functional generalized linear regression and functional quantile linear regression. The study of distributed functional regression is a methodologically important and technically non-trivial extension of the method to estimation in an infinite-dimensional Hilbert space. Statistical rates of the distributed estimator that match the rates of the standard estimator are established using all data.

E0954: **Functional-edged network modeling**
Presenter: **Chen Zhang**, Tsinghua University, China

Contrasting with existing works, which all consider nodes to be functions and use edges to represent the relationships between different functions, the target is network modelling, whose edges are functional data and transform the adjacency matrix into a functional adjacency tensor, introducing an additional dimension dedicated to function representation. Tucker functional decomposition is used for the functional adjacency tensor, and to further consider the community between nodes, the basis matrices are regularized to be symmetrical. Furthermore, to deal with irregular observations of the functional edges, model inference is conducted to solve a tensor completion problem. It is optimized by a Riemann conjugate gradient descent method. Besides these, several theorems are also derived to show the desirable properties of the functional edged network model. Finally, the efficacy of the proposed model is evaluated using simulation data and real metro system data from Hong Kong and Singapore.

E0996: **Statistical inference for functional data over high-dimensional domain**
Presenter: **Lijian Yang**, Tsinghua University, China
Co-authors: Qirui Hu

The purpose is to develop inference tools for the mean function of functional data over a high-dimensional domain. Tensor product spline is used to recover individual trajectories, leading to an efficient two-step mean estimator, meaning that it is asymptotically indistinguishable from the infeasible estimator using unobservable trajectories. A data-driven SCR with preset asymptotic coverage and uniformly adaptive width of order $n^{-1/2}$ is established, supported by consistent estimates of covariance function and quantile of the maximal deviation process. The asymptotic theory extends to two samples without extra difficulty. Extensive Monte Carlo experiments corroborate the theory, and satellite ocean data collected by CMEMS illustrates how the proposed SCR is used.

E1005: **A nonlinear mixed-effects functional regression model based on variable selection**
Presenter: **Yan-fu Li**, Tsinghua University, China

The mixed-effects functional regression (MFR) model offers a valuable tool for analyzing dynamic data with individual-specific variations. However, challenges arise in scenarios with non-linear relationships and variable selection among covariates. A novel extension to the MFR model is proposed. The approach incorporates non-linear components using bivariate splines, enabling a robust framework for complex relationship modeling. For variable selection in high-dimensional regression, a group-minimax concave penalty (MCP) that treats parameters from the same spline basis as a group is employed, ensuring accurate and unbiased selection. This methodology allows for estimating fixed and random effects in a two-step procedure. A flexible and comprehensive framework accommodates non-linear covariates and an MCP variable selection approach. Empirical validations and theoretical justifications support the effectiveness of our proposed methodology. In summary, the approach provides a versatile and efficient tool for modeling functional responses in the presence of non-linear relationships and mixed effects.

EO200 Room 313 RECENT DEVELOPMENTS IN TIME SERIES ANALYSIS AND RELATED TOPICS Chair: Sangyeol Lee

E0341: **Maximum likelihood estimation of elliptical multivariate regular variation**
Presenter: **Moosup Kim**, Keimyung University, Korea, South

The focus is on the efficient estimation of the elliptical tail. Initially, the density function of the spectral measure of an elliptical distribution is derived concerning a dominating measure on the unit sphere, which consequently leads to the density function of the elliptical tail. Subsequently, a maximum likelihood estimation is proposed based on the derived density function class. The resulting maximum likelihood estimator (MLE) is proven to be consistent and asymptotically normal. Moreover, it is demonstrated that the MLE is asymptotically efficient, with the added advantage that its asymptotic covariance matrix can be feasibly estimated at a low computational cost. A simulation study and real data analysis are conducted to illustrate the efficacy of the proposed method.

E0481: **Accelerated failure time model based on nonparametric Gaussian scale mixtures**
Presenter: **Byungtae Seo**, Sungkyunkwan University, Korea, South

When some parametric error distributions are assumed, such as normal for the accelerated failure time model, estimators typically suffer from misspecification problems. To relax this problem, a nonparametric Gaussian scale mixture model is proposed to flexibly specify the error distribution. Unlike existing nonparametric or semiparametric estimation methods such as rank based procedures, the proposed method enables us to use an explicit likelihood function while avoiding potential misspecification problems. This model is presented with specific estimating algorithms and some numerical examples.

