PROGRAMME AND ABSTRACTS

5th International Conference on

Econometrics and Statistics (EcoSta 2022)

http://cmstatistics.org/EcoSta2022

Ryukoku University, Kyoto, Japan

4 – 6 June 2022

Econometrics and Statistics

COMPUTATIONAL STATISTICS & DATA ANALYSIS

RYUKOKU UNIVERSITY

CMStatistics
Computational and Methodological Statistics
http://www.CMStatistics.org
http://www.CFEnetwork.org
Computational and Financial Econometrics
CFEnetwork

ISBN: 978-9925-7812-1-8

Co-chairs:

Monica Billio, Masayuki Hirukawa, Xinyuan Song and Toshiaki Watanabe.

EcoSta Editors:

Ana Colubi, Erricos J. Kontoghiorghes and Manfred Deistler.

Scientific Programme Committee:

Manabu Asai, Eric Beutner, Jan Pablo Burgard, Stefano Castruccio, Cathy W.S. Chen, Andreas Christmann, Bertrand Clarke, Abdelaati Daouia, Fabrizio Durante, Takeshi Emura, James Flegal, Subir Ghosh, Michele Guindani, Alain Hecq, Scott Holan, Yen-Tsung Huang, Ci-Ren Jiang, Sungkyu Jung, MingHung Kao, Yata Kazuyoshi, Abbas Khalili, Donggyu Kim, Yuta Koike, Pavel Krupskiy, Degui Li, Wai-Keung Li, Yi Li, Yuanyuan Lin, Shujie Ma, Koichi Maekawa, Qing Mai, Tapabrata Maiti, Matthieu Marbac, Cristina Mollica, Ryo Okui, Hernando Ombao, Taesung Park, Igor Pruenster, Anuradha Roy, Etsuro Shioji, Mike K.P. So, Peter Song, John Stufken, Jianguo Sun, Anneleen Verhasselt, Linbo Wang, Yuedong Wang, Raymond Wong, Yingcun Xia, Philip Yu, Emma Jingfei Zhang, Yuan Zhang, Yi Zhao and Lixing Zhu.

Local Organizing Committee:

Ryukoku University, EcoSta, CMStatistics and CFEnetwork.

Dear Colleagues,

It is a great pleasure to welcome you to the 5th International Conference on Econometrics and Statistics (EcoSta 2022). These years we are passing through extraordinary events that significantly affect our personal and professional lives. In order to cope with the uncertainty caused by the pandemic, we have implemented a hybrid format, so that the participants can select to attend in person or virtually according to their circumstances and the local restrictions. The programme has been dynamically adapted to allow the delegates to present their results and network in the best possible way. Despite the challenges, we are happy to host the biggest meeting of the conference series in terms of the number of participants and presentations, with about 850 attenders. We appreciate the efforts of all those involved in the conference, especially the session organizers, who have worked under the uncertainty of being able to hold their sessions in person or virtually, but have succeeded in putting together an excellent programme, though.

The conference is co-organized by the working group on Computational and Methodological Statistics (CMStatistics), the network of Computational and Financial Econometrics (CFEnetwork), the journal Econometrics and Statistics (EcoSta) and Ryukoku University. Following the success of the last editions, the aim is for the conference to become a leading meeting in econometrics, statistics and their applications.

The EcoSta 2022 consists of over 200 sessions, three keynote talks, four invited sessions, and about 770 presentations. These numbers confirm the support of the involved research communities for this important initiative. It is indeed promising that the EcoSta conference will become a successful medium for disseminating high-quality research in Econometrics and Statistics and facilitating networking.

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

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

The Elsevier journals of Econometrics and Statistics (EcoSta) and Computational Statistics & Data Analysis (CSDA) are associated with CFEnetwork, CMStatistics, and the EcoSta 2022 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.

Ana Colubi, Masayuki Hirukawa 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 (Japan time, GMT+9)

2022-06-04	**2022-06-05**	**2022-06-06**
Opening, 08:10 - 08:25	**F** EcoSta2022 08:00 - 10:05	**K** EcoSta2022 07:40 - 09:20
A EcoSta2022 08:25 - 09:40	**Coffee Break** 10:05 - 10:35	**L** EcoSta2022 09:30 - 11:10
Coffee Break 09:40 - 10:10	**G** EcoSta2022 10:35 - 12:15	**Coffee Break** 11:10 - 11:40
B EcoSta2022 10:10 - 11:50	**Lunch Break** 12:15 - 13:15	**M - Keynote** EcoSta2022 11:40 - 12:30
Lunch Break 11:50 - 13:00	**H** EcoSta2022 13:15 - 14:55	**Lunch Break** 12:30 - 13:30
C EcoSta2022 13:00 - 14:40	**Coffee Break** 14:55 - 15:25	**N** EcoSta2022 13:30 - 14:45
D - Keynote EcoSta2022 14:50 - 15:40	**I** EcoSta2022 15:25 - 16:40	**Coffee Break** 14:45 - 15:15
Coffee Break 15:40 - 16:10	**J** EcoSta2022 16:50 - 18:30	**O** EcoSta2022 15:15 - 16:55
E EcoSta2022 16:10 - 18:15		**P - Keynote** EcoSta2022 17:05 - 17:55
		Closing, 17:55 - 18:05

VIRTUAL TUTORIAL, MEETINGS, SOCIAL EVENTS AND ACCESS TO THE CONFERENCE

VIRTUAL TUTORIAL

The tutorial “Stochastic volatility and realized stochastic volatility models” will take place on Friday, 3rd of June 2022, 15:00-19:30 (GMT+9). It will be delivered by Prof. Yasuhiro Omori. Only participants who had subscribed for the tutorial can attend. Registered participants will be able to access the virtual tutorial through the website.

SPECIAL MEETINGS by invitation to group members

The EcoSta (Econometrics and Statistics) and CSDA (Computational Statistics and Data Analysis) Editorial Board meetings will take place on Friday, 3rd of June 2022, 15:45-16:30 (GMT+9). Indications to attend the Editorial Board meetings will be sent to the AEs participating in the conference in due course.

ACCESS TO THE CONFERENCE

- Participants can attend virtually or in person, provided that the conditions imposed by the host are satisfied. The in-person access to Ryukoku University for conference participants is restricted to those who had confirmed their in-person participation in the doodle sent by email.
- The in-person venue is the Fukakusa campus of Ryukoku University, 67 Tsukamoto-cho, Fukakusa, Fushimi-ku, Kyoto 612-8577, Japan.
- The **registration** will be located in the entrance of the ground floor of Building 22 during the weekend and in Room J301 of Jojukan on Monday (see map on page VIII and floor maps on page IX).
- The **coffee breaks and lunches** will take place in the café on the basement floor of Building 4 during the weekend and in Rooms J301-J306 and J401-403 of Jojukan on Monday (see map on page VIII and floor maps on page IX).
- The conference is live streaming, and it will not be recorded. The oral presentations will take place through Zoom, while the virtual social events and poster presentations will run in Gather Town. The conference programme time is set at GMT+9.
- 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 your wish to attend and then on the session. If it is hybrid, click on the blinking room of the floor map. 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).
- Detailed indications for virtual and in-person attendance, hybrid sessions, speakers, chairs, posters, networking, test sessions, as well as FAQ, can be found on the webpage.

Presentation instructions

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

Posters

The poster sessions will take place through Gather Town. The posters should be sent in **png format** to info@CMStatistics.org by the 2nd of June 2022. Landscape orientation is advisable. Detailed indications for the poster presentations can be found on the website.

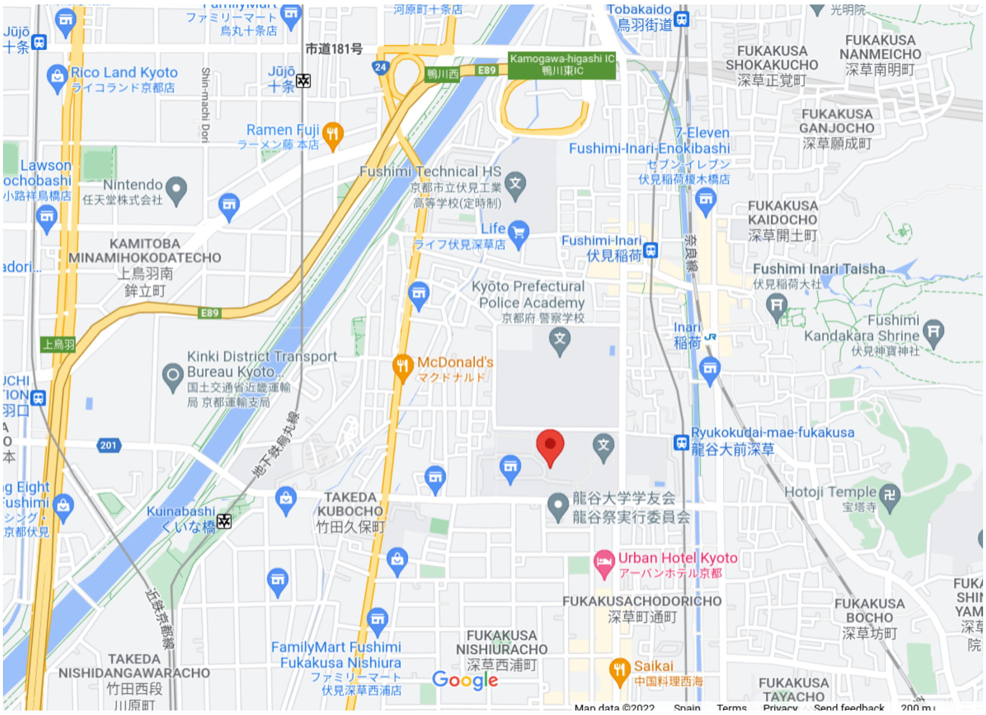
Session chairs

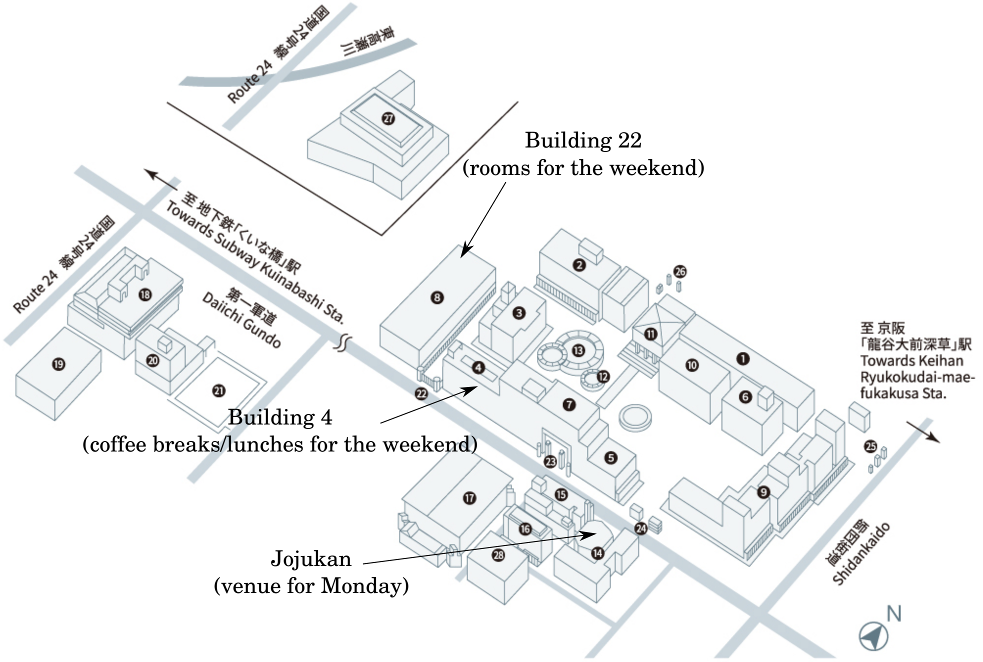
The session chairs will be responsible for introducing the session, the speakers and coordinating the discussion time. A member of the conference staff, identified by the name *Angel* followed by the room number, will assist in giving the rights to participate as the chair requests it. If any speaker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

Test session

A test session will be set up for Sunday, the 29th of May 2022, from 15:00 to 15:30 GMT+9 (Japan time). The participants will be able to enter through the Virtual Room R1 in the programme to test their presentations, (e.g., through Parallel session A) to test their presentations, video, micro and audio. Detailed indications for the test session can be found on the website.

Map of the venue and nearby area


Jūjō 十条
Rico Land Kyoto ライコランド京都店
市道181号
Jūjō 十条
Kamogawa-higashi IC 鴨川東IC
Tobakaido 鳥羽街道
FUKAKUSA SHOKAKUCHO 深草正覚町
FUKAKUSA NANMEICHO 深草南明町
Ramen Fuji ラーメン藤 本店
Lawson Kochobashi 小路祥鳥橋店
Nintendo 任天堂株式会社
Fushimi Technical HS 京都市立伏見工業高等学校(定時制)
Fushimi-Inari-Enokibashi
7-Eleven セブン-イレブン 伏見稲荷榎木橋店
FUKAKUSA GANJOCHO 深草願成町
FUKAKUSA KAIDOCHO 深草開土町
Life ライフ伏見深草店
Fushimi-Inari 伏見稲荷
KAMITOBA MINAMIHOKODATECHO 上鳥羽南鉾立町
Kyoto Prefectural Police Academy 京都府 警察学校
Fushimi Inari Taisha 伏見稲荷大社
Inari 稲荷
Fushimi Kandakara Shrine 伏見神寶神社
Kinki District Transport Bureau Kyoto... 国土交通省近畿運輸局 京都運輸支局
McDonald's マクドナルド
Ryukokudai-mae-fukakusa 龍谷大前深草
Kuinabashi くいな橋
TAKEDA KUBOCHO 竹田久保町
龍谷大学学友会 龍谷祭実行委員会
Hotoji Temple 宝塔寺
Urban Hotel Kyoto アーバンホテル京都
FUKAKUSACHODORICHO 深草町通町
FUKAKUSA BOCHO 深草坊町
TAKEDA NISHIDANGAWARACHO 竹田西段川原町
FamilyMart Fushimi Fukakusa Nishiura ファミリーマート 伏見深草西浦店
FUKAKUSA NISHIURACHO 深草西浦町
Saikai 中国料理西海
FUKAKUSA TAYACHO
Google
Map data ©2022



❶和顔館／Wagenkan
❷2号館／Building 2
❸3号館／Building 3
❹4号館／Building 4
❺5号館／Building 5
❻8号館／Building 8
❼21号館／Building 21
❽22号館／Building 22
❾紫英館／Shieikan
❿図書館／Library

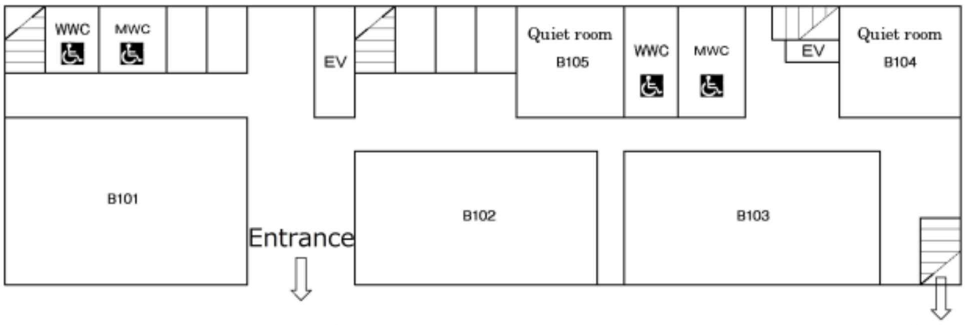
⓫顕真館／Kenshinkan
⓬カフェ樹林／Cafe Jurin
⓭ステージ／Stage
⓮成就館／Jojukan
⓯紫朋館／Shihokan
⓰紫陽館／Shiyokan
⓱体育館／Gymnasium
⓲紫光館／Shikokan
⓳紫光館別館／Shikokan Annex
⓴至心館／Shishinkan

㉑テニスコート／Tennis Court
㉒西門／West Gate
㉓正門／Main Gate
㉔通用門／Side Gate
㉕東門／East Gate
㉖北門／North Gate
㉗専精館／Senshokan
㉘ミトラ館／mitrakan

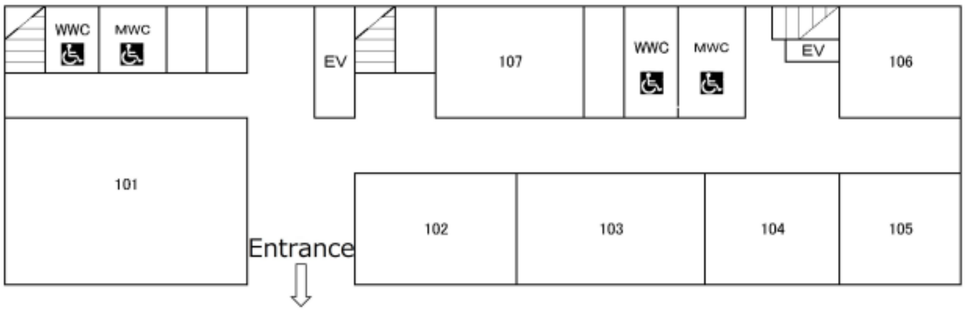
〒612-8577 京都市伏見区深草塚本町67
67 Tsukamoto-cho, Fukakusa, Fushimi-ku, Kyoto 612-8577
Tel. 075-642-1111

Floor maps

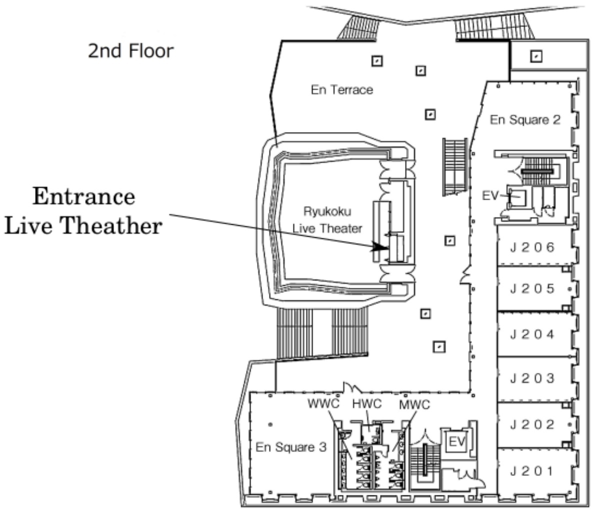
1st Basement Floor - Weekend quiet rooms (B104 and B105)

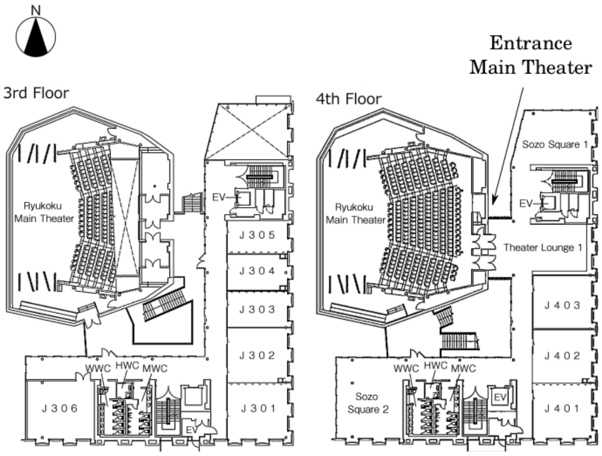


1st Floor (Ground floor) - Weekend hybrid sessions and registration



Building 22





Registration - J301, Quiet Rooms - J401, J402
Coffee breaks - J301-J306 and J401-403

Jojukan

PUBLICATION OUTLETS

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

Econometrics and Statistics (EcoSta)

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

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

- **Part A: Econometrics.** Emphasis will be given to methodological and theoretical papers containing substantial econometrics derivations or showing a potential of a significant impact in the broad area of econometrics. Topics of interest include the estimation of econometric models and associated inference, model selection, panel data, measurement error, Bayesian methods, and time series analyses. Simulations are to be considered when they involve an original methodology. Innovative papers in financial econometrics and its applications will be considered. The topics to be covered include portfolio allocation, option pricing, quantitative risk management, systemic risk and market microstructure. Interest will be focused as well on well-founded applied econometric studies that demonstrate the practicality of new procedures and models. Such studies should involve the rigorous application of statistical techniques, including estimation, inference and forecasting. Topics will include volatility and risk, credit risk, pricing models, portfolio management, and emerging markets. Innovative contributions in empirical finance and financial data analysis that use advanced statistical methods are encouraged. The results of the submissions should be replicable. Applications consisting only of routine calculations will not be of interest to the journal.

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

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

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

Saturday 04.06.2022	14:50 - 15:40	Room: 101 (Hybrid 1)	Chair: Erricos Kontoghiorghes	Keynote talk 1

Bayesian estimation of multivariate stochastic volatility models using dynamic factors
Speaker: **Yasuhiro Omori, University of Tokyo, Japan**

In the stochastic volatility models for multivariate daily stock returns, it has been found that the estimates of parameters become unstable as the dimension of returns increases. We first describe various multivariate stochastic volatility models and discuss how we can overcome the difficulties in the estimation of model parameters and latent variables using Markov chain Monte Carlo simulation. Then, we focus on the model based on the factor structure of multiple returns with two additional sources of information: first, the realized stock index associated with the market factor, and second, the realized covariance matrix calculated from high-frequency data. How to remove biases of realized volatilities and realized correlation matrices are also illustrated. The proposed dynamic factor model with the leverage effect and realized measures is applied to ten of the top stocks composing the exchange-traded fund linked with the investment return of the S&P500 index and the model is shown to have a stable advantage in portfolio performance. The relative weight of the measurement equation for the realized covariances is found to be larger than that for the daily returns. The estimates of leverage effect and the correlation coefficients among asset returns are found to be higher without realized covariances, which is considered to be the bias due to insufficient information.

Monday 06.06.2022	11:40 - 12:30	Room: Main Theater (Hybrid 1)	Chair: Masayuki Hirukawa	Keynote talk 2

An empirical evaluation of some long-horizon macroeconomic forecasts (virtual)
Speaker: **Kenneth D West, University of Wisconsin, United States** Kurt Lunsford

Over a century of annual cross-country data are used to evaluate pseudo-out-of-sample forecasts of six mean-reverting variables and four quite persistent variables for horizons up to 50 years. The mean-reverting variables are real per capita GDP growth, CPI inflation, labor productivity growth, population growth, money growth and equity returns; the quite persistent variables are real exchange rates, the investment-to-output ratio, and long- and short-term nominal interest rates. Our models for forecasting include simple time series models and frequency domain methods recently developed. We consider both point estimation and coverage of 68% intervals for forecasts. For the six mean-reverting variables, a simple AR(1) and a frequency domain model are best in terms of point estimation and, as well, have well-calibrated 68% forecast intervals; calibration of forecast intervals, does, however, distinctly degrade when increasing the horizon from 25 to 50 years. For the four very persistent variables, a random walk is perhaps the best choice, though forecast intervals are not well calibrated at any horizon for any of the models. We conclude that forecasting over very long horizons is viable, at least for data that are not very persistent.

Monday 06.06.2022	17:05 - 17:55	Room: Main Theater (Hybrid 1)	Chair: Ana Colubi	Keynote talk 3

Multiple change-point detection for functional data (virtual)
Speaker: **Jeng-Min Chiou, Academia Sinica, Taiwan**

Detecting abrupt structural changes in a data sequence is interesting in many applications. Multiple change-point detection of a functional data sequence is discussed through two recent approaches: dynamic segmentation and greedy segmentation. They comprise the detection and the significance testing procedures under the least-squares segmentation criterion intending to identify the locations and the number of change points. We discuss the asymptotic properties of the methods and explore their finite sample performance through simulation studies and data applications.

Saturday 04.06.2022 08:25 - 09:40 Parallel Session A – EcoSta2022

EV452 Room Virtual R12 CONTRIBUTIONS IN METHODOLOGICAL STATISTICS AND ECONOMETRICS Chair: Ci-Ren Jiang

E0527: **The stochastic frontier model with ordered multiple choice**
Presenter: **Yi-Wun Chen**, Binghamton University, State University of New York, United States
The problem of biased and inconsistent estimates due to the sample selection bias has been frequently considered. If we encompass all observations with different regression coefficients (heterogeneous observations) into one regression equation, statistically it also implies the sample selection bias uses the stochastic frontier (SF) model with endogenous switching to discuss the issue of heterogeneity among firms. The SF model with two regimes/selections is extended to multiple regimes/selections to broaden the empirical application. We derive the closed-form of the likelihood function for the proposed model and the estimator of the technical efficiency (TE) index and then estimate it by maximum likelihood estimation. In the empirical study, we study the operating cost efficiency of doctoral-granting universities in the United States and applied it to the proposed model with three regimes.

E0259: **Actual events vs. actual reporting: Modeling firm performance under environmental uncertainty using machine learning**
Presenter: **Minh Nguyen**, University of Hawaii at Manoa, United States
Not all companies respond the same to natural disaster events. Two ways that natural disasters affect firm performance are studied: actual events vs. actual reporting. We consider the billion-dollar natural disasters in the United States as the actual events and the number of words related to natural disasters in the Management Discussion and Analysis section in Form 10-Ks filing by the US public companies as the actual reporting. The aim is also to compare the performances of classification and regression trees (CART) and neural networks with the benchmark model-based linear regression model in predicting the performance of U.S. public companies under environmental uncertainty. We find that both actual events and actual reporting of natural disasters in year t negatively affect the return on assets (ROA) in year $t+1$. Also, the actual natural disasters in year t negatively affect sales growth in year $t+1$. Moreover, we find that the environmental uncertainty variables are much less important than the traditional financial statement variables in predicting firm performance with the CART model. Comparing CART, neural networks, and linear regression models, we find that CART and neural networks outperform linear regression models in predicting firm performance.

E0699: **Partition-Mallows model and its inference for rank aggregation**
Presenter: **Wanchuang Zhu**, The University of Sydney, Australia
Learning how to aggregate ranking lists has been an active research area for many years and its advances have played a vital role in many applications ranging from bioinformatics to internet commerce. The problem of discerning the reliability of rankers based only on the rank data is of great interest to many practitioners but has received less attention from researchers. By dividing the ranked entities into two disjoint groups, i.e., relevant and irrelevant/background ones, and incorporating the Mallows model for the relative ranking of relevant entities, we propose a framework for rank aggregation that can not only distinguish quality differences among the rankers but also provide the detailed ranking information for relevant entities. Theoretical properties of the proposed approach are established, and its advantages over existing approaches are demonstrated via simulation studies and real-data applications. Extensions of the proposed method to handle partial ranking lists and conduct covariate-assisted rank aggregation are also discussed.

EV465 Room Virtual R7 CONTRIBUTIONS IN TIME SERIES I Chair: Donggyu Kim

E0431: **Spatial aggregation on high dimensional multivariate time series analysis**
Presenter: **William WS Wei**, Temple University, United States
The vector autoregressive (VAR) and vector autoregressive moving average (VARMA) models have been widely used to model multivariate time series, because of their ability to represent the dynamic relationships among variables in a system and their usefulness in forecasting unknown future values. However, when the number of dimensions is very large, the number of parameters often exceeds the number of available observations, and it is impossible to estimate the parameters. A suitable solution is clearly needed. After introducing some existing methods, we will suggest the use of spatial aggregation as a dimension reduction method, which is very natural and simple to use. We will compare our proposed method with other existing methods in terms of forecast accuracy through both simulations and empirical examples.

E0567: **A framework for time series aggregation and seasonality using marked point processes**
Presenter: **Tucker McElroy**, Census Bureau, United States
Co-authors: Anindya Roy
Stochastic processes with a meager sampling rate are studied, wherein the number of observations per epoch or region is small. The goal is to construct a model whereby temporal or vertical aggregation of such meager processes results in a data stream that more closely resembles a Gaussian process, such as is commonly used to model higher-aggregate economic data. The approach involves Marked Point Processes, where each point corresponds to a single transaction in the marketplace, and a given measure that governs the distribution of points determines the mean and covariance structure of the observed process. We discuss applications to sampling, warping through a change of measure, and modeling of data with trend and seasonality.

E1024: **Metallgesellschaft's hedging revisited: A superior predictive ability test analysis**
Presenter: **Janette Goodridge**, Utah State University, United States
Co-authors: Tyler Brough
In the fall of 1993, German industrial conglomerate Metallgesellschaft AG experienced a derivatives related loss of over one billion dollars. This was one of the largest, if not the largest, losses of its kind at this time. Naturally, this raised interest regarding what caused these huge losses and called into question the events and circumstances surrounding and leading up to this disaster. There are two basic schools of thought regarding Metallgesellschafts strategy. The first is that it was a good strategy and had the potential to be successful and profitable. The other is that its strategy was risky and speculative. Despite the abundance of literature surrounding this event, there has been so resolution as to which viewpoint is correct. The aim is to remedy this via time-series data simulation, the stationary bootstrap, and Hansens Superior Predictive Ability (SPA) test. SPA tests can be used to determine if one specific method or procedure is outperformed by a benchmark method. Applied to Metallgesellschafts situation, SPA can answer once and for all if Metallgesellschaft unnecessarily took on too much risk, or had a good strategy.

EO165 Room 101 (Hybrid 1) HIGH-DIMENSIONAL INFERENCE AND REPRODUCIBLE LEARNING Chair: Daoji Li

E0975: **High-dimensional robust inference via the debiased rank lasso**
Presenter: **Yoshimasa Uematsu**, Hitotsubashi University, Japan
Co-authors: Kazuma Sawaya

An inferential framework robust to heavy-tailed error distributions in high-dimensional linear regression models is proposed. A key ingredient of the robustness is the rank lasso, but the estimator is not asymptotically normally distributed thanks to the regularization. We propose the debiased rank lasso estimator, which can establish the asymptotic normality. Furthermore, using this estimator, we develop a method for the robust simultaneous inference that can discover important variables in the linear regression models with the false discovery rate (FDR) controlled under the preassigned level. We also confirm the performance through extensive numerical simulations. We find that our procedure controls the FDR and exhibits higher power than the original method when the error distribution is heavy-tailed.

E1012: **Reproducible learning for censored data via deep knockoffs**
Presenter: **Daoji Li**, California State University Fullerton, United States

A new feature selection procedure with guaranteed false discovery rate (FDR) control for censored data is introduced. By using deep knockoffs, the proposed procedure can handle covariates with arbitrary and unspecified data distributions. It also can deal with both continuous and categorical covariates. We provide theoretical justifications by showing that the FDR is controlled at the target level. Extensive numerical experiments confirm the generality, effectiveness, and power of our method.

E0776: **Asymptotic behaviors of hierarchical clustering under high dimensional settings**
Presenter: **Kento Egashira**, University of Tsukuba, Japan
Co-authors: Kazuyoshi Yata, Makoto Aoshima

Hierarchical clustering has been approved as a useful tool for the analysis of gene expression microarray data on behalf of high-dimensional, low-sample-size (HDLSS) data. While three asymptotic behaviors of hierarchical clustering are deliberated under asymptotic settings from moderate dimension through HDLSS, it is considered that the conditions required are strict for HDLSS data due to having discussions under several asymptotic settings at once. Given this background, this presentation focuses on HDLSS settings and we prove the asymptotic properties of hierarchical clustering under mild and practical settings for HDLSS data. We proceed with the current comprehension of hierarchical clustering under HDLSS settings without assuming normality. Finally, numerical simulation studies are given and we discuss the performance of the hierarchical clustering for high dimensional data.

EO351 Room 102 (Hybrid 2) CHANGE-POINT DETECTION AND VARIABLE SELECTION FOR LARGE-SCALE DATA Chair: Tao Zou

E0251: **High-dimensional change-point detection using generalized homogeneity metrics**
Presenter: **Shubhadeep Chakraborty**, University of Washington, Seattle, USA, United States
Co-authors: Xianyang Zhang

Change-point detection has been a classical problem in statistics and econometrics. The focus is on the problem of detecting abrupt distributional changes in the data-generating distribution of a sequence of high-dimensional observations, beyond the first two moments. This has remained a substantially less explored problem in the existing literature, especially in the high-dimensional context, compared to detecting changes in the mean or the covariance structure. We develop a nonparametric methodology to (i) detect an unknown number of change-points in an independent sequence of high-dimensional observations and (ii) test for the significance of the estimated change-point locations. The approach essentially rests upon nonparametric tests for the homogeneity of two high-dimensional distributions. We construct a single change-point location estimator via defining a cumulative sum process in an embedded Hilbert space. As the key theoretical innovation, we rigorously derive its limiting distribution under the high dimension medium sample size (HDMSS) framework. Subsequently we combine our statistic with the idea of wild binary segmentation to recursively estimate and test for multiple change-point locations. The superior performance of our methodology compared to other existing procedures is illustrated via extensive simulation studies as well as overstock prices data observed during the period of the Great Recession in the United States.

E0772: **Subbagging variable selection for massive data**
Presenter: **Xian Li**, The Australian National University, Australia
Co-authors: Tao Zou, Xuan Liang

Massive datasets usually possess the features of large N (the number of observations) and large p (the number of variables). We propose a subbagging variable selection approach to select relevant variables from massive datasets. Subbagging (subsample aggregating) is an aggregation approach originally from the machine learning literature, which is well suited to the recent trends of massive data analysis and parallel computing. Specifically, we propose a subbagging loss function based on a collection of subsample estimators, which uses a quadratic form to approximate the full sample loss function. The shrinkage estimation and variable selection can be further conducted based on this subbagging loss function. We then theoretically establish the root N-consistency and selection consistency for this approach. It is also proved that the resulting estimator possesses the oracle property. However, variance inflation is found in its asymptotic variance compared to the full sample estimator. A modified BIC-type criterion is further developed specifically to tune the hyperparameter in this method. An extensive numerical study is presented to illustrate the finite sample performance and computational efficiency.

E0960: **High-dimensional cluster/anomaly detection using scan statistics**
Presenter: **Tung-Lung Wu**, Mississippi State University, United States

Scan statistics have been proved to be powerful in detecting local clusters/anomalies. Scan statistics have been successfully applied in various models, including Bernoulli, Poisson, Normal, Exponential and other models. However, little work has been done for more than 3-dimensional data due to computational complexity. We will study the properties of scan statistics in detecting high-dimensional clusters/anomalies with the aid of random projection, a method for dimension reduction.

EO357 Room 103 (Hybrid 3) ANALYTICAL TOOLS FOR BIOMEDICAL DATA WITH COMPLEX STRUCTURES Chair: Shuo Chen

E0238: **High-dimension to high-dimension screening for detecting genome-wide epigenetic regulators of gene expression**
Presenter: **Tianzhou Ma**, University of Maryland, United States

The advancement of high-throughput technology characterizes a wide range of epigenetic modifications across the genome that regulate gene expression. The high dimensionality of both epigenetic and gene expression data make it challenging to identify the epigenetic regulators of genes over the whole genome. Conducting a univariate test for each epigenetic-gene pair is subject to serious multiple comparison burden, and direct application of regularization methods to select important epigenetic-gene pairs is computationally infeasible for both high-dimensional predictors and responses. Applying fast screening to reduce the dimension first before regularization is more efficient and stable than applying regularization methods alone. We propose a high-dimension to high-dimension screening method based on robust partial correlation, namely rPCor, in a multivariate regression model for detecting epigenetic regulators of gene expression over the whole genome. Compared to existing screening methods, our method can reduce the dimension of both predictor and response, and screen at both node (epigenetic features or genes) and edge (epigenetic-gene pairs) levels. We develop data-driven procedures to determine the conditional sets and the optimal screening threshold and implement a fast iterative algorithm. Simulations and two real data applications in cancer studies illustrate the validity and advantage of our method.

E0444: **Analytical tools for whole-brain networks: Fusing statistics and network science to understand brain function**
Presenter: **Sean Simpson**, Wake Forest School of Medicine, United States
Co-authors: Mohsen Bahrami, Chal Tomlinson, Paul Laurienti

Brain network analyses have exploded in recent years, and hold great potential in helping us understand normal and abnormal brain function. Network science approaches have facilitated these analyses and our understanding of how the brain is structurally and functionally organized. However, the development of statistical methods that allow relating this organization to health outcomes has lagged behind. We have attempted to address this need by developing analytical tools that allow relating system-level properties of brain networks to outcomes of interest. These tools serve as synergistic fusions of statistical approaches with network science methods, providing needed analytic foundations for whole-brain network data. We delineate two recent approaches–a mixed-modeling framework for dynamic network analysis and a regression framework for relating distances between brain network features to covariates of interest–that expand the suite of analytical tools for whole-brain networks and aid in providing complementary insight into brain function.

E0674: **The mediating role of neuroimaging data in age-related cognitive decline**
Presenter: **Shuo Chen**, University of Maryland, School of Medicine, United States

Understanding age-related cognitive decline has been central to aging neuroscience. Aging affects brain functions and structures and consequently causes decayed neurocognitive performance. To further understand this process, we investigate the mediation role of multivariate neuroimaging variables in age-related cognitive decline. Considering the fact that cognition is exclusively determined by the brain, we propose a new multivariate mediation model by maximizing the mediation effect of brain imaging data. Specifically, we decompose the total effect of aging on cognitive function into natural direct and indirect effects and maximize the indirect effect with a parsimonious set of neuroimaging variables. We implement the optimization problem by alternating the direction method of multipliers and considering the aggregate effect of selected imaging variables as a functional brain age score (FBAS). The simulation results show that our method can accurately select imaging variables and estimate the mediation effect in comparison to existing methods. We further apply the proposed method to the whole-brain cortical thickness and white-mater integrity data of 37,441 UK Biobank participants and found that the mediation effect of brain imaging variables explains more than 90% of the age effect on cognitive decline.

EO413 Room 104 (Hybrid 4) HIGH-DIMENSIONAL ASSOCIATIONS: APPLICATIONS IN SPATIAL STATISTICS Chair: Yumou Qiu

E0980: **Inference for nonparanormal partial correlation via rank-based nodewise regression with applications to spatial data**
Presenter: **Yumou Qiu**, Iowa State University, United States

A statistical inference procedure is proposed for partial correlations under the high-dimensional nonparanormal (NPN) model where the observed data are normally distributed after certain monotone transformations. The nonparanormal partial correlation is the partial correlation of the normal transformed data under the NPN model, which is a more general measure of conditional dependence. We estimate the NPN partial correlations by regularized nodewise regression based on the empirical ranks of the original data. A multiple testing procedure is proposed to identify the nonzero NPN partial correlations. The proposed method can be carried out by a simple coordinate descent algorithm for lasso optimization. It is easy to implement and computationally more efficient compared to the existing methods for estimating NPN graphical models. Theoretical results are developed to show the asymptotic normality of the proposed estimator and to justify the proposed multiple testing procedure. An application of the proposed method to identify spatial dependence structures in data is discussed. Numerical simulations and a case study on brain imaging data demonstrate the utility of the proposed procedure and evaluate its performance compared to the existing methods.

E0994: **An improved doubly robust estimator using partially recovered unmeasured spatial confounder**
Presenter: **Yuzhen Zhou**, University of Nebraska Lincoln, United States
Co-authors: Sayli Pokal, Yawen Guan, Honglang Wang, Yuzhen Zhou

Studies in environmental and epidemiological sciences are often spatially varying and observational in nature with the aim of establishing cause and effect relationships. One of the major challenges with such studies is the presence of unmeasured confounders. Spatial confounding is the phenomenon in which the spatial residuals are correlated to the spatial covariates in the model. While there is extensive literature studying the effect of spatial confounding bias in the case of continuous covariates, not much work has been done in scenarios where the covariate of interest is binary. A novel method is developed which adjusts for the spatial confounding bias under a spatial-causal inference framework when the covariate of interest is binary. By combining tools from spatial statistics and causal inference literature we propose a method that reduces the bias due to spatial confounding. Through simulation studies, we demonstrate that the proposed improved doubly robust estimator outperforms the existing methods and has the lowest bias and close to nominal coverage in most scenarios. Finally, we implement our method to estimate the effect of installing SCR/SNCR NOx emission control technologies on ambient ozone concentrations.

E1017: **Testing and signal identification for two-sample high-dimensional covariances via multi-level thresholding**
Presenter: **Bin Guo**, Southwestern University of Finance and Economics, China

Testing and signal identification is considered for covariance matrices from two populations. A multi-level thresholding procedure is proposed for testing the equality of two high-dimensional covariance matrices, which is designed to detect sparse and faint differences between the covariances. A novel U-statistic composition is developed to establish the asymptotic distribution of the thresholding statistics in conjunction with the matrix blocking and the coupling techniques. It is shown that the proposed test is more powerful than the existing tests in detecting sparse and weak signals in covariances. Multiple testing procedures are constructed to discover different covariances and the sub-groups of variables with different covariance structures between the two populations. The proposed procedures are based on the multi-level thresholding test, which is able to control the false discovery proportion (FDP) with high power. Simulation experiments and a case study on the returns of the SP500 stocks before and after the COVID-19 pandemic are conducted to demonstrate and compare the utilities of the proposed methods.

EO369 Room 105 (Hybrid 5) ADVANCES IN STATISTICAL LEANING THEORY AND LARGE-SCALE INFERENCE Chair: Peter Radchenko

E0637: **Spatially adaptive false discovery rate thresholding for sparse estimation**
Presenter: **Gourab Mukherjee**, University of Southern California, United States
Co-authors: Wenguang Sun, Jiajun Luo

A new false discovery rate (FDR) is developed based threshold estimator by extending the elegant FDR based estimator to spatial settings. The idea is to first construct robust and structure-adaptive weights by estimating local sparsity levels, and thereafter to set spatially adaptive thresholds using the weighted Benjamini-Hochberg procedure. We present asymptotic results demonstrating the superior performance of the proposed method. Through numerical experiments we illustrate the importance of spatial adaptation by studying the finite sample performance of the proposed estimator.