E0602: **Block wild bootstrap based Ljung-Box test for VAR model with time-varying variance**
Presenter: **Younjae Lee**, Hankuk University of Foreign Studies, Korea, South
Co-authors: Taewook Lee

The goodness-of-fit test of vector autoregressive (VAR) models is investigated with time-varying variance. While the conventional Ljung-Box (LB) test is assumed to follow the chi-square distribution asymptotically, it has been found that the LB test can deviate significantly from the chi-square distribution for VAR models with time-varying variance. The block wild bootstrap-based (BWB) LB test is proposed, and it has been proved that our BWB-LB test achieves a correct critical region for the goodness-of-fit test of VAR models with time-varying variance. The simulation study shows that the BWB-LB test achieves the correct sizes and comparable powers in finite samples.

E0778: **Shapley values for identifying fault variables in MSPC**
Presenter: **Sungim Lee**, Dankook University, Korea, South
Co-authors: Juwhan Kim, Johan Lim

In multivariate statistical process control, Hotelling's T-squared (HT) control chart is widely used for detecting changes in the mean vector. However, its effectiveness is reduced if it fails to identify fault variables when an out-of-control status is signaled. A novel approach is presented using Shapley values, a powerful technique for explaining predictions in deep learning models. In general, Shapley values are usually estimated using the Shapley sampling method or the KernelSHAP algorithm. These methods are compared with existing methods, such as the Mason-Tracy-Young procedure and adaptive step-down procedure, which are prevalent in fault variable identification for HT control charts. The numerical studies demonstrate that the approach significantly outperforms existing methods in terms of sensitivity and specificity, particularly when changes are significant, but not remarkable. It is found that the approach improves fault variable identification.

E0702: **Robust estimation for bounded bivariate time series models of counts based on density power divergence**
Presenter: **Minyoung Jo**, Seoul National University, Korea, South
Co-authors: Sangyeol Lee

Two types of bounded bivariate time series models are investigated, which are suitable for analyzing time series of counts with values within a finite range. They are a modified version of bounded bivariate integer-valued ARCH models and bounded bivariate INAR models. All these models are constructed based on the bivariate binomial distribution of Type II. The main focus is on providing a robust estimation method for these models. For this, the minimum density power divergence estimator (MDPDE) is employed as a robust estimator. To assess the performance of MDPDE and validate its effectiveness, both Monte Carlo simulations and real data analysis are conducted using monthly earthquake data in the United States. Findings collectively confirm the efficacy and suitability of the methods.

EO151 Room 405 RECENT ADVANCES AND CHALLENGES IN INFERENCE AND LEARNING Chair: Ansgar Steland

E1008: **Likelihood asymptotics for stationary Gaussian arrays**
Presenter: **Fabian Mies**, Delft University of Technology, Netherlands
Co-authors: Carsten Chong

Arrays of stationary Gaussian time series can arise naturally in econometric applications, e.g., the discretization of continuous-time stochastic processes, or be introduced artificially to model persistence via so-called local-to-unity models, i.e., linear time series models with parameters close to a unit root. For the parametric statistical estimation of these stationary models, the spectral density plays a central role. In particular, classical results in time series analysis suggest that the Gaussian likelihood and Fisher information may be approximated in terms of the spectral density, and conditions for the efficiency of the MLE have been formulated in the literature. Unfortunately, these general results do not cover arrays of time series. The contribution is to show in which way the asymptotic likelihood theory needs to be adapted for the array case and to demonstrate that this yields a straightforward approach to studying a broad class of processes. As a motivating example, the mixed fractional Brownian motion estimation is investigated based on high-frequency observations. Findings reveal that the achievable convergence rates depend intricately on the size of the various components and their intertemporal and cross-temporal dependence structure.

E1032: **Weak convergence of the function-indexed sequential empirical process for nonstationary time series**
Presenter: **Florian Scholze**, RWTH Aachen University/ University of Bamberg, Germany

Sequential empirical processes have several statistical applications, including change detection and goodness-of-fit testing. Studying their behavior in different settings is, therefore, of both theoretical and practical interest. So far, the literature on the weak convergence of the function-indexed sequential empirical process under dependence seems to be limited to the stationary case. To partially close this gap, its weak convergence is studied in a nonstationary setting, and it is shown to be asymptotically equicontinuous, provided suitable maximal inequalities are available for the increments of its nonsequential counterpart. The assumed maximal inequalities implicitly contain some nonspecific dependence restrictions, but no additional dependence restrictions are imposed. Thereby, a certain level of generality is achieved, which enables a range of possible applications. Limitations, possible extensions and statistical applications are discussed.