E0839: **Sparse high-dimensional regression with discrete optimization**
Presenter: **Peter Radchenko**, University of Sydney, Australia

Recent applications of discrete optimization techniques are discussed in high-dimensional regression, concentrating on the algorithmic framework for grouped variable selection. While there exist appealing approaches based on convex relaxations and nonconvex heuristics, we will focus on optimal solutions for the L0-regularized formulation, a problem that is less explored due to computational challenges. The proposed methodology covers nonparametric sparse additive modelling with smooth components and allows for pairwise interactions. Experiments based on the US Census Planning Database demonstrate that our methods automatically identify useful interactions among key factors that have been reported in earlier work by the US Census Bureau. In addition to being useful from an interpretability standpoint, our models lead to predictions that are comparable to popular black-box machine learning methods based on gradient boosting and neural networks.

E0465: **A burden shared is a burden halved: A fairness-adjusted approach to classification**
Presenter: **Bradley Rava**, University of Southern California, United States

The focus is on fairness in classification, where one wishes to make automated decisions for people from different protected groups. When individuals are classified, the decision errors can be unfairly concentrated in certain protected groups. We develop a fairness-adjusted selective inference (FASI) framework and data-driven algorithms that achieve statistical parity in the sense that the false selection rate (FSR) is controlled and equalized among protected groups. The FASI algorithm operates by converting the outputs from black-box classifiers to R-values, which are intuitively appealing and easy to compute. Selection rules based on R-values are provably valid for FSR control, and avoid disparate impacts on protected groups. The effectiveness of FASI is demonstrated through both simulated and real data.

EO443 Room Virtual R1 ADVANCED NONPARAMETRIC AND SEMIPARAMETRIC METHODS Chair: Yuanyuan Guo

E0459: **A unified approach to variable selection for Cox's proportional hazards model with interval-censored failure time data**
Presenter: **Mingyue Du**, Hong Kong Polytechnic University Shenzhen Research Institute, China

Cox's proportional hazards model is the most commonly used model for regression analysis of failure time data and some methods have been developed for its variable selection under different situations. We consider a general type of failure time data, case K interval-censored data, that include all of the other types discussed as special cases, and propose a unified penalized variable selection procedure. In addition to its generality, another significant feature of the proposed approach is that unlike all of the existing variable selection methods for failure time data, the proposed approach allows dependent censoring, which can occur quite often and could lead to biased or misleading conclusions if not taken into account. For the implementation, a coordinate descent algorithm is developed and the oracle property of the proposed method is established. The numerical studies indicate that the proposed approach works well for practical situations and it is applied to a set of real data arising from the Alzheimer's Disease Neuroimaging Initiative study.

E0568: **An information ratio-based goodness-of-fit test for copula models on censored data**
Presenter: **Tao Sun**, Renmin University of China, China
Co-authors: Yu Cheng, Ying Ding

A copula is a popular method for modeling the dependence among marginal distributions in multivariate censored data. As many copula models are available, it is essential to check if the chosen copula model fits the data well for analysis. Existing approaches to testing the fitness of copula models are mainly for complete or right-censored data. No formal goodness-of-fit (GOF) test exists for interval-censored or recurrent events data. We develop a general GOF test for copula-based survival models using the information ratio (IR) to address this research gap. It can be applied to any copula family with a parametric form, such as the frequently used Archimedean and Gaussian families. The test statistic is easy to calculate, and the test procedure is straightforward to implement. We establish the asymptotic properties of the test statistic. The simulation results show that the proposed test controls the type-I error well and achieves adequate power when the dependence strength is moderate to high. Finally, we apply our method to test various copula models in analyzing multiple real datasets. Our method consistently separates different copula models for all these datasets in terms of model fitness.

E0867: **Bayesian approach for interval-censored data with time-varying covariates effects**
Presenter: **Bin Zhang**, Cincinnati Children's Hospital Medical Center, United States

Cox regression is one of the most commonly used methods in the analysis of interval-censored failure time data. In many practical studies, the covariate effects on the failure time may not be constant over time. Time-varying coefficients are therefore of great interest due to their flexibility in capturing the temporal covariate effects. To analyze spatially correlated interval-censored time-to-event data with time-varying covariate effects, a Bayesian approach with a dynamic Cox regression model is proposed. The coefficient is estimated as a piecewise constant function and the number of jump points estimated from the data. A conditional autoregressive distribution is employed to model the spatial dependency. The posterior summaries are obtained via an efficient reversible jump Markov chain Monte Carlo algorithm. The properties of our method are illustrated by simulation studies as well as an application to smoking cessation data in southeast Minnesota.

EO341 Room Virtual R10 RECENT ADVANCES WITH SCALABLE AND HIGH-DIMENSIONAL METHODS Chair: Sayar Karmakar

E0682: **Bayesian sparse Gaussian mixture model in high dimensions**
Presenter: **Yanxun Xu**, Johns Hopkins University, United States

A Bayesian method is proposed to estimate high-dimensional Gaussian mixture models whose component centers exhibit sparsity using a continuous spike-and-slab shrinkage prior. We establish the minimax risk for parameter estimation in sparse Gaussian mixture models and show that the posterior contraction rate of the proposed Bayesian model is minimax optimal. Computationally, the posterior inference can be implemented via an efficient Gibbs sampler with data augmentation, circumventing the challenging frequentist nonconvex optimization-based algorithms. We also obtain a contraction rate for the misclustering error by using tools from matrix perturbation theory. The validity and usefulness of the proposed approach are demonstrated through simulation studies and the analysis of a single-cell sequencing dataset.

E0676: **Long-term prediction for high-dimensional regression**
Presenter: **Sayar Karmakar**, University of Florida, United States

Time-aggregated prediction intervals are constructed for a univariate response time series in a high-dimensional regression regime. A simple quantile-based approach to the LASSO residuals seems to provide reasonably good prediction intervals. We allow for a very general possibly heavy-tailed, possibly long-memory and possibly non-linear dependent error process and discuss both the situations where the predictors are assumed to form a fixed or stochastic design. Finally, we construct prediction intervals for hourly electricity prices over horizons spanning 17 weeks and compare them to selected Bayesian and bootstrap interval forecasts

E0664: **Generative multiple-purpose sampler for weighted M-estimation**
Presenter: **Minsuk Shin**, University of South Carolina, United States
Co-authors: Shijie Wang, Jun Liu

To overcome the computational bottleneck of various data perturbation procedures such as the bootstrap and cross-validations, we propose the Generative Multiple-purpose Sampler (GMS), which constructs a generator function to produce solutions of weighted M-estimators from a set of given weights and tuning parameters. The GMS is implemented by a single optimization without having to repeatedly evaluate the minimizers of weighted losses and is thus capable of significantly reducing the computational time. We demonstrate that the GMS framework enables the implementation of various statistical procedures that would be unfeasible in a conventional framework, such as the iterated bootstrap, bootstrapped cross-validation for penalized likelihood, and bootstrapped empirical Bayes with nonparametric maximum likelihood, etc. To construct a computationally efficient generator function, we also propose a novel form of neural network called the *weight multiplicative multilayer perceptron* to achieve fast convergence. The numerical results demonstrate that the new neural network structure enjoys a few orders of magnitude speed advantage in comparison to the conventional one. An R package called GMS is provided, which runs under Pytorch to implement the proposed methods and allows the user to provide a customized loss function to tailor to their own models of interest.

EO397 Room Virtual R11 RECENT DEVELOPMENTS IN FUNCTIONAL AND TIME SERIES DATA ANALYSIS Chair: Tianhao Wang

E0326: **Simultaneous warping and clustering of functional electrocardiogram**
Presenter: **Wei Yang**, University of Pennsylvania, United States
Co-authors: Wensheng Guo

The goal of clustering functional data is to identify distinct functional patterns in the entire domain. These functional data are usually subjected to phase variability distorting the observed patterns and requires curve registration to remove the phase variability. Curve registration requires a target to which a functional object is aligned. A natural target is the cross-sectional mean of the functional objects within the same cluster, which is not available prior to clustering. There is also a trade-off between flexible warping and clustering data into more clusters. The more the phase variability is removed through curve registration, the less the remaining variability in the data, which often leads to a smaller number of clusters. Consequently, the number of clusters based on the amplitude variability and flexibility of the warping are confounded. External information is required to determine the number of clusters and warping flexibility. We proposed an iterative method that performs simultaneous curve registration and clustering. We also proposed a unified criterion for selecting the number of clusters and the penalty parameter of the warping functions. The criterion is derived from the classification likelihood, evaluating the association of the cluster membership with an outcome variable, which penalizes the uncertainties of cluster memberships. We evaluated the method through simulation and applied it to the digital electrocardiographic data.

E0428: **A regression approach for large portfolio allocation**
Presenter: **Lei Huang**, Southwest Jiaotong University, China

Portfolio allocation is an important topic in financial data analysis. Based on the mean-variance optimization principle, we propose a synthetic regression model for the construction of portfolio allocation, and an easy-to-implement approach to generate the synthetic sample for the model. Compared with the regression approach in existing literature for portfolio allocation, the proposed method of generating the synthetic sample provides a more accurate approximation for the synthetic response variable when the number of assets under consideration is large. Due to the embedded leave-one-out idea, the synthetic sample generated by the proposed method has a weaker within sample correlation, which makes the resulting portfolio allocation closer to the optimal one. This intuitive conclusion is theoretically confirmed to be true by the asymptotic properties established. We have also conducted intensive simulation studies in this article to compare the proposed method with the existing ones, and found the proposed method works better. Finally, we apply the proposed method to real datasets. The yielded returns look very encouraging.

E0608: **Joint curve registration for longitudinal and survival data with application to Alzheimer's disease onset prediction**
Presenter: **Tianhao Wang**, Rush University Medical Center, United States

In studies of Alzheimer's disease (AD), there is great interest in understanding the progression of cognitive markers and developing a prognostic model for AD onset using the longitudinal cognitive markers. The conventional joint modeling approach for longitudinal and survival data requires a predetermined time scale, typically the time since baseline, in which every subject is assumed to have a comparable risk profile. However, in many observational AD studies, the participants entered the studies with heterogeneous cognition status at baseline, which leads to heterogeneous and incomparable risk profiles in time since baseline. We introduce a novel joint modeling approach based on the functional curve registration method. It assumes the longitudinal trajectories follow a flexible common shape function with a person-specific disease progression pattern characterized by a random curve registration function, which is further used to create a homogenous and comparable time scale for survival analysis. We propose a personalized dynamic prediction framework that can be updated as new observations are collected to reflect the patient's latest cognition status. Simulation studies and application to data from the Rush Religious Orders Study and Memory and Aging Project demonstrate the effectiveness of this new approach.

EO436 Room Virtual R2 NONPARAMETRIC/ HIGH DIMENSIONAL METHODS: NEUROIMAGING AND POINT CLOUDS Chair: Luo Xiao

E0552: **Tensor quantile regression with application to association between neuroimages and human intelligence**
Presenter: **Cai Li**, St. Jude Children's Research Hospital, United States
Co-authors: Heping Zhang

Human intelligence is usually measured by well-established psychometric tests. The recorded cognitive scores are continuous but usually heavy-tailed with potential outliers and violating the normality assumption. Meanwhile, magnetic resonance imaging (MRI) provides an unparalleled opportunity to study brain structures and cognitive ability. Motivated by association studies between MRI images and human intelligence, we propose a tensor quantile regression model, which is a general and robust alternative to the commonly used scalar-on-image linear regression. Moreover, we take into account rich spatial information of brain structures, incorporating low-rankness and piecewise smoothness of imaging coefficients into a regularized regression framework. Extensive numerical studies are conducted to examine the empirical performance of the proposed method and its competitors. Finally, we apply the proposed method to the Human Connectome Project. We are able to identify the most activated brain subregions associated with quantiles of human intelligence. The prefrontal and anterior cingulate cortex are found to be mostly associated with lower and upper quantile of fluid intelligence. The insular cortex associated with the median of fluid intelligence is a rarely reported region.

E1006: **Incorporation of spatial- and connectivity-based cortical brain distances in regularized regression**
Presenter: **Jaroslaw Harezlak**, Indiana University School of Public Health-Bloomington, United States
Co-authors: Timothy Randolph, Damian Brzyski, Joaquin Goni, Aleksandra Steiner

The aim is to address the problem of adaptive incorporation of spatial information and connectivity-based information in brain imagining data in the multiple linear regression setting. In the example considered, we model scalar outcomes as functions of the brain cortical properties, e.g. cortical thickness and cortical area. We utilize both connectivity and spatial proximity information to build adaptive penalty terms in the regularized regression problem. The general idea of incorporating external information in the regularization approach via linear mixed model representation has been recently established in our prior work, specifically in the method called: ridgified Partially Empirical Eigenvectors for Regression (riPEER). We incorporate multiple sources of information, including structural connectivity network structure as well as the spatial distance between the cortical regions to estimate the regression parameters with multiple penalty terms via a riPEER extension called disPEER (distance-based Partially Empirical Eigenvectors for Regression). We present a simulation study testing various realistic scenarios and apply disPEER to data arising from the Human Connectome Project (HCP) study.

E1007: **An efficient spline smoothing for 3D point cloud learning**
Presenter: **Xinyi Li**, Clemson University, United States
Co-authors: Shan Yu, Yueying Wang, Guannan Wang, Ming-Jun Lai, Lily Wang

Over the past two decades, we have seen an exponentially increased amount of point clouds of irregular shapes collected in various areas. Motivated by the importance of solid modeling for point clouds, we develop a novel and efficient smoothing tool based on multivariate splines over the tetrahedral partitions to extract the underlying signal and build up a 3D solid model from the point cloud. The proposed method can be used to denoise or deblur the point cloud effectively and provide a multi-resolution reconstruction of the actual signal. In addition, it can handle sparse and irregularly distributed point clouds and recover the underlying trajectory from globally and locally missing data. Furthermore, we establish the theoretical guarantees of the proposed method. Specifically, we derive the convergence rate and asymptotic normality of the proposed estimator and illustrate that the convergence rate achieves the optimal nonparametric convergence rate, and the asymptotic normality holds uniformly. We demonstrate the efficacy of the proposed method over traditional smoothing methods through extensive simulation examples.

EO333 Room Virtual R3 INNOVATIVE WEIGHTING METHODS FOR CAUSAL INFERENCE Chair: Steve Yadlowsky

E0446: **Assessing external validity over worst-case subpopulations**
Presenter: **Hongseok Namkoong**, Columbia University, United States

Study populations are typically sampled from limited points in space and time, and marginalized groups are underrepresented. To assess the external validity of randomized and observational studies, we propose and evaluate the worst-case treatment effect (WTE) across all subpopulations of a given size, which guarantees positive findings remain valid over subpopulations. We develop a semiparametrically efficient estimator for the WTE that analyzes the external validity of the augmented inverse propensity weighted estimator for the average treatment effect. Our cross-fitting procedure leverages flexible nonparametric and machine learning-based estimates of nuisance parameters and is a regular root-n estimator even when nuisance estimates converge more slowly. On real examples where external validity is of core concern, our proposed framework guards against brittle findings that are invalidated by unanticipated population shifts.

E0950: **The basis for inference based on synthetic control methods**
Presenter: **David Hirshberg**, Emory University, United States

Synthetic Control methods are becoming popular far beyond the context of comparative case studies in which they were first proposed. It is no longer the rule that they are used only when we have one (or few) treated units. But despite recent attention, there is little consensus on when they work and how to do inference based on them. That there is no one way to think about panel data makes this difficult. In some interpretations, we are solving what is essentially a matrix completion problem with noise that is completely unrelated to the selection of treatment; in others, we are inverse propensity weighting to adjust for the selection of treatment based on past outcomes, noise and all. We will discuss some results characterizing synthetic control estimation based on these two interpretations, drawing on synthetic control estimators for panel data as well as that on covariate balancing or calibrated inverse propensity weighting estimators for cross-sectional data. We will highlight some issues that become apparent when we try to mix these perspectives, approaching inference based on the selection of treatment from a perspective in which behaviors specific to individual units, i.e. fixed effects—interactive or otherwise, are needed to explain the heterogeneity of the data.

E0440: **Quantile-based test for heterogeneous treatment effects**
Presenter: **EunYi Chung**, University of Illinois at Urbana Champaign, United States
Co-authors: Mauricio Olivares

One way to look at the distributional effects of a policy intervention comprises estimating the quantile treatment effect at different quantiles. We exploit this idea and develop a new permutation test for heterogeneous treatment effects based on a modified quantile process. To establish the asymptotic validity of the test, we transform the test statistic using a martingale transformation so that its limit behavior is distribution-free. Numerical evidence shows our permutation test outmatches other popular quantile-based tests in terms of size and power performance. We discuss a fast implementation algorithm and illustrate our method using experimental data from a welfare reform.

EO111 Room Virtual R4 ADVANCES IN STATISTICAL LEARNING FOR COMPLEX DATA Chair: Wenbo Wu

E0307: **Conditional probability tensor decompositions for multivariate categorical response regression**
Presenter: **Xin Zhang**, Florida State University, United States

In many modern regression applications, the response consists of multiple categorical random variables whose probability mass is a function of a common set of predictors. We consider a new method for modeling such a probability mass function in settings where the number of response variables, the number of categories per response, and the dimension of the predictor are large. We introduce a latent variable model which implies a low-rank tensor decomposition of the conditional probability tensor. We derive an efficient and scalable penalized expectation-maximization algorithm to fit this model and examine its statistical properties.

E0658: **Functional sparse group lasso**
Presenter: **Jun Song**, Korea University, Korea, South

A method will be presented for functional predictor selection and the estimation of smooth functional coefficients simultaneously in a scalar-on-function regression problem under a high-dimensional multivariate functional data setting. In particular, we develop a method for functional group-sparse regression under a generic Hilbert space of infinite dimension. Then we show the convergence of algorithms and the consistency of the estimation and selection under infinite-dimensional Hilbert spaces. Simulation and fMRI data application will be presented at the end to show the effectiveness of the methods in both the selection and estimation of functional coefficients.

E0806: **Pseudo sufficient dimension reduction with ill-conditioned sample covariance matrix**
Presenter: **Wenbo Wu**, University of Texas at San Antonio, United States

In high-dimensional data problems, the sample covariance matrix of the predictors is often singular either due to correlations among the predictors or due to an $n << p$ setting. Most sufficient dimension reduction methods rely on the inverse of the sample covariance as part of the estimation process. To conquer the challenge brought by the singular or near-singular sample covariance matrix, we propose a pseudo estimation approach by artificially adding random noises to the observed data. We show that with careful control of the added noises, the resulting estimator based on the perturbed data can still be consistent. In addition, a new variable selection procedure is proposed based on the pseudo estimator. The advantages of the proposed method are demonstrated by both simulation studies and real data analyses.

EO159 Room Virtual R5 L0-CONSTRAINED STATISTICAL LEARNING Chair: Ziwei Zhu

E0883: **Best subset selection is robust against design dependence**
Presenter: **Yongyi Guo**, Princeton University, United States
Co-authors: Ziwei Zhu, Jianqing Fan

Best subset selection(BSS) is among the most classical variable selection methods for high-dimensional linear regression. Nevertheless, BSS has recently received far less interest than its convex relaxed forms, largely because of NP-hardness. The aim is to invoke a renaissance in BSS by providing a computationally efficient implementation with superior variable selection performance. We first analyze the variable selection properties of BSS with known sparsity. We show that an identifiability margin condition is sufficient and nearly necessary for BSS to recover the true model. This condition is free of the restricted eigenvalues of the design, suggesting the robustness of BSS against design dependence. A relaxed version of this condition is sufficient for BSS to achieve the sure screening property when the true sparsity is overestimated. Next, we show that the established properties for BSS carry over to any near best subset. In particular, an approximate BSS algorithm based on two-stage iterative hard thresholding(IHT) can find a sparse sure screening subset within logarithmic steps. Based on this, the true model can be recovered easily. The simulations and real data examples show that this algorithm yields lower false discovery rates and higher true positive rates than competing approaches, especially under highly correlated design.

E0899: **Sure early selection by searching for the best subset**
Presenter: **Shihao Wu**, University of Michigan, Ann Arbor, China
Co-authors: Ziwei Zhu

In scientific discovery, it is often statistically intangible to identify all the important features with no false discovery, let alone the intimidating expense of experiments to test their significance. Such realistic limitation calls for a statistical guarantee for the early discovery of a model selector to navigate scientific adventure on the sea of big data. We focus on the early solution path of best subset selection (BSS), where the sparsity constraint is set to be lower than the true sparsity. Under a sparse high-dimensional linear model, we establish the sufficient and (near) necessary condition for BSS to achieve sure early selection, or equivalently, zero false discovery throughout its entire early path. Essentially, this condition boils down to a lower bound of the minimum projected signal margin that characterizes the fundamental gap in signal capturing between sure selection models and those with spurious discovery. Defined through projection operators, this margin is independent of the restricted eigenvalues of the design, suggesting the robustness of BSS against collinearity. On the computational aspect, we introduce a screen-then-select (STS) strategy to search for the best subset. Theoretical guarantee for sure early selection using the STS strategy is established. Numerical experiments show that the early solution paths of STS exhibit a much lower false discovery rate than competing approaches.

E0987: **Best subset selection in reduced rank regression**
Presenter: **Canhong Wen**, University of Science and Technology of China, China

Sparse reduced-rank regression is one of the most fundamental statistical approaches for investigating the association between large numbers of predictors and responses. While the advance in theory and algorithm is rapid, there still exists a gap between the algorithmic solution and theoretical guarantee, and no literature studies the computation complexity for achieving the statistical convergence rate. We propose a new method by constructing the algorithmic solution to estimate the sparse reduced-rank regression, which is motivated by the primal-dual formulation. Owing to the primal-dual mechanism, the main update of the algorithm is restricted to a small subset of predictors and thus its computation is efficient, especially in high dimensions. Under some mild conditions, we show that the algorithmic solution enjoys nice sampling properties including the consistency of estimation and support set recovery. Therefore, the new method fills up the gap between the theory and algorithm. Moreover, we propose a generalized-type information criterion for tuning the rank and sparsity level. Extensive numerical studies on synthetic and real data show that the proposal enjoys a nice performance on estimation, variable selection, and computation efficiency.

EO167 Room Virtual R6 ENVIRONMENTAL DATA SCIENCE Chair: Soutir Bandyopadhyay

E0610: **Spatial regression with nonparametric modeling of Fourier coefficients**
Presenter: **Chae Young Lim**, Seoul National University, Korea, South

Modeling of Fourier coefficients, known as a spectral density function, are considered to represent spatial dependence of a stationary spatial random field and use it for spatial regression under a Bayesian framework. Especially, we switch from the space domain to the frequency domain and introduce a Gaussian process prior to the log spectral density. As we do not impose any further assumption on log spectral density, the resulting covariance function is not of a parametric form and/or isotropic assumption. A simulation study supports that our approach is robust over various parametric covariance models. Also, our approach gives comparable or better prediction results over conventional spatial prediction under most parametric covariance models that we considered. Even though we need to estimate spectral density at all Fourier frequencies during the Bayesian procedure, our approach does not lose much computational efficiency compared to estimating only a few parameters in the parametric covariance models. We also compare our approach with some other existing spatial prediction approaches using two datasets of Korean ozone concentration. Our approach performs reasonably good in terms of mean absolute error and root mean squared error.

E0611: **Advantages of model misspecification for block data**
Presenter: **Soutir Bandyopadhyay**, Colorado School of Mines, United States

In many applications, spatial data are typically collected at areal levels (i.e., block data), while inferences and predictions are desired about the variable at points or blocks different from those at which the variable has been observed. The inferences and predictions typically depend on integrals that are often analytically intractable, and numerically expensive to approximate to high levels of accuracy. One may consider a naive approach to analyzing block data by converting it to a point-referenced counterpart by assuming that the whole mass (i.e., the integrated value) is observed at the centroid of each block. Such simplifications completely avoid the computational complexity associated with the analysis of change of support problems. We assess the extent to which both the block design and underlying process properties can affect the accuracy and stability of estimation and prediction tasks performed using this misspecification (relative to when these tasks are performed using a correctly specified observational model) and provide guidance for practitioners as to when this misspecification is inappropriate.

E0688: **Periodogram regression, a semi-parametric mixed effects approach for modelling non-stationary tropical cyclone frequency**
Presenter: **Sourav Das**, James Cook University, Australia
Co-authors: Lyuyuan Zhang, Guoqi Qian

Tropical cyclones (TC) are significant indicators of evolving climate dynamics. Two primary responses of interest are the cyclone frequency and intensity. We propose a novel integrated modelling framework for simultaneous modelling of TC frequency across several meteorological regions in Australasia. We take a two-stage semi-parametric approach where large scale environmental variation is modelled using generalized linear models and stochastic spatio-temporal variation is estimated using spectral analysis of time series. This framework offers flexibility in modelling and leads to a confluence of several disciplines in statistical methodology including, hierarchical modelling, generalized additive mixed-effects modelling and periodogram estimation, concluding with a linear prediction for integer-valued time series with forecast uncertainties.

EO217 Room Virtual R8 RECENT ADVANCES IN STATISTICAL METHODS FOR PRECISION MEDICINE Chair: Subharup Guha

E1015: **Unbiased multigroup comparisons by integrating multiple observational studies: A new concordant population approach**
Presenter: **Subharup Guha**, University of Florida, United States
Co-authors: David Christiani, Yi Li

The effective synthesis of information from multiple observational studies to make meta-analytic comparisons of multiple group responses is a challenging problem. Existing weighting and matching techniques cannot incorporate domain knowledge or directly analyze multiple cohorts with three or more groups (e.g., races), preventing the generalizability of the results to a natural population. We propose a new class of generalized balancing weights that incorporate known attributes of a larger population of interest into the target population, adjusting for under-sampled groups. Optimizing over any unknown attributes, we obtain the concordant target population. For censored outcomes, we propose balance-weighted Kaplan-Meier estimators to calculate confidence intervals and quantiles of the marginal survival curves of multiple groups. We devise small-sample procedures for uncertainty quantification and assess the performance through simulation studies. We apply the method to compare race-specific cancer survival among several TCGA glioblastoma multiforme (GBM) patient cohorts by adapting to the known racial decomposition of GBM in the U.S. population, and find that Blacks are more vulnerable and endure significantly worse prognoses.

E1020: **Semiparametric latent-class models for multivariate longitudinal and survival data**
Presenter: **Kin Yau Wong**, Hong Kong Polytechnic University, Hong Kong
Co-authors: Donglin Zeng, Danyu Lin

In long-term follow-up studies, data are often collected on repeated measures of multivariate response variables as well as on-time to the occurrence of a certain event. To jointly analyze such longitudinal data and survival time, we propose a general class of semiparametric latent-class models that accommodates a heterogeneous study population with flexible dependence structures between the longitudinal and survival outcomes. We combine nonparametric maximum likelihood estimation with sieve estimation and devise an efficient EM algorithm to implement the proposed approach. We establish the asymptotic properties of the proposed estimators through a novel use of modern empirical process theory, sieve estimation theory, and semiparametric efficiency theory. Finally, we demonstrate the advantages of the proposed methods through extensive simulation studies and provide an application to a motivating cohort study

E1026: **Propensity score matching and stratification for multiple and ordinal treatments: Application to an EHR-derived study**
Presenter: **Stacia DeSantis**, University of Texas Health Science Center at Houston, United States

Currently, methods for conducting multiple or ordinal treatments propensity scoring in the presence of high-dimensional covariate spaces that result from big data are lacking. The most prominent method relies on inverse probability treatment weighting (IPTW), which has limitations. We present a novel propensity scoring framework that uses the entire propensity score vector for multiple or ordinal treatments to establish a scalar balancing score that can achieve covariate balance in the presence of high-dimensional covariates. Specifically, we fit a one-parameter power function to the cumulative distribution function of the propensity score vector, resulting in a scalar balancing score that is used for matching and/or stratification. We present simulation results that show excellent performance in achieving covariate balance and estimating average treatment effects in the presence of multiple treatments. We then apply the approach to a study derived from electronic health records to determine the causal relationship between three different vasopressors and mortality in patients with non-traumatic aneurysmal subarachnoid hemorrhage. Results suggest that the method performs well when applied to large observational studies with multiple treatments that have large covariate spaces.

EO287 Room Virtual R9 ADVANCES IN HIGH-DIMENSIONAL SAMPLING METHODS Chair: Shiwei Lan

E0726: **Scaling up Bayesian uncertainty quantification for inverse problems using deep neural networks**
Presenter: **Shiwei Lan**, Arizona State University, United States

Due to the importance of uncertainty quantification (UQ), the Bayesian approach to inverse problems has recently gained popularity in applied mathematics, physics, and engineering. However, traditional Bayesian inference methods based on Markov Chain Monte Carlo (MCMC) tend to be computationally intensive and inefficient for such high dimensional problems. To address this issue, a surrogate-based method, calibration-emulation-sampling (CES), has recently been proposed for large dimensional UQ problems. We propose a novel CES approach for Bayesian inference based on deep neural network models for the emulation phase. The resulting algorithm is computationally more efficient and more robust against variations in the training set. Further, by using an autoencoder (AE) for dimension reduction, we have been able to speed up our Bayesian inference method up to three orders of magnitude. Overall, our method, henceforth called Dimension-Reduced Emulative Autoencoder Monte Carlo (DREAMC) algorithm, is able to scale Bayesian UQ up to thousands of dimensions for inverse problems. Using two low-dimensional (linear and nonlinear) inverse problems we illustrate the validity of this approach. Next, we apply our method to two high-dimensional numerical examples (elliptic and advection-diffusion) to demonstrate its computational advantages over existing algorithms.

E0770: **A quantum parallel Markov chain Monte Carlo**
Presenter: **Andrew Holbrook**, UCLA, United States

A novel quantum computing strategy is proposed for parallel MCMC algorithms that generate multiple proposals at each step. This strategy makes parallel MCMC amenable to quantum parallelization by using the Gumbel-max trick to turn the generalized accept-reject step into a discrete optimization problem. This allows us to embed target density evaluations within a well-known extension of Grover's quantum search algorithm. Letting P denote the number of proposals in a single MCMC iteration, the combined strategy reduces the number of target evaluations required from $O(P)$ to $O(P^1/2)$. We review both the rudiments of quantum computing and the Gumbel-max trick in order to elucidate their combination for as wide an audience as possible.

E0965: **Sampling via birth-death dynamics**
Presenter: **Yulong Lu**, University of Massachusetts, United States

Birth-death dynamics for sampling multimodal probability distributions are discussed. At the continuum level, this dynamics is described by an infinite-dimensional nonlinear and nonlocal ODE, which takes the target distribution as the unique invariant measure. The advantage of the birth-death dynamics is that it allows the global movement of the mass of a probability density directly from one mode to another in the phase space according to their relative weights, without the difficulty of going through low probability regions, suffered by any local dynamics such as the overdamped Langevin MCMC. We prove that the birth-death dynamics converge to the unique invariant measure with a uniform rate provided that the initial distribution, compared to the target, has a strictly lower bound. We also propose a practical interacting particle sampling scheme as a numerical implementation of the kernelized version of birth-death dynamics. The acceleration effect of birth-death for overdamped Langevin dynamics will be demonstrated via some analytical and numerical examples.

Saturday 04.06.2022 10:10 - 11:50 Parallel Session B – EcoSta2022

EI007 Room 101 (Hybrid 1) INNOVATIONS IN FUNCTIONAL DATA ANALYSIS (VIRTUAL) Chair: Fang Yao

E0161: **Testing stationarity of functional time series in the frequency domain**
Presenter: **Alexander Aue**, UC Davis, United States
Co-authors: Anne van Delft

Interest in functional time series has spiked in the recent past with both methodology and applications. A new stationarity test is discussed for functional time series based on frequency-domain methods. The proposed test statistic is based on joint dimension reduction via functional principal components analysis across the spectral density operators at all Fourier frequencies, explicitly allowing for frequency-dependent levels of truncation to adapt to the dynamics of the underlying functional time series. The properties of the test are derived both under the null hypothesis of stationary functional time series and under the smooth alternative of locally stationary functional time series. The methodology is theoretically justified through asymptotic results. Evidence from simulation studies and an application to annual temperature curves suggests that the test works well in finite samples.

E0432: **Semiparametric functional regression models with multivariate functional predictors**
Presenter: **Yehua Li**, University of California at Riverside, United States

Motivated by a crop yield prediction application using temperature trajectories and other scalar predictors, we consider two classes of semiparametric functional regression models. We jointly model cross-correlated functional predictors using multivariate functional principal component analysis (mFPCA), and use the mFPCA scores as predictors in a second stage semiparametric regression. In the proposed partially linear functional additive models (PLFAM), we predict the scalar response by both the parametric effects of the multivariate predictor and additive nonparametric effects of the mFPCA scores, and adopt the component selection and smoothing operator (COSSO) penalty to select relevant components and regularize the fitting. In the second class of semiparametric functional regression models, we also consider the interactions between the functional and multivariate predictors, where we assume the interaction depends on a nonparametric, single-index structure of the multivariate predictor to avoid the curse of dimensionality. We establish theoretical properties for both models, letting the number of principal components diverge to infinity. A fundamental difference between our framework and the existing high-dimensional semiparametric regression models is that the mFPCA scores are estimated with errors, the magnitudes of which increase with the order of FPC. The practical performances of the proposed methods are illustrated through analysis of the motivating crop yield data.

E1027: **Wrapped Gaussian process functional regression model for batch data on Riemannian manifold**
Presenter: **Jian Qing Shi**, Southern Univesity of Science and Technology, China

Regression is an essential and fundamental methodology in statistical analysis. Plenty of literature focuses on linear and nonlinear regression in the context of the Euclidean space. However, regression models in non-Euclidean spaces deserve more attention since people observed enormous manifold-valued data. Taking the advantage of massive manifold valued data, this talk will discuss a concurrent functional regression model for batch data on Riemannian manifolds by estimating both mean structure and covariance structure simultaneously. The response variable is considered to follow a wrapped Gaussian process distribution. A nonlinear relationship between manifold valued response variables and multiple Euclidean covariates can be captured by this model in which the covariates could be functional and scalar. The performance of the model has been tested on both simulated data and real data, which endorses it is an effective and efficient tool in conducting functional data regression on Riemannian manifolds.

EO141 Room 102 (Hybrid 2) NEW METHODS FOR CAUSAL INFERENCE Chair: Luke Keele

E0239: **Cumulative probability models and their utility in semi-parametric estimation of causal effects**
Presenter: **Andrew Spieker**, Vanderbilt University Medical Center, United States
Co-authors: Bryan Shepherd, Caroline Birdrow

G-computation is a longitudinal generalization of standardization designed to accommodate time-varying confounding. While especially useful for estimating causal effects of time-dependent treatments, the parametric g-formula is sometimes criticized for its sensitivity to departures from assumptions. Cumulative probability models have recently been developed as a semi-parametric approach to modeling continuous outcome, through which one is able to model the CDF of an outcome conditional on covariates. We will show how cumulative probability models can be embedded within g-computation in order to bypass overly stringent parametric assumptions. We will then illustrate the utility of this methodology through Monte-Carlo illustrations and an application to a large cohort of women with endometrial cancer in order to compare cumulative medical costs associated with various adjuvant treatment strategies.

E0482: **Bayesian nonparametric methods for causal mediation analysis**
Presenter: **Jason Roy**, Rutgers University, United States

In many settings, interest is not just in the effect of an exposure on an outcome, but also on possible mechanisms. Causal mediation analysis aims to estimate how much of the impact of the exposure on the outcome is due to the exposure's impact on intermediate variable(s). However, the inference is challenging due to both the need for strong identifying assumptions and the need for multiple models. We describe some Bayesian nonparametric models that were motivated by the desire to avoid the risk of bias that comes with parametric modeling. We illustrate the methods with real examples for both the single mediator and multiple mediator scenarios.

E0483: **Estimating the causal effect of policy interventions in the presence of spillovers**
Presenter: **Nandita Mitra**, University of Pennsylvania, United States

Public policy interventions are commonly evaluated using the difference-in-differences approach. However, this approach does not directly account for the effect of the policy “spilling over‘ to neighboring regions such as nearby cities or states. For example, the implementation of an excise tax on sugar-sweetened and artificially sweetened beverages in the city of Philadelphia was shown to be associated with a substantial decrease in volume sales of taxed beverages in Philadelphia but also showed an increase in beverage volume sales in bordering counties which were not subject to the excise tax. The latter association could potentially be explained by cross-border shopping behaviors of Philadelphia city residents. To address these important concerns, we extend difference-in-differences methods to identify the causal effects of policy interventions under various spillover conditions. We propose doubly robust estimators for the average treatment effect on the treated and on the neighboring control. The new estimators relax the standard assumptions on interference and model specification. In addition, we formally define a new causal estimator for the average treatment effect on the treated as a function of neighborhood exposure to the policy intervention. Importantly, our approach allows one to generalize the causal effect of a policy change to other target populations that may be different from the original study population.

E0358: **Measuring racial disparities in emergency general surgery via approximate balancing weights**
Presenter: **Luke Keele**, University of Pennsylvania, United States

The basic research design for the study of racial disparities in surgical care uses statistical methods to compare black and white patients that are similar in terms of baseline characteristics. Differences in outcomes are interpreted as due to differences in care as a function of race. We develop a new form of approximate balancing weights for this purpose. Approximate balancing weights are generalizations of inverse propensity score weights that are designed to directly target covariate balance in the estimation process. This class of weighting methods solve a convex optimization problem to find a set of weights that target a specific loss function. The approximate balancing weights we develop rely on a hyper-parameter that governs the bias-variance trade-off in weighting. We also develop a data-driven method for hyper-parameter selection and review how outcome modeling can be applied for additional bias reduction. We conduct a series of simulation studies to understand bias reduction properties. We apply this method to study racial disparities in emergency general surgery. In one comparison, we only compare patients in terms of risk factors. In a second analysis, we compare patients on risk factors within the same hospital. We find that racial disparities in outcomes persist when we compare whites and blacks with similar risk factors. Racial disparities are eliminated when we compare similar patients within the same hospital.

EO447 **Room 103 (Hybrid 3)** **RECENT ADVANCES IN INFERENCE FOR COMPLEX STATISTICAL MODELS** **Chair: Jason Xu**

E0443: **ZAP: z-value adaptive procedures for false discovery rate control with side information**
Presenter: **Dennis Leung**, University of Melbourne, Australia

Adaptive multiple testing with covariates is an important research direction that has gained major attention in recent years, as it has been widely recognized that leveraging side information provided by auxiliary covariates can improve the power of testing procedures for controlling the false discovery rate (FDR), e.g. in the differential expression analysis of RNA-sequencing data, the average read depths across samples can provide useful side information alongside individual p-values, and incorporating such information promises to improve the power of existing methods. However, for two-sided hypotheses, the usual data processing step that transforms the primary statistics, generally known as z-values, into p-values not only leads to a loss of information carried by the main statistics but can also undermine the ability of the covariates to assist with the FDR inference. Motivated by this and building upon recent advances in false discovery rate research, we develop ZAP, a z-value based covariate-adaptive methodology. It operates on the intact structural information encoded jointly by the z-values and covariates, to mimic an optimal oracle testing procedure that is unattainable in practice; the power gain of ZAP can be substantial in comparison with p-value based methods.

E0804: **Likelihood-based inference for stochastic epidemic models via data augmentation**
Presenter: **Jason Xu**, Duke University, United States

Stochastic epidemic models such as the Susceptible-Infectious-Removed (SIR) model are widely used to model the spread of disease at the population level, but fitting these models to data presents significant challenges. In particular, the marginal likelihood is typically considered intractable in the presence of missing data, as practitioners resort to simulation methods or approximations. We discuss some recent contributions that enable direct inference using the likelihood of observed data, focusing on a perspective that makes use of latent variables to explore configurations of the missing data within a Bayesian framework. Motivated both by count data from large outbreaks and high-resolution contact data from mobile health studies, we show how a data-augmented MCMC approach successfully learns the interpretable epidemic parameters and scales to handle realistic data settings efficiently

E0868: **Spatio-temporal Bayesian modeling of crime in Philadelphia**
Presenter: **Shane Jensen**, The Wharton School of the University of Pennsylvania, United States

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

E0894: **Incorporating mechanistic knowledge in causal inference**
Presenter: **Alexander Volfovsky**, Duke University, United States

At their core, the assumptions needed for causal inference are concerned with removing the effects of potentially unobserved quantities. We may know that a drug is given, but maybe not when, or we may observe where a disease is transmitted but maybe not exactly from whom, yet in both settings, we might be interested in causal questions: Does the drug have an effect? Does a mitigation strategy work to prevent future transmission? Because these processes are governed by established biological mechanisms, mechanistic models can provide invaluable insights into the interactions between biological objects (drug diffusion in the body, transmission probabilities between individuals). Conditioning on these models can provide more credibility to the necessary assumptions for causal inference. We present two case studies of leveraging these types of models for causal inference: (1) we analyze observational data of critically ill patients and identify the effect of seizures if they were not treated, and (2) we employ a mechanistic model of disease transmission to help design a trial for evaluating a non-pharmaceutical intervention.

EO115 **Room 104 (Hybrid 4)** **METHODS FOR SURVIVAL DATA ANALYSIS I** **Chair: Takeshi Emura**

E0456: **Evaluating association between two event times with observations subject to informative censoring**
Presenter: **Dongdong Li**, Harvard Medical School, United States

The aim is to evaluate the association between two event times without specifying the joint distribution parametrically. This is particularly challenging when the observations on the event times are subject to informative censoring due to a terminating event such as death. We link the joint distribution of the two event times and the informative censoring time using a nested copula function. We use flexible functional forms to specify the covariate effects on both the marginal and joint distributions. In a semiparametric model for the bivariate event time, we estimate simultaneously the association parameters, the marginal survival functions, and the covariate effects. A byproduct of the approach is a consistent estimator for the induced marginal survival function of each event time conditional on the covariates. We develop an easy-to-implement pseudolikelihood-based inference procedure, derive the asymptotic properties of the estimators, and conduct simulation studies to examine the finite-sample performance of the proposed approach. For illustration, we apply our method to analyze data from the breast cancer survivorship study that motivated this research.