E1046: **An estimator for dynamic linear panel data models based on nonlinear moment conditions**
Presenter: **Joachim Schnurbus**, University of Passau, Germany
Co-authors: Andrew Adrian Yu Pua, Markus Fritsch

An instrumental variables (IV) estimator based on nonlinear (in parameters) moment conditions that may resolve identification problems regarding the lag parameter when estimating linear dynamic panel data models is proposed. The estimator is applicable in the unit root, near-unit root, and non-unit root case. Consistency and asymptotic normality of the estimator are shown when the cross-section dimension is large and the time series dimension is either fixed or large, and the improvement in the rate of convergence compared to existing estimators is discussed. While the estimator point identifies the lag parameter when the lag parameter is one, it yields two distinct solutions otherwise. A selection rule is proposed and analyzed for the latter case, which is supposed to identify the consistent root asymptotically.

E1088: **Sequential conformal prediction for time series**
Presenter: **Chen Xu**, Georgia Institute of Technology, United States
Co-authors: Yao Xie

A new distribution-free conformal prediction algorithm is presented for sequential data (e.g., time series), called the sequential predictive conformal inference (SPCI). The nature that time series data are non-exchangeable is specifically accounted for, and thus many existing conformal prediction algorithms are not applicable. The main idea is to adaptively re-estimate the conditional quantile of non-conformity scores (e.g., prediction residuals) upon exploiting the temporal dependence among them. More precisely, the problem of conformal prediction interval is cast as predicting the quantile of a future residual, given a user-specified point prediction algorithm. Theoretically, asymptotic valid conditional coverage is established upon extending consistency analyses in quantile regression. Using simulation and real-data experiments, a significant reduction is demonstrated in the interval width of SPCI compared to other existing methods under the desired empirical coverage. Extensions to multivariate time series are also discussed.

EO256 Room 408 ECONOMETRICS AND CONTEMPORARY ISSUES IN ECONOMICS AND FINANCE (VIRTUAL) Chair: Namhyun Kim

E0659: **Optimal predictor and transformation selection for macroeconomic forecasting using variable importance in random forests**
Presenter: **Maurizio Daniele**, ETH Zurich, KOF Swiss Economic Institute, Switzerland
Co-authors: Philipp Kronenberg, Tim Reinicke

A novel recursive group variable importance measure in random forests (RF) is proposed to select the most relevant indicators for predicting key macroeconomic variables. In contrast to existing RF-based importance measures, the method enhances the modeling flexibility by accounting for general types of time series structures in economic data. In an out-of-sample forecasting experiment using a large dimensional macroeconomic dataset based on the FRED-MD database, significant improvements are illustrated in forecasting US inflation when employing the RF-based selection approach for extracting the optimal predictors and data transformations compared to existing selection methods relying on conventional-regularization techniques, e.g. the lasso and elastic net. Moreover, the findings reveal that optimal variable transformations uniformly enhance the predictive accuracy of various modeling approaches, including regularization methods, (dynamic) factor models, neural networks, and random forests. The observed forecasting improvements highlight the importance of considering alternative transformations beyond the conventional choices recommended in the FRED-MD dataset. Furthermore, theoretical insights are provided on the RF-based selection criterion in an additive model framework.

E0955: **Superstar firms and aggregate fluctuations**
Presenter: **Oscar Pavlov**, University of Tasmania, Australia

The rise of market power in the last decades is primarily driven by the largest firms. A theory of these superstar firms is proposed, in which their technology involves the ability to produce multiple products. Superstars interact with smaller competitors, and market share reallocations and product creation generate heterogeneous markup dynamics across firms. Higher market shares of superstars increase the parameter space for macroeconomic indeterminacy. Bayesian estimation of the general equilibrium model suggests the importance of the endogenous amplification of the product creation channel and animal spirits play a non-trivial role in driving U.S. business cycles.