E0871: **Statistical methods for interval-censored multi-state data and mismeasured covariates with application in HIV care**
Presenter: **Hongbin Zhang**, CUNY (SPH), United States

In 2015, WHO announced the Treat All policy which recommends immediate antiretroviral therapy (ART) treatment of HIV infected people, regardless of disease severity. In evaluating the impact of adopting the Treat All policy at a national level, the relationship between the biomarkers such as CD4 counts and WHO clinical stages (1: asymptomatic; 2 mild; 3: advanced; 4: severe; 5: mortality) is investigated to assess the magnitude of Treat All effects that would go through (or not go through) CD4 counts, a strong proxy of ART treatment. The WHO clinical stage data are interval-censored as the exact time of stage to stage transition between the clinical visits is unobservable. The CD4 covariate can have a substantial measurement error. We proposed statistical methods for multi-state data subject to interval-censoring and mismeasured time-varying covariates: 1) two-steps method where the prediction of the true time-varying covariates was plugged into the outcome model for the estimation; and 2) joint model methods in which parameters from the longitudinal covariates model and from the survival model were simultaneously estimated where we implemented a computationally efficient method using the stochastic version of EM (StEM). The methods were applied to real-world service delivery data in Central Africa and evaluated with simulation.

E0451: **Disease progression based feature screening for ultrahigh-dimensional survival-associated biomarkers**
Presenter: **Mengjiao Peng**, East China Normal University, China, China
Co-authors: Liming Xiang

The increased availability of ultrahigh-dimensional biomarker data and the high demand for identifying biomarkers importantly related to survival outcomes made feature screening methods commonplace in the analysis of cancer genome data. In the presence of progression-free survival (PFS), typically interpreted as a surrogate endpoint for overall survival (OS) in cancer studies, the association between both OS and PFS has suggested a high concordance in both survival endpoints, namely, patients with higher PFS would most likely have longer OS times. We propose a novel feature screening method by incorporating information on PFS into the selection of important biomarker predictors for a more accurate inference of OS subsequent to disease progression. The proposal is based on the rank of correlation between individual features and the conditional distribution of OS given observations of PFS. It is advantageous for its flexible model nature, which requires no marginal model assumption for OS or PFS, and the minimal computational cost for implementation. Theoretical results show its ranking consistency, sure screening and false rate control properties. Simulation results demonstrate that the proposed screener leads to more accurate feature selection than the method without considering the prior information about PFS. An application to breast cancer genome data illustrates its practical utility and facilitates disease classification using selected biomarker predictors.

E0979: **Dependent Dirichlet processes for analysis of a generalized shared frailty model**
Presenter: **Chong Zhong**, The Hong Kong Polytechnic University, Hong Kong
Co-authors: Zhihua Ma, Junshan Shen, Catherine Liu

The Bayesian paradigm takes advantage of well fitting complicated survival models and feasible computing in survival analysis owing to the superiority in tackling the complex censoring scheme, compared with the frequentist paradigm. The aim is to display the latest tendency in Bayesian computing, in the sense of automating the posterior sampling, through Bayesian analysis of survival modeling for multivariate survival outcomes with complicated data structure. Motivated by relaxing the strong assumption of proportionality and the restriction of a common baseline population, we propose a generalized shared frailty model which includes both parametric and nonparametric frailty random effects to incorporate both treatment-wise and temporal variation for multiple events. We develop a survival-function version of the ANOVA dependent Dirichlet process to model the dependency among baseline survival functions. The posterior sampling is implemented by No-U-Turn sampler in Stan, a contemporary Bayesian computing tool, automatically. The proposed model is validated by analysis of the bladder cancer recurrences data. The estimation is consistent with existing results. Our model and Bayesian inference provide evidence that the Bayesian paradigm fosters complex modeling and feasible computing in survival analysis and Stan relaxes posterior inference.

EO327 Room 105 (Hybrid 5) ADVANCES IN STATISTICAL METHODS FOR OBSERVATIONAL STUDIES Chair: Rajarshi Mukherjee

E0464: **Personalized treatment selection using causal heterogeneity**
Presenter: **Kinjal Basu**, LinkedIn, United States

Randomized experimentation (also known as A/B testing) is widely used in the internet industry to measure the metric impact obtained by different treatment variants. A/B tests identify the treatment variant showing the best performance, which then becomes the chosen or selected treatment for the entire population. However, the effect of a given treatment can differ across experimental units and a personalized approach to treatment selection can greatly improve upon the usual global selection strategy. We develop a framework for personalization through (i) estimation of heterogeneous treatment effect at either a cohort or member-level, followed by (ii) selection of optimal treatment variants for cohorts (or members) obtained through (deterministic or stochastic) constrained optimization. We perform a two-fold evaluation of our proposed methods. First, a simulation analysis is conducted to study the effect of personalized treatment selection under carefully controlled settings. This simulation illustrates the differences between the proposed methods and the suitability of each with increasing uncertainty. We also demonstrate the effectiveness of the method through a real-life example related to serving notifications on Linkedin. The solution significantly outperformed both heuristic solutions and the global treatment selection baseline leading to a sizable win on top-line metrics like member visits.

E0469: **A splitting Hamiltonian Monte Carlo method for efficient sampling**
Presenter: **Lin Liu**, Shanghai Jiao Tong University, China

A splitting Hamiltonian Monte Carlo (SHMC) algorithm is proposed which can be computationally efficient when combined with the random mini-batch strategy. By splitting the potential energy into numerically nonstiff and stiff parts, one makes a proposal using the nonstiff part of U, followed by a Metropolis rejection step using the stiff part that is often easy to compute. The splitting allows efficient sampling from systems with singular potentials (or distributions with degenerate points) and/or with multiple potential barriers. In our SHMC algorithm, the proposal only based on the nonstiff part in the splitting is generated by the Hamiltonian dynamics, which can be potentially more efficient than the overdamped Langevin dynamics. We also use random batch strategies to reduce the computational cost to $O(1)$ per time step in generating the proposals for problems arising from many-body systems and Bayesian inference, and prove that the errors of the Hamiltonian induced by the random batch approximation are $O(\sqrt{\Delta t})$ in the strong and $O(\Delta t)$ in the weak sense, where Δt is the time step. Numerical experiments are conducted to verify the theoretical results and the computational efficiency of the proposed algorithms in practice.

E0759: **Optimal dynamic treatment regimes via smooth surrogate losses**
Presenter: **Nilanjana Laha**, Harvard University, United States
Co-authors: Aaron Sonabend, Rajarshi Mukherjee, Tianxi Cai

Large health care data repositories such as electronic health records (EHR) open new opportunities to derive individualized treatment strategies for complicated diseases. We will discuss the problem of estimating sequential treatment rules tailored to a patient's individual characteristics, often referred to as dynamic treatment regimes (DTRs). Our main objective is to find the optimal DTR that maximizes a discontinuous value function through direct maximization of Fisher's consistent surrogate losses. We demonstrate that a large class of concave surrogates fails to be Fisher consistent – a behavior that differs from traditional binary classification problems. We further characterize a non-concave family of Fisher consistent smooth surrogates that can be optimized via gradient descent using off-the-shelf machine learning algorithms. Compared to the existing direct search approaches under the support vector machine framework, our proposed method is more computationally scalable to large sample sizes and allows for broader functional classes for treatment policies. We establish theoretical properties for our proposed DTR estimator and obtain a sharp upper bound on the regret. The finite sample performance of our proposed estimator is evaluated through extensive simulations. Finally, we illustrate the working principles and benefits of our method using EHR data from sepsis patients admitted to intensive care units.

E0854: **Efficient and accurate genome-wide survival association analysis controlling for sample relatedness in biobanks**
Presenter: **Rounak Dey**, Harvard University, United States

With decades of electronic health records linked to genetic data, large biobanks provide unprecedented opportunities for systematically understanding the genetics of complex diseases. Genome-wide survival association analysis can identify genetic variants associated with ages of onset, disease progression, and lifespan. Apart from the obvious computational challenge that such analyses entail, statistical methods also need to adjust for unknown genetic ancestry structures and familial relatedness among the biobank participants. Further, due to the cohort-based recruitment strategy typically followed in biobanks, most phenotypes have severe heavy-censoring which can lead to extreme type I error inflation in standard asymptotic tests of no genetic effects. We developed an efficient and accurate frailty model approach for genome-wide survival association analysis of censored time-to-event (TTE) phenotypes by accounting for both population structure and relatedness. Our method utilizes state-of-the-art optimization strategies to reduce the computational cost, and the saddlepoint approximation to allow for the analysis of heavily censored phenotypes (>90%) and low-frequency genetic variants (down to minor allele count 20). We demonstrated the performance of our method through extensive simulation studies and analysis of five TTE phenotypes, including lifespan, with heavy censoring rates (90.9% to 99.8%) on 400,000 UK Biobank participants and 180,000 individuals in FinnGen.

EO421 Room 106 (Hybrid 6) RECENT ADVANCES IN QUANTILE REGRESSION METHODS Chair: Jeong Hoon Jang

E0284: **Forecasting conditional distributions of stock returns: Quantile regression with machine learning**
Presenter: **Cindy Yu**, Iowa State University, United States
Co-authors: Haitao Li, Guoliang Ma

Machine learning methods are developed to forecast conditional distributions of stock returns by estimating conditional quantiles over fine grid points on the unit interval through quantile regression. Machine learning makes it possible to model conditional quantiles as highly nonlinear functions of a large number of return predictors. We adopt Bayesian optimization with a Gaussian process that significantly improves the efficiency of hyperparameter tuning in machine learning. Simulation studies show that our methods can accurately predict the conditional distributions of complicated data generating processes. Empirical results show that our methods can identify stocks with extreme positive or negative returns and achieve superior performance in long-short investing.

E0339: **Zero-inflated quantile rank-score based test with application to scRNA-seq differential gene expression analysis**
Presenter: **Wodan Ling**, Fred Hutchinson Cancer Research Center, United States
Co-authors: Wenfei Zhang, Bin Cheng, Ying Wei

Differential gene expression analysis based on scRNA-seq data is challenging due to two unique characteristics of scRNA-seq data. First, multimodality and other heterogeneity of the gene expression among different cell conditions lead to divergences in the tail events or crossings of the expression distributions. Second, scRNA-seq data generally have a considerable fraction of dropout events, causing zero inflation in the expression. To account for the first characteristic, existing parametric approaches targeting the mean difference in gene expression are limited, while quantile regression that examines various locations in the distribution will improve the power. However, the second characteristic, zero inflation, makes the traditional quantile regression invalid and underpowered. We propose a quantile-based test that handles multimodality and zero inflation simultaneously. The proposed quantile rank-score based test for differential distribution detection (ZIQRank) is derived under a two-part quantile regression model. It comprises a test in logistic modeling for the zero counts and a collection of rank-score tests adjusting for zero inflation at multiple prespecified quantiles of the positive part. The testing decision is based on the combined p-value of the marginal tests. ZIQRank is asymptotically justified and shown to improve and complement the existing approaches through extensive simulation and real data studies.

E0466: **Estimation of causal quantile effects with a binary instrumental variable and censored data**
Presenter: **Bo Wei**, University of Michigan, United States
Co-authors: Limin Peng, Zhang Mei-Jie, Jason Fine

The causal effect of a treatment is of fundamental interest in the social, biological, and health sciences. Instrumental variable (IV) methods are commonly used to determine causal treatment effects in the presence of unmeasured confounding. We study a new binary IV framework with randomly censored outcomes where we propose to quantify the causal treatment effect by the concept of complier quantile causal effect (CQCE). The CQCE is identifiable under weaker conditions than the complier average causal effect when outcomes are subject to censoring, and it can provide useful insight into the dynamics of the causal treatment effect. Employing the special characteristic of the binary IV and adapting the principle of the conditional score, we uncover a simple weighting scheme that can be incorporated into the standard censored quantile regression procedure to estimate CQCE. We develop a robust nonparametric estimation of the derived weights in the first stage, which permits stable implementation of the second stage estimation based on existing software. We establish rigorous asymptotic properties for the proposed estimator, and confirm its validity and satisfactory finite-sample performance via extensive simulations. The proposed method is applied to a bone marrow transplant dataset to evaluate the causal effect of rituximab in diffuse large B-cell lymphoma patients.

E0918: **Function-on-function quantile regression model for predicting glucose levels and excursions**
Presenter: **Jeong Hoon Jang**, Yonsei University, Korea, South

A hypoglycemic event is characterized by a low excursion of blood glucose level (less than 70 mg/dL) that can lead to serious problems, including seizures or unconsciousness. The goal is to build a functional regression model that leverages glucose curve data (real-time continuous glucose measurements) collected by a wearable glucose monitor to enable accurate and timely prediction of future glucose levels and hypoglycemic events, ultimately providing patients with sufficient time to take necessary preemptive actions. One challenge is that hypoglycemic events denote low glucose excursions that may not be well captured and predicted by the means. Hence, we develop a function-on-function quantile regression model that can reveal how the entire distribution of the functional response (future glucose curve) varies with the functional predictor (current and recent glucose curve) in ways that might not be captured by mean regression. The model incorporates random effects to account for subject-specific glucose patterns and also allows the model parameters to depend on latent classes to capture characteristics of distinct unobserved subgroups of the population. An efficient Bayesian estimation scheme based on asymmetric Laplace likelihood is presented. The predictive performance of the proposed method is examined through simulations and real patient data collected by Indiana University Hospital.

EO273 Room 107 (Hybrid 7) RECENT ADVANCES IN LATENT VARIABLE ANALYSIS AND PSYCHOMETRICS Chair: Gongjun Xu

E0614: **Using Bayesian IRT for multi-cohort repeated measure design to extract latent change scores**
Presenter: **Chun Wang**, University of Washington, United States

Repeated measure data design has been used extensively in a wide range of fields. Oftentimes, such data may be collected from multiple study cohorts and harmonized, with the intention of gaining higher statistical power and enhanced external validity. Traditional analysis may fit a unidimensional item response theory (IRT) model to data from one time point and one cohort to obtain item parameters and use such parameters throughout the analysis. Such a simplified approach ignores the item residual dependencies in the repeated measure design on one hand, and on the other hand, it does not exploit the accumulated information from different cohorts. Instead, we propose two approaches: an integrative approach using a two-tier bi-factor model via concurrent calibration, and if such calibration fails to converge, a Bayesian sequential calibration approach that uses highly informative priors on overlapping items across studies to establish a common scale. The three approaches are demonstrated using Alzheimer's Diseases Neuroimage Initiative cognitive battery data. Interestingly, the latent change scores extracted from the two proposed approaches are better separated from measurement errors, hence they contain more signals to identify people with mild cognitive impairment at the greatest risk of conversion to Alzheimer's disease.

E0879: **A network approach to assessment data: Mapping item-response interactions in interaction maps**
Presenter: **Minjeong Jeon**, UCLA, United States

Conventional item response data analysis typically relies on several assumptions, such as local item independence, respondent independence, and homogeneity. However, these assumptions are often violated in practice and are difficult to verify. To weaken the reliance on these assumptions, we propose a new perspective on item response data - to view them as network data representing relationships between two types of actors, respondents, and items. In this network view on item response data, a tie between the two types of actors is made when a correct response is given to the item by the respondent. The probability of a tie between a respondent and an item is then modeled as a function of a person attribute, an item attribute, and a distance between the person and the item in a low-dimensional Euclidean space. In this latent space item response model, the probability of a tie is determined by the person and the item attributes as well as how closely or distantly the person is located from the item in latent space. We will explain how the conventional assumptions of local item independence, respondent independence, and homogeneity are relaxed in the proposed latent space item response model. Additional benefits of the proposed network perspective on item response data and the proposed modeling approach are discussed with empirical data examples.

E1016: **Determining the number of factors in high-dimensional generalized latent factor models**
Presenter: **Xiaoou Li**, University of Minnesota, United States

As a generalization of the classical linear factor model, generalized latent factor models are useful for analyzing multivariate data of different types, including binary choices and counts. An information criterion is proposed to determine the number of factors in generalized latent factor models. The consistency of the proposed information criterion is established under a high-dimensional setting where both the sample size and the number of manifest variables grow to infinity, and data may have many missing values. An error bound is established for the parameter estimates, which plays an important role in establishing the consistency of the proposed information criterion. This error bound improves several existing results and may be of independent theoretical interest. We evaluate the proposed method by a simulation study and an application to Eysenck's personality questionnaire.

E0876: **Learning large Q-matrix by restricted Boltzmann machines**
Presenter: **Gongjun Xu**, University of Michigan, United States

Estimation of the large Q-matrix in Cognitive Diagnosis Models (CDMs) with many items and latent attributes from observational data has been a huge challenge due to its high computational cost. Borrowing ideas from deep learning literature, we propose to learn the large Q-matrix by Restricted Boltzmann Machines (RBMs) to overcome the computational difficulties. Key relationships between RBMs and CDMs are identified. Consistent and robust learning of the Q-matrix in various CDMs is shown to be valid under certain conditions. The simulation studies under different CDM settings show that RBMs not only outperform the existing methods in terms of learning speed, but also maintain good recovery accuracy of the Q-matrix. In the end, we illustrate the applicability and effectiveness of our method through a real data analysis.

EO035 Room Virtual R1 FINANCIAL BIG DATA MODELING Chair: Donggyu Kim

E0218: **High-dimensional high-frequency regression**
Presenter: **Donggyu Kim**, KAIST, Korea, South
Co-authors: Minseok Shin

A novel high-dimensional regression inference procedure is developed for high-frequency financial data. Unlike usual high-dimensional regression for low-frequency data, we need to additionally handle the time-varying coefficient problem. To accomplish this, we employ the Dantzig selection scheme and apply a debiasing scheme, which provides well-performing unbiased instantaneous coefficient estimators. With these schemes, we estimate the integrated coefficient, and to further account for the sparsity of the beta process, we apply thresholding schemes. We call this Thresholding dEbiased Dantzig Integrated Beta (TEDI Beta). We establish asymptotic properties of the proposed TEDI Beta estimator. In the empirical analysis, we apply the TEDI Beta procedure to analyzing high-dimensional factor models using high-frequency data.

E0455: **Large volatility matrix analysis using global and national factor models**
Presenter: **Sung Hoon Choi**, University of Connecticut, United States
Co-authors: Donggyu Kim

Several large volatility matrix inference procedures have been developed, based on the latent factor model. They often assumed that there are a few common factors, which can account for volatility dynamics. However, previous research demonstrated that there are local factors. Especially, when analyzing the global stock market, we often observe that national-specific factors explain their own volatility dynamics. To account for this, we propose a Double Principal Orthogonal complEment Thresholding (Double-POET) method, based on multi-level factor models. We establish its asymptotic properties. Furthermore, we demonstrate the drawback of using the regular principal orthogonal component thresholding (POET) when the local factor structure exists. We also describe the blessing of dimensionality using Double-POET for local covariance matrix estimation. Finally, we investigate the performance of Double-POET estimators in an out-of-sample portfolio allocation study using international stocks from 13 financial markets.

E0241: **Factor and idiosyncratic VAR-Ito volatility models for heavy-tailed high-frequency financial data**
Presenter: **Minseok Shin**, KAIST, Korea, South
Co-authors: Donggyu Kim, Yazhen Wang, Jianqing Fan

Various parametric models have been developed to predict large volatility matrices, based on the approximate factor model structure. They mainly focus on the dynamics of the factor volatility with some finite high-order moment assumptions. However, empirical studies have shown that idiosyncratic volatility also has a dynamic structure and it comprises a large proportion of the total volatility. Furthermore, we often observe that the financial market exhibits heavy tails. To account for these stylized features in financial returns, we introduce a novel Ito diffusion process for both factor and idiosyncratic volatilities whose eigenvalues follow the vector auto-regressive (VAR) model. We call it the factor and idiosyncratic VAR-Ito (FIVAR-Ito) model. To handle the heavy-tailedness and curse of dimensionality, we propose a robust parameter estimation method for a high-dimensional VAR model. We apply the robust estimator to predicting large volatility matrices and investigate its asymptotic properties. Simulation studies are conducted to validate the finite sample performance of the proposed estimation and prediction methods. Using high-frequency trading data, we apply the proposed method to large volatility matrix prediction and minimum variance portfolio allocation and showcase the new model and the proposed method.

E0253: **Effect of the U.S.-China trade war on stock markets: A financial contagion perspective**
Presenter: **Minseog Oh**, KAIST, Korea, South
Co-authors: Donggyu Kim

The effect of the U.S.-China trade war on stock markets is investigated from a financial contagion perspective, based on high-frequency financial data. Specifically, to account for risk contagion between the U.S. and China stock markets, we develop a novel jump-diffusion process. For example, we consider three channels for volatility contagion-such as integrated volatility, positive jump variation, and negative jump variation-and each stock market is able to affect the other stock market as an overnight risk factor. We develop a quasi-maximum likelihood estimator for model parameters and establish its asymptotic properties. Furthermore, to identify contagion channels and test the existence of a structural break, we propose hypothesis test procedures. From the empirical study, we find evidence of financial contagion from the U.S. to China and evidence that the risk contagion channel has changed from integrated volatility to negative jump variation.