E1002: **Forecasting agricultural land-use in England by using spatially highly resolution data**
Presenter: **Namhyun Kim**, University of Exeter, United Kingdom

The aim is to introduce a new analytical framework for conducting an empirical analysis of agricultural land-use shares. The framework includes a nonparametric method of modelling the spatial patterns in the land-use shares, which capture an agglomeration and complementary effect across land grids in the most general way. This enables a more effective analysis of the interplay between land-use and its important determinants, for example, climate and environment policies, than those often used in existing works. The method is a semiparametric generalization of the well-known spatial lag dependence model in spatial econometrics. Furthermore, allowing the data to speak for themselves leads to a more accurate measure of the importance of spatial-physical environment, e.g. soil characteristics, textures, altitude and slope, and therefore helps to re-focus environmental policies.

E0884: **Cross-national comparisons of COVID-19 lockdown effectiveness: The spatial functional data analysis approach**
Presenter: **Pipat Wongsa-art**, Cardiff University, United Kingdom

Although studying the cross-national effectiveness of lockdown strategies in reducing the transmission of COVID-19 is necessary and extremely important, it is far from being straightforward. An endogeneity of policy choices and reverse causality are obvious examples of obstructions that may impede progress. The purpose is to transform the problem into analyzing spatially dependent discrete longitudinal data of COVID-19 cases and deaths, which are often used by governments as the basis for making policy decisions. In the context of the analysis, the spatial dependence is extended beyond the concept of physical contiguity of neighborhoods, which is often the main focus in spatial econometrics, to COVID-19 response contiguity. Furthermore, a novel functional data analysis approach is suggested that can help to disentangle a component of the data series of COVID-19 cases/deaths, which is due to the COVID-19 response contiguity, from another component that is country-specific. The usefulness of the latter resides in its ability to capture information about the effectiveness of government policies. The method is used to perform cross-national comparisons of COVID-19 lockdown effectiveness in 36 OECD countries and provide a number of insights that are not yet available in

the literature.

EO086 Room 411 (Virtual sessions) ANALYTICS IN FINANCE AND INSURANCE Chair: Tak Kuen Siu

E0297: **Risk management for aquaculture businesses in the presence of multiple risk sources**
Presenter: **Christian Oliver Ewald**, Umea Uiversity, Sweden
An overview is provided about current modelling and analysis relevant to the financial, biological and ESG-related risks in the optimal management of aquaculture businesses with a particular focus on the Norwegian salmon farming industry. These results are based on recent studies, as well as research that is currently in development.

E0465: **Modelling spot and option-on-futures prices of the EU carbon allowance**
Presenter: **Rogemar Mamon**, University of Western Ontario, Canada
The behavior of the spot price of carbon emission allowance is investigated in the European Union Emissions Trading Scheme (EU ETS). Motivated by the volatility clustering phenomenon, a regime-switching mechanism is embedded into four stochastic models governed by a hidden Markov chain, which enables time-dependent parametrization. The pricing of European-style futures call options is examined under the proposed modelling setups. The models are assessed by comparing the pricing errors using Bloomberg's call option data on EUA futures. The proposed regime-switching geometric Brownian motion is deemed the best-fitting model among the developed alternatives.

E0626: **Threshold autoregressive nearest-neighbor models for claims reserving**
Presenter: **Tak Kuen Siu**, Macquarie University, Australia
Motivated by claims reserving in non-life insurance, a class of threshold autoregressive nearest-neighbor (TAR-NN) models is discussed. The TAR-NN model extends a major class of parametric nonlinear time series models, namely threshold autoregressive (TAR) models, by introducing a random field structure. It also generalizes nearest-neighbor (NN) models by introducing a regime-switching mechanism. Attention is given to a class of self-exciting threshold autoregressive nearest-neighbor (SETAR-NN) models and their applications to claims reserving. The (strict) stationarity and geometric ergodicity of the SETAR-NN model are discussed. The conditional least-square (CLS) method is used to estimate the SETAR-NN model and some of its nested models. Simulation studies on the parameter estimates are conducted. Applications of the SETAR-NN model and the nested models for projecting future claims liabilities are presented based on insurance claims data.