EO037 Room Virtual R10 MODERN STATISTICS FOR CAUSALITY ANALYSIS AND HIGH-DIMENSIONAL INFERENCE Chair: Shujie Ma

E0245: **Estimating the average treatment effect in randomized clinical trials with all-or-none compliance**
Presenter: **Zhiwei Zhang**, National Cancer Institute, United States
Co-authors: Zonghui Hu, Dean Follmann, Lei Nie

Noncompliance is a common intercurrent event in randomized clinical trials that raises important questions about analytical objectives and approaches. Motivated by the Multiple Risk Factor Intervention Trial (MRFIT), we consider how to estimate the average treatment effect (ATE) in randomized trials with all-or-none compliance. Confounding is a major challenge in estimating the ATE, and conventional methods for confounding adjustment typically require the assumption of no unmeasured confounders, which may be difficult to justify. Using randomized treatment assignment as an instrumental variable, the ATE can be identified in the presence of unmeasured confounders under suitable assumptions, including an assumption that limits the effect-modifying activities of unmeasured confounders. We describe and compare several estimation methods based on different modeling assumptions. Some of these methods are able to incorporate information from auxiliary covariates for improved efficiency without introducing bias. The different methods are compared in a simulation study and applied to the MRFIT.

E0247: **Doubly robust interval estimation for optimal policy evaluation in online learning**
Presenter: **Hengrui Cai**, North Carolina State University, United States
Co-authors: Ye Shen, Rui Song

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. We aim to overcome these difficulties in policy evaluation for online learning. We explicitly derive the probability of exploration that quantifies the probability of exploring the non-optimal actions under commonly used bandit algorithms. We use this probability 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.

E0257: **Statistical exploitation of unlabeled data under high dimensionality**
Presenter: **Jiwei Zhao**, University of Wisconsin-Madison, United States

The benefits of unlabeled data in the semi-supervised learning setting under high dimensionality are considered, for parameter estimation and statistical inference. In particular, we address the following two important questions. First, can we use the labeled data as well as the unlabeled data to construct a semi-supervised estimator such that its convergence rate is faster than the supervised estimator? Second, can we construct confidence intervals or hypothesis tests that are guaranteed to be more efficient or powerful than the supervised estimator? We show that the semi-supervised estimator with a faster convergence rate exists under some conditions, and the implementation of this optimal estimator needs a reasonably good estimation of the conditional mean function. For statistical inference, we mainly propose a safe approach that is guaranteed to be no worse than the supervised estimator in terms of statistical efficiency. Not surprisingly, if the conditional mean function is well estimated, our safe approach becomes semi-parametrically efficient. After the theory development, we will also present some simulation results as well as a real data analysis.

E0406: **High-dimensional causal mediation analysis**
Presenter: **Yeying Zhu**, University of Waterloo, Canada

Causal mediation analysis has become popular in recent years, in which researchers not only aim to estimate the causal effect of a treatment, but also try to understand how the treatment affects the outcome through intermediate variables, namely mediators. We propose a set of generalized structural equations to estimate the direct and indirect effects for mediation analysis when the number of mediators is of high dimensionality. Specifically, a two-step procedure is considered where the penalization framework can be adopted to perform variable selection. The obtained estimators can be interpreted as causal effects without imposing the linear assumption on the model structure. The performance of Sobel's method in obtaining the standard error and confidence interval for the estimated joint indirect effect is also evaluated in simulation studies. The proposed method is applied to investigate how DNA methylation plays a role in the regulation of human stress reactivity impacted by childhood trauma.

EO247 Room Virtual R11 RECENT ADVANCES IN BIOMEDICAL AND EHR DATA ANALYSIS Chair: Wenlin Dai

E0779: **Estimating heterogeneous gene regulatory networks from zero-inflated single-cell expression data**
Presenter: **Xiangyu Luo**, Renmin University of China, China
Co-authors: Qiuyu Wu

Inferring gene regulatory networks can elucidate how genes work cooperatively. The gene-gene collaboration information is often learned by Gaussian graphical models (GGM) that aim to identify whether the expression levels of any pair of genes are dependent on other genes' expression values. One basic assumption that guarantees the validity of GGM is data normality, and this often holds for bulk-level expression data which aggregate biological signals from a collection of cells. However, fine-grained cell-level expression profiles collected in single-cell RNA-sequencing (scRNA-seq) reveal non-normality features—cellular heterogeneity and zero-inflation. We propose a Bayesian latent mixture GGM to jointly estimate multiple gene regulatory networks accounting for the zero-inflation and unknown heterogeneity of single-cell expression data. The proposed approach outperforms competing methods on synthetic data in terms of network structure and precision matrix estimation accuracy and provides biological insights when applied to two real-world scRNA-seq datasets.

E0782: **A modern theory for high-dimensional Cox regression models**
Presenter: **Huijuan Zhou**, Shanghai University of Finance and Economics, China

The proportional hazards model has been extensively used in many fields such as biomedicine to estimate and perform statistical significance testing on the effects of covariates influencing the survival time of patients. The classical theory of maximum partial-likelihood estimation (MPLE) is used by most software packages to produce inference, e.g., the coxph function in R and the PHREG procedure in SAS. We investigate the asymptotic behavior of the MPLE in the regime in which the number of parameters p is of the same order as the number of samples n. The main results are (i) the existence of the MPLE undergoes a sharp 'phase transition'; (ii) the classical MPLE theory leads to invalid inference in the high-dimensional regime. We show that the asymptotic behavior of the MPLE is governed by a new asymptotic theory. These findings are further corroborated through numerical studies. The main technical tool in our proofs is the Convex Gaussian Min-max Theorem (CGMT), which has not been previously used in the analysis of partial likelihood. Our results thus extend the scope of CGMT and shed new light on the use of CGMT for examining the existence of MPLE and non-separable objective functions.

E0875: **Multivariate varying-coefficient models via tensor decomposition**
Presenter: **Kejun He**, Renmin University of China, China
Co-authors: Raymond Ka Wai Wong, Ya Zhou, Fengyu Zhang

Multivariate varying-coefficient models (MVCM) are popular statistical tools for analyzing the relationship between multiple responses and covariates. Nevertheless, estimating large numbers of coefficient functions is challenging, especially with a limited amount of samples. We propose a reduced-dimension model based on the Tucker decomposition, which unifies several existing models. In addition, sparse predictor effects, in the sense that only a few predictors are related to the responses, are exploited to achieve an interpretable model and sufficiently reduce the number of unknown functions to be estimated. All the above dimension-reduction and sparsity considerations are integrated into a penalized least squares problem on the constraint domain of 3rd-order tensors. To compute the proposed estimator, we propose a block updating algorithm with ADMM and manifold optimization. We also establish the oracle inequality for the prediction risk of the proposed estimator. A real data set from Framingham Heart Study is used to demonstrate the good predictive performance of the proposed method.

E1014: **Use of electronic health records data for research: Challenges and opportunities**
Presenter: **Hulin Wu**, University of Texas Health Science Center at Houston, United States

The challenges and opportunities from the real-world Electronic Health Records (EHR) data are introduced and discussed from a Big Data perspective. In particular, we propose a 9-step procedure that describes the whole lifecycle of EHR research projects from project initiation and data extraction to the result dissemination: 1) Initiate a project: proposing a research topic with potential high-impact biomedical and clinical questions or hypotheses; 2) Data queries and data extraction; 3) Data cleaning; 4) Data processing; 5) Data preparation; 6) Data analysis, modeling and prediction; 7) Result validation; 8) Result interpretation; and 9) Publication and dissemination. This procedure is quite similar to the data mining procedure for knowledge discoveries in databases (KDD). From each of these steps, we will discuss the challenges and opportunities for statisticians. Real data examples from a large nationwide EHR database in the USA will be used to illustrate the principles and concepts of EHR data processing and analysis.

EO445 Room Virtual R12 RECENT ADVANCES IN RELIABILITY AND COUNTING PROCESSES Chair: Tony Sit

E0246: **Distributed censored quantile regression**
Presenter: **Tony Sit**, The Chinese University of Hong Kong, Hong Kong

An extension of censored quantile regression to a distributed setting is discussed. With the growing availability of massive datasets, it is oftentimes an arduous task to analyse all the data with limited computational facilities efficiently. The proposed method, which attempts to overcome this challenge, consists of two key steps, namely: (i) estimation of both Kaplan-Meier estimator and model coefficients in a parallel computing environment; (ii) aggregation of coefficient estimations from individual machines. We study the upper limit of the order of the number of machines for this computing environment, which, if fulfilled, guarantees that the proposed estimator converges at a comparable rate to that of the oracle estimator. In addition, we also provide two further modifications for distributed systems including (i) a divide-and-conquer approximation and (ii) a nonparametric counterpart for censored quantile regression. Numerical experiments are conducted to compare the proposed and the existing estimators. The promising results demonstrate the computation efficiency of the proposed methods. Finally, for practical concerns, a cross-validation procedure is also developed which can better select the hyperparameters for the proposed methodologies.

E0249: **Likelihood inference for one-shot device testing under frailty models**
Presenter: **Man Ho Ling**, The Education University of Hong Kong, Hong Kong

A device that performs its intended function only once is referred to as a one-shot device. The actual lifetimes of one-shot devices under life tests cannot be observed, and thus the lifetime information under test is very limited. In addition, one-shot devices often consist of multiple components that could cause the failure of the device. The components are coupled together in the manufacturing process or assembly, resulting in the failure modes possessing latent heterogeneity and dependence. Frailty models facilitate an easily understandable interpretation of the dependence between components. However, finding the maximum likelihood estimates of frailty models based on completely censored data is challenging. An efficient expectation-maximization algorithm is presented to find the maximum likelihood estimates of model parameters, on the basis of one-shot device testing data with multiple failure modes under a constant stress accelerated life test, with the dependent components having exponential lifetime distributions under gamma frailty. The maximum likelihood estimate and confidence intervals for the mean lifetime of the k-out-of-M structured one-shot device under normal operating conditions are also discussed. The performance of the proposed inferential methods is finally evaluated through Monte Carlo simulations.

E1013: **Censored quantile regression with time-dependent covariates**
Presenter: **Chi Wing Chu**, City University of Hong Kong, Hong Kong
Co-authors: Tony Sit, Zhiliang Ying

A class of censored quantile regression models is proposed for right-censored failure time data with time-dependent covariates. Upon a quantile-based transformation, a system of martingale-type functional estimating equations for the quantile parameters is derived. While time-dependent covariates naturally arise in time to event analysis, the little existing literature requires either an independent censoring mechanism or a fully observed covariate process even after the event has occurred. The proposed formulation extends the existing censored quantile regression model so that only the covariate history up to the observed event time is required. A recursive algorithm is developed to evaluate the estimator numerically. Asymptotic properties including uniform consistency and weak convergence of the proposed estimator as a process of the quantile level are established. Monte Carlo simulations and numerical studies on the clinical trial data of the AIDS Clinical Trials Group are presented to illustrate the numerical performance of the proposed estimator.

E1021: **On the computation of system signature**
Presenter: **Ping Shing Ben Chan**, The Chinese University of Hong Kong, Hong Kong

The computation of the signature of a complex system is often challenging, as it may involve a large number of components and a complex architecture. We will discuss two different algorithms to compute the signature of a system. The first algorithm is proposed to compute the signature of a system with exchangeable components. This new algorithm relies on the information of minimal cut sets or minimal path sets, which is very intuitive and efficient. The next algorithm is used for computing the signature of a system consisting of subsystems with shared components. It relies on a new concept called decomposed survival signature. Applications in cyber systems and transportation systems are highlighted.

EO095 Room Virtual R13 LEARNING FROM COMPLEX DATA: NEW DIRECTIONS AND INNOVATIONS Chair: Shan Yu

E0822: **Fusion learning of functional linear regression with application to genotype-by-environment interaction studies**
Presenter: **Shan Yu**, University of Virginia, United States
Co-authors: Aaron Kusmec, Lily Wang, Dan Nettleton

A sparse multi-group functional linear regression model is proposed to simultaneously estimate multiple coefficient functions and identify groups, such that coefficient functions are identical within groups and distinct across groups. By borrowing information from relevant subgroups of subjects, our method enhances estimation efficiency while preserving heterogeneity in model parameters and coefficient functions. We use an adaptive fused lasso penalty to shrink coefficient estimates to a common value within each group. We also establish some theoretical properties of the proposed estimators. To enhance computation efficiency and incorporate neighborhood information, we propose to use a graph-constrained adaptive lasso with a computationally efficient algorithm. Two Monte Carlo simulation studies have been conducted to study the finite-sample performance of the proposed method. The proposed method is applied to sorghum flowering-time data and hybrid maize grain yields from the Genomes to Fields consortium.

E0853: **Statistical inference for mean functions of 3D functional objects**
Presenter: **Lily Wang**, George Mason University, United States
Co-authors: Yueying Wang, Guannan Wang

Functional data analysis has become a powerful tool for statistical analysis for complex objects, such as curves, images, shapes, and manifold-valued data. Among these data objects, 2D or 3D images obtained using medical imaging technologies have been attracting researchers' attention. In general, 3D complex objects are usually collected within the irregular boundary, whereas the majority of existing statistical methods have been focused on a regular domain. To address this problem, we model the complex data objects as functional data and propose trivariate spline smoothing based on tetrahedralizations for estimating the mean functions of 3D functional objects. The asymptotic properties of the proposed estimator are systematically investigated where consistency and asymptotic normality are established. We also provide a computationally efficient estimation procedure for covariance function and corresponding eigenvalue and eigenfunctions and derive uniform consistency. Motivated by the need for statistical inference for complex functional objects, we then present a novel approach for constructing simultaneous confidence corridors to quantify estimation uncertainty. Extension of the procedure to a two-sample case is discussed together with numerical experiments and a real-data application using Alzheimer's Disease Neuroimaging Initiative database.

E0878: **Variable selection in mixture models via stochastic partitioning**
Presenter: **Mahlet Tadesse**, Georgetown University, United States

Identifying latent classes and component-specific relevant predictors can shed important insights when analyzing high-dimensional data. We will present methods to address this problem in a unified manner by combining ideas of mixture models and variable selection and fitting the models via stochastic partitioning. We will illustrate the performance of the methods in different application areas.

E0952: **A Gaussian copula function-on-scalar regression in reproducing kernel Hilbert spaces**
Presenter: **Linglong Kong**, University of Alberta, Canada

To relax the linear assumption in function-on-scalar regression, we borrow the strength of copula and propose a novel Gaussian copula function-on-scalar regression. Our model is more flexible to characterize the dynamic relationship between functional response and scalar predictors. Estimation and prediction are fully investigated. We develop a closed-form for the estimator of coefficient functions in a reproducing kernel Hilbert space without the knowledge of marginal transformations. Valid, distribution-free, finite-sample prediction bands are constructed via conformal prediction. Theoretically, we establish the optimal convergence rate on the estimation of coefficient functions and show that our proposed estimator is rate-optimal under fixed and random designs. The finite-sample performance is investigated through simulations and illustrated in real data analysis.

EO099 Room Virtual R2 ADVANCES IN BAYESIAN METHODOLOGY AND COMPUTATION Chair: James Flegal

E0544: **Globally-centered autocovariances in MCMC**
Presenter: **Dootika Vats**, Indian Institute of Technology, Kanpur, India
Co-authors: Medha Agarwal

Autocovariances are a fundamental quantity of interest in Markov chain Monte Carlo (MCMC) simulations with autocorrelation function (ACF) plots being an integral visualization tool for performance assessment. Unfortunately, for slow-mixing Markov chains, the empirical autocovariance can highly underestimate the truth. For multiple-chain MCMC sampling, we propose a globally-centered estimator of the autocovariance function (G-ACvF) that exhibits significant theoretical and empirical improvements. We show that the bias of the G-ACvF estimator is smaller than the bias of the current state-of-the-art. The impact of this improved estimator is evident in three critical output analysis applications: (1) ACF plots, (2) estimates of the Monte Carlo asymptotic covariance matrix, and (3) estimates of the effective sample size. Under weak conditions, we establish strong consistency of our improved asymptotic covariance estimator and obtain its large-sample bias and variance. The performance of the new estimators is demonstrated through various examples.

E0551: **Bayesian latent class model for multi-source domain adaptation**
Presenter: **Zehang Li**, University of California, Santa Cruz, United States

Distribution shift is a major challenge to deploying statistical and machine learning algorithms in many fields. We will discuss the challenge of distribution shift across different domains and methods to mitigate the issue in the context of assigning causes of death using verbal autopsies (VA). Worldwide, two-thirds of deaths do not have a cause assigned. VA is a well-established tool to collect information describing deaths outside of hospitals by conducting surveys with 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. A Latent Class model framework is discussed for VA data that jointly models VAs collected over multiple heterogeneous domains, assigns the cause of death for out-of-domain observations, and estimates cause-specific mortality fractions for a new domain. We introduce a parsimonious representation of the joint distribution of the collected symptoms using nested latent class models and develop an efficient algorithm for posterior inference.

E0754: **Lugsail lag windows for estimating time-average covariance matrices**
Presenter: **James Flegal**, University of California - Riverside, United States
Co-authors: Dootika Vats

Lag windows are commonly used in time series, econometrics, steady-state simulation, and Markov chain Monte Carlo to estimate time-average covariance matrices. In the presence of a positive correlation of the underlying process, estimators of this matrix almost always exhibit significant negative bias, leading to undesirable finite-sample properties. We propose a new family of lag windows specifically designed to improve finite-sample performance by offsetting this negative bias. Any existing lag window can be adapted into a lugsail equivalent with no additional assumptions. We use these lag windows within spectral variance estimators and demonstrate their advantages in a linear regression model with autocorrelated and heteroskedastic residuals. We further employ the lugsail lag windows in weighted batch means estimators due to their computational efficiency on large simulation output. We obtain bias and variance results for these multivariate estimators and significantly weaken the mixing condition on the process. Superior finite-sample properties are illustrated in a vector autoregressive process and a Bayesian logistic regression model.

E0943: **Semiparametric Bayesian discrete event time modeling**
Presenter: **Adam King**, California State Polytechnic University, Pomona, United States

Many event time outcomes are discrete, either because the process is inherently discrete (e.g., number of semesters to graduation) or because an underlying continuous process has been discretized (e.g., time to cessation of drug use recorded as whole number of months). Existing approaches for handling such discrete data include treating the survival times as continuous (with adjustments for tied outcomes), or using discrete models that omit important features like random effects. We present a general Bayesian discrete-time proportional hazards model, incorporating a number of features popular in continuous-time models such as competing risks, frailties, and generalized additive models style semiparametric incorporation of time-varying covariates (including flexible baseline hazards for time effects). These methods are implemented in a freely available R package called brea. We illustrate with analyses of college graduation rates and time to illicit drug use cessation.

EO209 Room Virtual R3 RECENT ADVANCES ON STATISTICAL MODELING OF COMPLEX DATA Chair: Youngdeok Hwang

E0966: **Joint semiparametric kernel machine network regression**
Presenter: **Byung-Jun Kim**, Michigan Technological University, United States
Co-authors: Inyoung Kim

A joint semiparametric kernel network regression model is developed for possibly nonlinear or non-additive associations and complicated interactions on both variable selection and network estimation. The approach is a unified and integrated method that can simultaneously identify important variables for a continuous outcome and build a network among the variables. The advantages of our proposed approach lie in flexibility, interactivity, and interpretability through the connection between the variable selection and the network estimation. We demonstrate our approach using real data application on genetic pathway-based analysis.

E0976: **Time delay estimation of traffic congestion based on statistical causality**
Presenter: **Sungil Kim**, Ulsan National Institute of Science and Technology (UNIST), Korea, South

Obtaining accurate time delay estimates is important in traffic congestion analysis because they can be used to address fundamental questions regarding the origin and propagation of traffic congestion. However, estimating the exact time delay during congestion is a challenge owing to the complex propagation process between roads and the high uncertainty regarding the future behavior of the process. To aid in accurate time delay estimation during congestion, we propose a novel time delay estimation method for the propagation of traffic congestion due to traffic accidents using lag-specific transfer entropy (TE). In the proposed method, nonlinear normalization with a sliding window is used to effectively reveal the causal relationship between the source and target time series in calculating the TE. Moreover, Markov bootstrap techniques are adopted to quantify uncertainty in the time delay estimator. To the best of our knowledge, our proposed method is the first to estimate the time delay based on the causal relationship between adjacent roads. We validated its efficacy using simulated data and real user trajectory data obtained from a major GPS navigation system applied in South Korea.

E0938: **Bayesian model calibration and sensitivity analysis for oscillating biological experiments**
Presenter: **Youngdeok Hwang**, City University of New York, United States

Most organisms exhibit various endogenous oscillating behaviors which provide crucial information as to how the internal biochemical processes are connected and regulated. Understanding the molecular mechanisms behind these oscillators requires interdisciplinary efforts combining both biological and computer experiments, as the latter can complement the former by simulating perturbed conditions with higher resolution. Harmonizing the two types of experiment, however, poses significant statistical challenges due to identifiability issues, numerical instability, and ill behavior in high dimensions. This article devises a new Bayesian calibration framework for oscillating biochemical models. The proposed Bayesian model is estimated using an advanced MCMC which can efficiently infer the parameter values that match the simulated and observed oscillatory processes. Also proposed is an approach to sensitivity analysis approach based on the intervention posterior. This approach measures the influence of individual parameters on the target process by utilizing the obtained MCMC samples as a computational tool. The proposed framework is illustrated with circadian oscillations observed in a filamentous fungus, Neurospora crassa.

E0591: **Blocking, rerandomization, and regression adjustment in randomized experiments with high-dimensional covariates**
Presenter: **Ke Zhu**, Tsinghua University, China
Co-authors: Hanzhong Liu, Yuehan Yang

Blocking, a special case of rerandomization, is routinely implemented in the design stage of randomized experiments to balance baseline covariates. Regression adjustment is highly encouraged in the analysis stage to adjust for the remaining covariate imbalances. Researchers have recommended combining these techniques; however, the research on this combination in a randomization-based inference framework with a large number of covariates is limited. Methods are proposed that combine blocking, rerandomization, and regression adjustment techniques in randomized experiments with high-dimensional covariates. In the design stage, we suggest the implementation of blocking or rerandomization or both techniques to balance a fixed number of covariates most relevant to the outcomes. For the analysis stage, we propose a regression adjustment method based on the Lasso to adjust for the remaining imbalances in the additional high-dimensional covariates. Moreover, we establish the asymptotic properties of the proposed estimator and outline conditions under which this estimator is more efficient than the unadjusted estimator. In addition, we provide a conservative variance estimator to facilitate valid inferences. Our analysis is randomization-based, allowing the outcome data generating models to be misspecified. Simulation studies and two real data analyses demonstrate the advantages of the proposed method.

EO223 Room Virtual R4 RECENT ADVANCES IN FINANCIAL TIME SERIES Chair: Cathy W-S Chen

E0242: **Online change point detection via copula based Markov models**
Presenter: **Li-Hsien Sun**, National Central University, Taiwan

Time series analysis is a critical issue in varied fields such as finance, industry, and biology. However, due to the possibility of the structure change, the corresponding problem such as loss or damage can be expected. See the stock market during the financial crisis in 2008 and also the COVID-19 in 2020 for instance. Hence, the corresponding change point for structural change is worth to study. In order to detect the change point online for time series data or correlated data, we propose the model for online change point detection via copula-based Markov models where the time serial data is described by copula-based Markov model and the change point detection based on the run-length distribution using the Bayesian approach. Finally, the performance of the proposed method is illustrated through numerical and empirical studies.

E0269: **A network autoregressive model with GARCH effects and its applications**
Presenter: **Shih-Feng Huang**, National University of Kaohsiung, Taiwan
Co-authors: Hsin-Han Chiang, Yu-Jun Lin

A network autoregressive model with GARCH effects, denoted by NAR-GARCH, is proposed to depict the return dynamics of stock market indices. A GARCH filter is employed to marginally remove the GARCH effects of each index, and the NAR model with the Granger causality test and Pearson's correlation test with sharp price movements is used to capture the joint effects caused by other indices with the most updated market information. The NAR-GARCH model is designed to depict the joint effects of nonsynchronous multiple time series in an easy-to-implement and effective way. The returns of 20 global stock indices from 2006 to 2020 are employed for our empirical investigation. The numerical results reveal that the NAR-GARCH model has satisfactory performance in both fitting and prediction for the 20 stock indices, especially when a market index has strong upward or downward movements.

E0310: **Pricing and hedging crypto options**
Presenter: **Huei-Wen Teng**, National Chiao Tung University, Taiwan
Co-authors: Wolfgang Karl Haerdle

Crypto options are financial derivatives where the underlying is related to a cryptocurrency. For example, the inverse Bitcoin option traded in Deribit exchange set up the strike price using the market value of a Bitcoin in USD but its payoff is converted in Bitcoin. Crypto options form a popular and inevitable asset class following the rapid and steady developments of cryptocurrencies. It has been documented that a stochastic volatility model with correlated jumps is dominating other GARCH-type or stochastic volatility models for cryptocurrencies under the physical measures. However, understanding the risk-neural pricing measure for crypto options remains scarce. Calibrating stochastic volatility models using option prices is challenging because there are no closed-form formulas of option prices. This issue becomes more prominent when jumps are involved in the model. We would like to compare the pricing and hedging performances of the stochastic volatility models with jumps for crypto options.

E0489: **A Bayesian analysis in long- and short-term financial volatility components with mixture distributions**
Presenter: **Edward Meng-Hua Lin**, Tunghai University, Taiwan

In forecasting market volatility studies, the consideration of incorporating external information associated with volatility components is an important and challenging issue. From related literature on volatility models, there is much evidence to show that using intraday range data as the volatility measure will construct more accurate volatility forecasts than using daily returns. Therefore, we will investigate whether using the mixed frequency data to incorporate the threshold conditional autoregressive range model with mixture distributions, which could capture the long- and short-term volatility components can improve the prediction ability of the volatility model. We conduct the following issues: (1) to propose a nonlinear range-based volatility model incorporating the long- and short-term volatility components to take volatility forecasting. (2) Considering the mixture GB2 distribution is ordered to capture more features of the implied volatility. (3) We employ the Bayesian approaches to estimate the different types of unknown parameters of the proposed nonlinear model simultaneously. (4) The GB2 density could be represented in various flexible distributions, then we conduct simulation studies to explore the effect of forecasting among its change of shape parameters. (5) We explore whether the long- and short-term volatility components and mixture distribution can improve the forecast performance through empirical analysis.

EO359 Room Virtual R5 TENSOR AND MULTILAYER NETWORKS Chair: Ivor Cribben

E0579: **Statistical limits for testing the correlation of hypergraphs**
Presenter: **Mingao Yuan**, North Dakota State University, United States
Co-authors: Zuofeng Shang

The focus is on the hypothesis testing of correlation between two -uniform hypergraphs on unlabelled nodes. Under the null hypothesis, the hypergraphs are independent, while under the alternative hypothesis, the hyperedges have the same marginal distributions as in the null hypothesis but are correlated after some unknown node permutation. We consider two scenarios: the hypergraphs are generated from the Gaussian-Wigner model and the dense Erdös-Rényi model. We derive the sharp information-theoretic testing threshold. Above the threshold, there exists a powerful test to distinguish the alternative hypothesis from the null hypothesis. Below the threshold, the alternative hypothesis and the null hypothesis are not distinguishable. The threshold involves and decreases as gets larger. This indicates testing the correlation of hypergraphs becomes easier than testing the correlation of graphs.

E0725: **Community detection with dependent connectivity**
Presenter: **Yubai Yuan**, University of California, Irvine, United States

In network analysis, within-community members are more likely to be connected than between-community members, which is reflected in that the edges within a community are intercorrelated. However, existing probabilistic models for community detection such as the stochastic block model(SBM) are not designed to capture the dependence among edges. We propose a new community detection approach to incorporate intracommunity-dependence of connectivities through the Bahadur representation. The proposed method does not require specifying the likelihood function, which could be intractable for correlated binary connectivities. In addition, the proposed method allows for heterogeneity among edges between different communities. In theory, we show that incorporating correlation information can achieve a faster convergence rate compared to the independent SBM, and the proposed algorithm has a lower estimation bias and accelerated convergence compared to the variational EM. Our simulation studies show that the proposed algorithm outperforms the existing multi-network community detection methods assuming conditional independence among edges. We also demonstrate the application of the proposed method to agricultural product trading networks from different countries and to brain fMRI imaging networks.

E0817: **Bayesian time-varying tensor vector autoregressive models for dynamic effective connectivity**
Presenter: **Wei Zhang**, Universita della Svizzera italiana, China
Co-authors: Ivor Cribben, Sonia Petrone, Michele Guindani

Recent developments in functional magnetic resonance imaging investigate how some brain regions directly influence the activity of other regions dynamically throughout the course of an experiment, namely dynamic effective connectivity. Time-varying vector autoregressive models have been employed to draw inferences for this purpose, but they are computationally intensive, since the number of parameters increases quadratically with the dimension of the time series. We propose a computationally efficient Bayesian time-varying VAR approach for modelling high-dimensional time series. The proposed framework employs a tensor decomposition for the VAR coefficient matrices at different lags. Dynamically varying connectivity patterns are captured by assuming that at any given time only a subset of components in the tensor decomposition is active. Latent binary time series select the active components at each time via a convenient Ising prior specification. The proposed prior structure encourages sparsity in the tensor structure and allows to ascertain model complexity through the posterior distribution. More specifically, sparsity-inducing priors are employed to allow for global-local shrinkage of the coefficients, to determine automatically the rank of the tensor decomposition and to guide the selection of the lags of the auto-regression. We show the performances of our model formulation via simulation studies and data from a real fMRI study involving a book reading experiment.

E0959: **Time series of weakly dependent tensors**
Presenter: **Dorcas Ofori-Boateng**, Portland State University, United States

There are several areas of application for time series of weakly dependent tensors across disciplines involving meteorological, brain imaging, power grid, electricity trading, and bitcoin data. Of course, for many of these datasets, we could consider full spatiotemporal covariances. For example, for meteorological data, the spacing in many cases is approximately a rectangular grid, which gives rise to a latitude against longitude or a latitude against longitude against the time tensor model. By considering the lower dimensional tensor, we can properly account for the dependence within the time component, allowing us to observe breaks/anomalies within the spatial structure of the tensor-related data. We extend tensor graphical models to weakly dependent data, that allows for the modelling of data with arbitrary tensor degree K. For example, in a functional magnetic resonance imaging (fMRI) data set that is collected over space, time, subjects, and multiple scans (or other modalities such as electroencephalography), our new estimator's three-way decomposition cannot only account for temporal dependence within a scan but also over longitudinal time. Next, we provide rates for statistical convergence for the new estimator. We demonstrate the usefulness of our developed algorithm on simulated datasets and against benchmark method results.

EO283 **Room Virtual R6** **RECENT DEVELOPMENT IN COMPLEX DATA ANALYSIS** **Chair: Yanlin Tang**

E0178: **Distributed variable selection for sparse regression under memory constraints**
Presenter: **Xuejun Jiang**, Southern University of Science and Technology, China

Variable selection is studied using the penalized likelihood method for distributed sparse regression with a large sample size n under a limited memory constraint, where the memory of one machine can only store a subset of data. This is a much-needed problem to be solved in the big data era. A naive divide-and-conquer method of solving this problem is to split the whole data into N parts and run each part on one of N machines, aggregate the results from all machines via averaging, and finally obtain the selected variables. However, it tends to select more noise variables, and the false discovery rate may not be well controlled. We improve it by a special designed weighted average in aggregation. Compared with the alternating direction method of multiplier (ADMM) to deal with massive data in the literature, our proposed methods reduce the computational burden a lot and perform better by mean square error in most cases. Theoretically, we establish asymptotic properties of the resulting estimators for the likelihood models with a diverging number of parameters. Under some regularity conditions, we establish oracle properties in the sense that our distributed estimator shares the same asymptotic efficiency as the estimator based on the full sample. Computationally, a distributed penalized likelihood algorithm is proposed to refine the results in the context of general likelihoods. Furthermore, the proposed method is evaluated by simulations and a real example.

E0191: **Robust distributed learning**
Presenter: **Xiaozhou Wang**, East China Normal University, China

The growing size of modern data brings new challenges to many classical statistical problems and calls for the development of distributed learning approaches. While in practice, distributed systems may be attacked or behave abnormally, which causes the distributed algorithms based on faultless systems invalid. We will introduce some research results on robust distributed learning. Algorithms and theoretical properties are given. Simulation studies are conducted to demonstrate the performance of the proposed methodologies.

E0302: **Bayesian empirical likelihood inference with complex survey data**
Presenter: **Puying Zhao**, Yunnan University, China

A Bayesian empirical likelihood approach is proposed to survey data analysis on a vector of finite population parameters defined through estimating equations. The method allows over-identified estimating equation systems and is applicable to both smooth and nondifferentiable estimating functions. Our proposed Bayesian estimator is design-consistent for general sampling designs and the Bayesian credible intervals are calibrated in the sense of having asymptotically valid design-based frequentist properties under single-stage unequal probability sampling designs with small sampling fractions. Large sample properties of the proposed Bayesian inference are established for both noninformative and informative priors under the design-based framework. We also propose a Bayesian model selection procedure with complex survey data and show that it works for general sampling designs. An efficient MCMC procedure is described for the required computation of the posterior distribution for general vector parameters. Simulation studies and an application to a real survey dataset are included to examine the finite sample performances of the proposed methods as well as the impact of different types of priors and different types of sampling designs.

E0351: **A regression framework of integrating genetic, imaging, and clinical data with applications**
Presenter: **Ting Li**, Shanghai University of Finance and Economics, China

The motivation comes from the joint analysis of genetic, imaging, and clinical (GIC) data collected in many large-scale biomedical studies, such as the UK Biobank study and the Alzheimer's disease neuroimaging initiative (ADNI) study. We propose a regression framework based on partially functional linear regression models to map high-dimensional GIC-related pathways for many phenotypes of interest. We develop a joint model selection and estimation procedure by embedding imaging data in the reproducing kernel Hilbert space and imposing the ℓ_0 penalty for the coefficients of scalar variables. We systematically investigate the theoretical properties of scalar and functional efficient estimators, including non-asymptotic error bound, minimax error bound, and asymptotic normality. We apply the proposed method to the ADNI dataset to identify important features from a large number of genetic polymorphisms and study the effects of a certain set of informative genetic variants and the surface data on the clinical outcomes.

EO313 Room Virtual R7 NONPARAMETRIC AND SEMIPARAMETRIC STATISTICS Chair: Yoshihiko Nishiyama

E0433: **Optimal minimax rates of specification testing with data-driven bandwidth**
Presenter: **Masamune Iwasawa**, Otaru University of Commerce, Japan
Co-authors: Yoshihiko Nishiyama, Kohtaro Hitomi

Optimal minimax rates of specification testing for linear and non-linear instrumental variable regression models are investigated. The test constructed by non-parametric kernel techniques can be rate optimal when bandwidths are selected appropriately. Since bandwidths are often selected in a data-dependent way in empirical studies, the rate optimality of the test with data-driven bandwidths are investigated. While the least-squares cross-validation selects bandwidths that are optimal for estimation, it is shown not to be optimal for testing. Thus, we propose a power maximizing bandwidth selection method that is optimal for testing.

E0439: **Improved confidence intervals for expectiles in risk management**
Presenter: **Yoshihiko Maesono**, Chuo University, Japan
Co-authors: Spiridon Penev

After the global financial crisis, bank regulators' activities were directed toward proposing measures of risk that could be an alternative to the VaR. The VaR is not a coherent risk measure mainly due to the fact that it does not satisfy the subadditivity property. Meanwhile, the elicitability property was pointed out as another essential requirement. It then turned out, that expectiles have it all as they are both coherent and elicitable. The focus is on the derivation of the Edgeworth expansion for the standardized and the studentized version of the kernel-based estimator of the expectile. Inverting the expansion allows us to construct accurate confidence intervals for the expectile when sample sizes are moderate. The methodology is illustrated with an application in risk management in finance for the estimation and accurate confidence interval construction for the coherent risk measure expectile-VaR.

E0429: **Asymmetric kernel density estimation for biased data**
Presenter: **Yoshihide Kakizawa**, Hokkaido University, Japan

For the data supported on $[0,\infty)$ or $[0,1]$, the so-called boundary bias problem is one of the interests, and asymmetric kernel density estimation has been well studied. We consider a situation where a random sample $\{X_1,\ldots,X_n\}$ is not directly available but the data $\{Y_1,\ldots,Y_n\}$ is instead observed from the length-biased distribution. Some previous works have been discussed without care of the boundary bias problem of the Rosenblatt-Parzen kernel density estimator. Indeed, an usual approximation (near the origin $x=0$) of a certain smooth function g as the convolution integral does not hold when $g(0)>0$, and that, even for the case $g(0)=0$, the order of the approximation (near the origin $x=0$) is slower when $g'(0)\neq 0$, This is the motivation that, instead of the location-scale kernel $k((x-\cdot)/h)/h$, we focus on an application of an asymmetric kernel, and then propose two density estimators.

E0425: **Higher-order asymptotic properties of the kernel density estimator with plug-in bandwidth**
Presenter: **Yoshihiko Nishiyama**, Kyoto University, Japan
Co-authors: Shunsuke Imai

The effect of bandwidth selection is investigated via the plug-in method on the asymptotic structure of the nonparametric kernel density estimator. We find that the plug-in method has no effect on the asymptotic structure of the estimator up to the order of $O(nh_0)^{1/2} = O(n^{L/(2L+1)})$ for a bandwidth h_0 and any kernel order L. We also provide the valid Edgeworth expansion up to the order of $O(nh_0)^{-1}$ and find that the plug-in method starts to have an effect on the term whose convergence rate is $O(nh_0)^{-1/2} * h0 = O(n^{(L+1)/(2L+1)})$. In other words, we derive the exact convergence rate of the deviation between the distribution functions of the estimator with a deterministic bandwidth and with the plug-in bandwidth. Monte Carlo experiments are conducted to see whether our approximation improves previous results.

EO319 Room Virtual R8 RECENT ADVANCES IN BAYESIAN DATA ANALYSIS Chair: Weixuan Zhu

E0566: **Ordinal causal discovery**
Presenter: **Yang Ni**, Texas AM University, United States

Causal discovery for purely observational, categorical data is a long-standing challenging problem. Unlike continuous data, the vast majority of existing methods for categorical data focus on inferring the Markov equivalence class only, which leaves the direction of some causal relationships undetermined. An identifiable ordinal causal discovery method is proposed that exploits the ordinal information contained in many real-world applications to uniquely identify the causal structure. The proposed method is applicable beyond ordinal data via data discretization. Through real-world and synthetic experiments, we demonstrate that the proposed ordinal causal discovery method combined with simple score-and-search algorithms has favorable and robust performance compared to state-of-the-art alternative methods in both ordinal categorical and non-categorical data. An accompanied R package OCD is freely available.

E0585: **Bayesian parameter inference and model selection for differential equation models**
Presenter: **Shijia Wang**, Nankai University, China

Nonlinear ordinary differential equations (ODEs) are used in a wide range of scientific problems to model complex dynamic systems, for example, transmission models for COVID-19. The differential equations often contain unknown parameters that are of scientific interest, which have to be estimated from noisy measurements of the dynamic system. Generally, there is no closed-form solution for nonlinear ODEs, and the likelihood surface for the parameter of interest is multi-modal and very sensitive to different parameter values. We will introduce our proposed sequential Monte Carlo (SMC) to conduct Bayesian inference for parameters in ODEs and model selection.

E1030: **Covariate dependent Beta-GOS process**
Presenter: **Weixuan Zhu**, Xiamen University, China

Covariate-dependent processes have been widely used in Bayesian nonparametric statistics thanks to their flexibility in incorporating covariate information and allowing for correlation among process realizations. Unlike most of the existing work that focuses on extensions of exchangeable species sampling processes such as the Dirichlet process, we propose a new class of covariate-dependent nonexchangeable priors by considering the generalization of the Beta-GOS model. We show that the proposed prior has an equivalent formulation under a continuous kernel mixture and its latent variable representation, which leads to a natural nonexchangeable parallel with the classical dependent Dirichlet process formulation. We further apply the proposed prior for regression and autoregressive models and show that its posterior sampling algorithm enjoys the same computational complexity as that of the Beta-GOS. We demonstrate the excellent numerical performance of our method via simulation and two real data examples.

E0511: **Bayesian outcome selection modelling**
Presenter: **Khue-Dung Dang**, University of Melbourne, Australia
Co-authors: Louise Ryan, Richard Cook, Tugba Akkaya-Hocagil, Sandra Jacobson, Joseph Jacobson

Psychiatric and social epidemiology often involves assessing the effects of environmental exposure on outcomes that are difficult to measure directly. To address this problem, it is common to measure outcomes using a comprehensive battery of different tests thought to be related to a common, underlying construct of interest. In the motivating application, researchers wanted to assess the impact of in-utero alcohol exposure on child cognition and neuropsychological development, which were evaluated using a range of different tests. Statistical analysis of the resulting multiple outcomes data can be challenging, not only because of the need to account for the correlation between outcomes measured on the same individual but because it is often unclear, a priori, which outcomes are impacted by the exposure under study. While researchers will generally have some hypotheses about which outcomes are important, a framework is needed to help identify outcomes that are sensitive to the exposure and to quantify the associated treatment or exposure effects of interest. We propose such a framework using a modification of stochastic search variable selection (SSVS), a popular Bayesian variable selection model and use it to quantify the overall effect of the exposure on the affected outcomes. We investigate the performance of the method via simulation and illustrate its application to data from a study involving the effects of prenatal alcohol exposure on child cognition.