EC269 Room 406 FINANCIAL ECONOMETRICS Chair: Wai-keung Li

E0988: **A general option pricing framework for affine fractionally integrated models**
Presenter: **Alex Badescu**, University of Calgary, Canada
Co-authors: Maciej Augustyniak, Jean-Francois Begin, Sarath Kumar Jayaraman
The purpose is to study the impact of fractional integration on volatility modelling and option pricing. A general discrete-time pricing framework is proposed based on affine multi-component volatility models. It not only nests a large variety of option pricing models from the literature but also allows for the introduction of novel covariance-stationary long-memory affine GARCH pricing models. Using an infinite sum characterization of the log-asset prices cumulant generating function, semi-explicit expressions are derived for the valuation of European-style derivatives under a general variance-dependent stochastic discount factor. Moreover, an extensive empirical analysis is carried out using returns and S&P 500 options from 1996-2019. Overall, once the informational content from options is incorporated into the parameter estimation process, including fractionally integrated dynamics in volatility, the out-of-sample option pricing performance is improved. The largest improvements in the implied volatility root-mean-square errors occur for options with maturities longer than one year, reaching 33% and 13% when compared to standard one- and two-component short-memory models, respectively.

E1009: **Extreme movements and volatility regimes: A latent factor regime switching perspective**
Presenter: **Yuyi Li**, University of Liverpool, United Kingdom
Co-authors: Ruijun Bu, Jie Cheng, Abdoulkarim Idi cheffou, Fredj Jawadi
The purpose is to empirically investigate asymmetric volatility regime switching in the context of extreme price movements, employing a novel latent-factor-driven endogenous regime-switching framework. The relationship is modeled between past shocks in observed returns and those in a latent factor using a copula function. This approach enables empirical capturing of more complex market features such as asymmetry, nonlinearity, and tail dependence in our endogeneity function. The empirical analysis extends to non-Gaussian state-dependent processes and state-dependent endogeneity, encompassing and expanding upon several existing endogenous regime-switching models. Through empirical study, the model is demonstrated to predict transitions from low to high volatility regimes more accurately than the symmetric model proposed in a past study. Findings suggest that symmetric models may significantly underestimate the likelihood of shifting to higher-risk regimes, particularly after substantial negative market shocks.

E0824: **Financial time series analysis with weighted quantile approach**
Presenter: **Tomas Tichy**, VSB-TU Ostrava, Czech Republic
Co-authors: Michal Holcapek, David Nedela
A rather complex sub-task of the research is to design inference mechanisms in a fuzzy-probability environment that will be well justified and interpreted because probability and fuzziness are different measures, and care must be taken when combining them. Inference mechanisms proposed in previous works are analyzed and integrated into the new probabilistic fuzzy models. They are also analyzed and extended to other inference mechanisms designed for fuzzy systems in fuzzy-probabilistic settings. In particular, in this contribution, weighted quantiles and the introduction of a very simple and fast method for their determination are studied. This method is implemented in an algorithm for probabilistic-fuzzy inference systems that allows the deriving of weighted quantiles for any element or interval in a given input domain for a fixed probability, for example, a median or median function, if the probability is equal to 0.5. In addition, a quantile function can be determined for any fixed element in a given domain. Subsequently, several examples of time series are investigated, and statistical properties of proposed probabilistic-fuzzy models are discussed and compared to recently popular models in financial time series analysis and forecasting.

E0708: **Testing the tail efficiency hypothesis: Extreme-value perspective on market efficiency**
Presenter: **Junshu Jiang**, King Abdullah University of Science and Technology, Saudi Arabia
Co-authors: Raphael Huser, Jordan Richards
In economics, the market efficiency hypothesis posits that asset prices reflect all available information. Several empirical investigations show that market efficiency under extreme situations is relatively low. Although many models for extremal dependence have been developed over the last few decades, these mainly focus on characterizing the behaviour of a random vector when it exhibits only positive extremal dependence, i.e., all components are jointly extreme in the same direction. This limitation makes them unsuitable for studying dependence in financial markets where it is not uncommon for data to exhibit both positive and negative extremal dependence. To jointly model positive and negative extremal dependence,

regular variation models are constructed on the entirety of Rd space and develop a bivariate measure for the asymmetry between the strength of extremal dependence in adjacent orthants. The so-called directional tail dependence (DTD) measure allows us to construct the tail efficiency hypothesis, an analogue of the market efficiency hypothesis, for behavior of the market in its tails. Asymptotic results for estimators of the DTD are described, and testing is discussed via permutation-based methods. Empirical results for China's derivatives market support that the market is generally tail-efficient, with a few exceptions which indicate potential arbitrage opportunities.

Authors Index