EO335 Room Virtual R9 DESIGN AND ANALYSIS OF EXPERIMENTS Chair: Boxin Tang

E0200: **Computing multiple-objective optimal regression designs via CVX**
Presenter: **Julie Zhou**, University of Victoria, Canada

Model-based optimal regression designs with multiple objectives are common in practice. The objectives are often competitive, such as functions from A-, c-, D-, E-, and I-optimality criteria. It is extremely hard to derive analytical solutions for optimal designs with multiple objectives, and there are also no general and efficient algorithms for searching such designs for user-specified nonlinear models and criteria. We propose a new and effective approach for finding approximate multiple-objective optimal designs via the CVX solver. It can efficiently find different types of multiple-objective optimal designs after the optimization problems are carefully formulated as convex optimization problems. This approach is flexible and can be applied to any regression model. We present applications for minimax and efficiency constrained multiple-objective optimal designs.

E0337: **Linear orthogonal arrays of strength three and their applications**
Presenter: **Yuanzhen He**, Beijing Normal University, China
Co-authors: Guanzhou Chen

Orthogonal arrays of high strength can provide attractive designs for both physical experiments and computer experiments. Nevertheless, only a few orthogonal arrays of high strength are available in design literature, and most of their descriptions are in the language of specialized domains and terminology. These technical barriers have prevented high-strength orthogonal arrays from being widely applied. The aim is to relieve the above problem and provide a convenient reference for researchers and practitioners by investigating and reviewing the construction methods of linear orthogonal arrays of strength three. Two applications of these linear orthogonal arrays, constructions of blocked orthogonal arrays of strength three and strong orthogonal arrays of strength three, are introduced to illustrate their power in generating desirable designs for physical and computer experiments. The key components for these designs can be found in the text or Appendix so that people can use them conveniently.

E0476: **A study of orthogonal array-based designs under a broad class of space-filling criteria**
Presenter: **Guanzhou Chen**, Simon Fraser University, Canada
Co-authors: Boxin Tang

Space-filling designs based on orthogonal arrays are attractive for computer experiments for they can be easily generated with desirable low-dimensional stratification properties. Nonetheless, it is not very clear how they behave and how to construct good such designs under other space-filling criteria. We justify orthogonal array-based designs under a broad class of space-filling criteria, which include commonly used distance-, orthogonality- and discrepancy-based measures. To identify designs with even better space-filling properties, we partition orthogonal array-based designs into classes by allowable level permutations and show that the average performance of each class of designs is determined by two types of stratifications, with one of them being achieved by strong orthogonal arrays of strength 2+. Based on these results, we investigate various new and existing constructions of space-filling orthogonal array-based designs, including some strong orthogonal arrays of strength 2+ and mappable nearly orthogonal arrays.

E1018: **Column-orthogonal strong orthogonal arrays**
Presenter: **Min-Qian Liu**, Nankai University, China

Strong orthogonal arrays were recently introduced as a new class of space-filling designs for computer experiments due to their better stratifications than orthogonal arrays. To further improve the space-filling properties in low dimensions while possessing the column orthogonality, column-orthogonal strong orthogonal arrays of strength two stars and three are proposed. Construction methods and characterizations of such designs are provided. The resulting strong orthogonal arrays, with the number of levels being increased, have their space-filling properties in one and two dimensions being strengthened. They can accommodate comparable or even larger numbers of factors than those in the existing literature, enjoy flexible run sizes, and possess column orthogonality. The construction methods are convenient and flexible, and the resulting designs are good choices for computer experiments.

Saturday 04.06.2022 13:00 - 14:40 Parallel Session C – EcoSta2022

EI009 Room 101 (Hybrid 1) Recent advances in Bayesian econometrics and statistics (virtual) Chair: Toshiaki Watanabe

E0153: **Bayesian analysis of multiple networks for financial risk management**
Presenter: **Mike So**, The Hong Kong University of Science and Technology, Hong Kong

A network is a common tool for studying relationships or links among a set of nodes, like individuals or companies. The relationships can change over time and consist of interactions in multiple types of links or views. This kind of multi-view data can be observed in financial markets where the nodes in networks can be listed companies in a stock market and the links among the nodes represent various co-movement of returns and other financial random variables. To analyze the relationship among listed companies or stocks, we propose a model using observed covariates to explain interactions in different links. We also introduce a latent space to display the unobserved relationship on a plot and provide insights into financial risk management.

E0154: **On the quantile market risk factor model with heteroskedasticity, skewness, and leptokurtosis**
Presenter: **Cathy W-S Chen**, Feng Chia University, Taiwan
Co-authors: Kai Y-K Wang

The Fama French three-factor model advances the capital asset pricing model by expanding size risk and value risk factors to market risk factors. This research introduces a new quantile Fama-French three-factor model associated with heteroskedasticity, skewness, and leptokurtosis that allows for various market risk factor estimates under different market conditions. We employ an adaptive Bayesian Markov chain Monte Carlo sampling scheme to estimate all parameters in the proposed model over various quantile levels while assessing the performance via a simulation study. We analyze some daily stock returns from NASDAQ and further select the best model via the posterior odds ratio. It is clear that the various market conditions and GARCH effect should be incorporated into the model. Findings show that the estimation of the size factor turns insignificant for lower quantiles - i.e., when the market is in a panic, investors ignore the size effect of a company's assets.

E0684: **Measuring regional economic uncertainty**
Presenter: **Jouchi Nakajima**, Hitotsubashi University, Japan

An econometric framework is proposed for measuring the time-varying uncertainty of regional economic activity. A dynamic factor model with stochastic volatility is exploited to forecast the regional economic activity, and the uncertainty is defined as the conditional volatility of forecasted errors in the model. The framework is illustrated using Japan's regional economic data. We provide its application to climate-change analysis and show that irregular climate events significantly impact the uncertainty of economic activity.

EO183 Room 102 (Hybrid 2) Empirical covariance operators and beyond Chair: Moritz Jirak

E0308: **Bootstrapping the operator norm in high dimensions: Error estimation for covariance matrices and sketching**
Presenter: **Miles Lopes**, UC Davis, United States
Co-authors: Benjamin Erichson, Michael Mahoney

Although the operator (spectral) norm is one of the most widely used metrics for covariance estimation, comparatively little is known about the fluctuations of error in this norm. To be specific, let $\hat{\Sigma}$ denote the sample covariance matrix of n i.i.d. observations in $\mathbb{R}^p$ that arise from a population matrix Σ, and let $T_n = \|\hat{\Sigma}-\Sigma\|_{\text{op}}$. In the setting where the eigenvalues of Σ have a decay profile of the form $\lambda_j(\Sigma) \asymp j^{-2\beta}$, we analyze how well the bootstrap can approximate the distribution of T_n. The main result shows that up to factors of $\log(n)$, the bootstrap can approximate the distribution of T_n with respect to the Kolmogorov metric at the rate of $n^{-\frac{\beta-1/2}{6\beta+4}}$, which does not depend on the ambient dimension p. In addition, we offer a supporting result of independent interest that establishes a high-probability upper bound for T_n based on flexible moment assumptions. More generally, we discuss the consequences of our work beyond covariance matrices, and show how the bootstrap can be used to estimate the errors of sketching algorithms in randomized numerical linear algebra (RandNLA).

E0603: **Optimal and adaptive invariant shrinkage estimators for general large covariance and precision matrices**
Presenter: **Xiucai Ding**, UC Davis, United States

Some recent results are presented on the estimation of high dimensional covariance and precision matrices using Stein's invariant estimators under various loss functions. Our estimators are optimal and adaptive for general population covariance and precision matrices.

E0813: **Quantitative limit theorems and bootstrap approximations for empirical spectral projectors**
Presenter: **Martin Wahl**, Humboldt University of Berlin, Germany
Co-authors: Moritz Jirak

The problem of finding distributional approximations of the spectral projectors of an empirical covariance operator is considered. The problem will be studied in a dimension-free framework in which the data lives in high-dimensional or infinite-dimensional spaces and the complexity is characterized by the so-called relative rank of the population covariance operator. In this framework, novel quantitative limit theorems and bootstrap approximations are presented subject to mild moment conditions. In many cases, these results improve existing limit theorems established in a Gaussian setting.

E0941: **Logarithmic law for large sample correlation matrices**
Presenter: **Nestor Parolya**, Delft University of Technology, Netherlands
Co-authors: Johannes Heiny, Dorota Kurowicka

The log determinant of the sample correlation matrix based on a data matrix of size n and dimension p is found to satisfy a CLT (central limit theorem) for p/n in $(0,1]$ and $p \leq n$. Explicit formulas for the asymptotic mean and variance are provided. In case the population mean is unknown, we show that after recentering by the empirical mean the obtained CLT holds with a shift in the asymptotic mean. This result is of independent interest in both large dimensional random matrix theory and high-dimensional statistical literature of large sample correlation matrices for non-normal data. At last, the obtained findings are applied for testing of uncorrelatedness of p random variables. Surprisingly, in the null case $R = I$, the test statistic becomes completely pivotal and the extensive simulations together with the new theory show that the obtained CLT also holds if the moments of order four do not exist at all, which conjectures a promising and robust test statistic for heavy-tailed high-dimensional data.

EO325 Room 103 (Hybrid 3) DATA SCIENCE IN SOCIAL SCIENCE Chair: Yasumasa Matsuda

E0350: **Panel data quantile regression for treatment effect models**
Presenter: **Takuya Ishihara**, Tohoku University, Japan

A novel estimation method is developed for quantile treatment effects (QTE) under rank invariance and rank stationarity assumptions. The identification of the nonseparable panel data model under these assumptions has been recently explored, and a parametric estimation based on the minimum distance method has been proposed. However, when the dimensionality of the covariates is large, the minimum distance estimation using this process is computationally demanding. To overcome this problem, we propose a two-step estimation method based on the quantile regression and minimum distance methods. We then show the uniform asymptotic properties of our estimator and the validity of the nonparametric bootstrap. The Monte Carlo studies indicate that our estimator performs well in finite samples. Finally, we present two empirical illustrations, to estimate the distributional effects of insurance provision on household production and TV watching on child cognitive development.

E0460: **Health transition after retirement: Empirical evidence from public pension reform in Japan**
Presenter: **Michio Yuda**, Tohoku University, Japan
Co-authors: Fengming Chen, Midori Wakabayashi

An important common issue among developed counties with aging populations to clarify the mechanisms of retirement and health transition is to consider a problem between the elderly's quality of life and social costs of medical and long-term care expenditures. We use the individual panel data from the four waves of the Japanese Study of Aging and Retirement from 2007 to 2013 to examine how retirement from the labor market affects the health transition of elderly males. In the empirical analysis, we focus on the natural experiment of gradual pensionable age increases for the earnings-related public pension system in Japan depending on birth cohort and use fixed-effect instrumental variable estimation to deal with the endogenous problem of the retirement decision. We find that retirement significantly improves mental health and chewing ability but make them more susceptible to lifestyle-related diseases. Our supplemental results indicate that a significant increase in dentistry utilization after retirement would contribute to chewing improvement but that other daily habits and health care utilization are not significantly affected by retirement.

E0602: **Convolutional regression for big spatial data**
Presenter: **Yasumasa Matsuda**, Tohoku University, Japan

It is now common to collect big spatial data on a national or continental scale at discrete time points. The aim is to present a regression model where both dependent and independent variables are big spatial data. Regarding spatial data as functions over a region, we propose a functional regression by a parametric convolution kernel together with the least-squares estimation on the frequency domain by applying Fourier transform, which makes it possible to handle massive datasets with asymptotic validations under the mixed asymptotics. The regression is applied to new weekly cases of coronavirus disease 2019 (COVID-19) and human mobility collected in Japanese cities. We find that an increase in human mobility results in an increase of COVID-19 cases in a time lag of two weeks.

E0926: **Indian buffet process factor model for counterfactual analysis**
Presenter: **Stanley Iat-Meng Ko**, Tohoku University, Japan

A factor model-based counterfactual analysis is proposed. We explicitly estimate the underlying factor structure of the outcome variables and estimate the counterfactual values of the unit subject to an intervention. With the help of the non-parametric Bayesian Indian Buffet Process prior, our approach is capable of exploring heterogeneous factor exposures, and the number of latent factors is endogenously determined in the estimation process. The flexible Markov Chain Monte Carlo algorithm utilizes the maximal intervention-free information provided by the data, whereas the original synthetic control only uses pre-intervention data. The counterfactual values are estimated by simulating from the posterior predictive distribution such that we may integrate out any parameter estimation uncertainty. We also calculate the posterior predictive upper- and lower-quantile bounds for inference. The two applications, namely the California's Tobacco Control Program and the West German Reunification, demonstrate the usefulness of our approach compared to the synthetic control method and the elastic net model.

EO113 Room 104 (Hybrid 4) REGIME SWITCHING AND CHANGE DYNAMICS Chair: Matus Maciak

E0479: **Infinitely stochastic micro forecasting**
Presenter: **Michal Pesta**, Charles University, Czech Republic
Co-authors: Matus Maciak, Ostap Okhrin

Stochastic forecasting and risk valuation are now front burners in a list of applied and theoretical sciences. We propose an unconventional tool for stochastic prediction of future expenses based on the individual (micro) developments of recorded events. Considering a firm, enterprise, institution, or any entity, which possesses knowledge about particular historical events, there might be a whole series of several related subevents: payments or losses spread over time. This all leads to an infinitely stochastic process at the end. The aim, therefore, lies in predicting future subevent flows coming from already reported, occurred but not reported, and yet not occurred events. The emerging forecasting methodology involves marked time-varying Hawkes process with marks being other time-varying Hawkes processes. The estimated parameters of the model are proved to be consistent and asymptotically normal under simple and easily verifiable assumptions. The empirical properties are investigated through a simulation study. In the practical part of our exploration, we elaborate on a specific actuarial application for micro claims reserving.

E0480: **Real-time changepoint detection in a nonlinear expectile model**
Presenter: **Matus Maciak**, Charles University, Czech Republic
Co-authors: Michal Pesta, Gabriela Ciuperca

Regime switching within advanced stochastic models attracts a lot of interest over the last years with many different strategies being applied in this direction. We introduce a complex online changepoint detection procedure based on conditional expectiles. Nonlinearity of the underlying model improves the overall flexibility of the overall model, the conditional expectiles—well-known in econometrics for being the only coherent and elicitable risk measure—bring in some additional robustness, and the proposed changepoint detection test is proved to be consistent while the distribution under the null hypothesis depends on neither the functional form of the underlying model nor the unknown parameters which ensure very simple and straightforward applicability for real-life situations. Important theoretical details are summarized and finite sample empirical properties are presented.

E0550: **Economic policy uncertainty with ada-net**
Presenter: **Ostap Okhrin**, Technische Universitaet Dresden, Germany
Co-authors: Niels Gillmann

A local vector autoregressive model is developed. It allows us to estimate time-varying multivariate models without requiring the model parameters to change at every point in time. The estimation is done by a number of locally homogenous intervals and thereby identifying structural breaks. The local intervals are determined in a sequential testing procedure. This approach is especially suited for short time series since our approach usually results in only a few changes in the coefficients over time. We illustrate the method with simulations and a real data application. Using monthly Economic Policy Uncertainty data for five countries over 20 years, we show that uncertainty is connected across the world in a network.

Furthermore, there seem to be three major breaks in our sample. Namely, the Global Financial Crisis, the European sovereign debt crisis and the election of Donald Trump as president of the USA.

E0834: **Nonparametric maximum likelihood and related methods in infinite-dimensional situations: Convex optimization aspects**
Presenter: **Ivan Mizera**, University of Alberta, Canada

Certain nonparametric methods, including, but not limited to nonparametric maximum likelihood, and applied to problems like density estimation in shape-constrained situations, nonparametric estimation of mixture models, and others, are reviewed from the theoretical point of view. These methods have been already demonstrated to work in practical problems; the focus of the present analysis, rather than statistical properties - which have been to some extent analyzed in the literature elsewhere - is on the aspects of convex optimization employed in their implementation, in particular on the theoretical vindication of certain approximate strategies, possible aspects of regularization, and the potential of extensions to high-dimensional situations. Some practical details are discussed as well and illustrated on data-analytic examples.

EO423 **Room 105 (Hybrid 5)** **JOINT MODELING OF COMPLEX DEPENDENT DATA** **Chair: Xinyuan Song**

E0348: **Conditional copula models for correlated survival endpoints in individual patient data meta-analysis**
Presenter: **Takeshi Emura**, Kurume University, Japan
Co-authors: Virginie Rondeau, Sofeu Casimir

Existing copula models for the joint distribution of two failure times impose a simplifying assumption: the measure of correlation does not depend on covariates. This assumption is problematic when one tries to measure Kendall's tau between surrogate endpoint and true endpoint in a meta-analytical setup. We suggest extending the existing copula models so that Kendall's tau depends on covariates. Our newly proposed model, a joint frailty-conditional copula model, can effectively implement meta-analyses. In order to facilitate our approach, we develop an original R function "condCox.reg" and make it available in the R package joint.Cox. We apply the proposed method to a gastric cancer dataset with 3288 patients in 14 randomized trials from the GASTRIC group. This data analysis concludes that Kendall's tau has different values between the surgical treatment arm and the adjuvant chemotherapy arm.

E0504: **A tree-based Bayesian accelerated failure time cure model for estimating heterogeneous treatment effect**
Presenter: **Rongqian Sun**, The Chinese University of Hong Kong, China
Co-authors: Xinyuan Song

Estimating heterogeneous treatment effects has drawn increasing attention in medical studies, considering that patients with divergent features can undergo a different progression of disease even with identical treatment. Such heterogeneity can co-occur with a cured fraction for biomedical studies with a time-to-event outcome and further complicates the quantification of treatment effects. A joint framework of Bayesian causal forest and accelerated failure time cure model is considered to capture the cured proportion and treatment effect heterogeneity through three separate Bayesian additive regression trees. Under the potential outcomes framework, conditional and sample average treatment effects within the uncured subgroup are derived on the scale of log survival time, and treatment effects on the scale of survival probability are derived for each individual. Bayesian backfitting Markov chain Monte Carlo algorithm with the Gibbs sampler is conducted to estimate the causal effects. Simulation studies show the satisfactory performance of the proposed method. The proposed model is then applied to a breast cancer dataset extracted from the SEER database to demonstrate its usage in detecting heterogeneous treatment effects and cured subgroups.

E0505: **Robust joint estimation of treatment effect via possible dependent instrumental variables**
Presenter: **Yiqi Lin**, The Chinese University of Hong Kong, Hong Kong
Co-authors: Qingliang Fan, Xinyuan Song

The instrumental variable (IV) estimation with potentially invalid IVs is extended to allow for weak IVs and scenarios where the majority or plurality rules are difficult to hold or verify. In empirical research, weak IVs are common. A novel estimator, called WIT, is proposed to deal with invalid IVs and improve estimation accuracy under many weak IVs. We show that the WIT estimator works remarkably well under more relaxed identification conditions, which is unachievable in previous literature. Theoretical properties are derived for the proposed estimator. The finite sample property is demonstrated on simulated data and an empirical study concerning the effect of trade on economic growth.

E0647: **Bayesian adaptive lasso factor analysis models with pre- and post-test binary data**
Presenter: **Junhao Pan**, Sun Yat-sen University, China
Co-authors: Lijin Zhang

Binary data is frequently encountered in behavioral, educational and medical research. We extend previous work on the Bayesian covariance Lasso confirmatory factor analysis (CFA) model on the following aspects: (1) take the binary data into account by assuming that they are coming from an underlying latent continuous distribution with a threshold specification; and (2) handle potentially local dependency in item clusters by assigned the adaptive covariance Lasso prior to blocked diagonal residual covariance structure, which achieves model parsimony and generally fits the data better, while keeping the factor structure intact. We develop the Bayesian inference method based on the parameter expansion and Markov Chain Monte Carlo procedures. The simulation studies showed that the Bayesian estimates of the unknown parameters of interest are reliable. Real data on Pre- and Post-test Genetic Knowledge Items were also analyzed to evaluate the validity and practical usefulness of the proposed procedure.

EO285 **Room 106 (Hybrid 6)** **ADVANCES IN ANALYSIS OF COMPLEX DEPENDENT DATA** **Chair: Takashi Owada**

E0618: **Large deviation principle for geometric and topological functionals and associated point processes**
Presenter: **Takashi Owada**, Purdue University, United States

A large deviation principle is proved for the point process associated with k-element connected components in the d-dimensional Euclidean space with respect to the connectivity radii as a function of sample size. The random points are generated from a homogeneous Poisson point process so that the connectivity radius is of the so-called sparse regime. The rate function for the obtained large deviation principle can be represented as relative entropy. As an application, we deduce large deviation principles for various functionals and point processes appearing in stochastic geometry and topology. As concrete examples of topological invariants, we consider persistent Betti numbers of geometric complexes and the number of Morse critical points of the min-type distance function.

E0752: **Adaptive estimators for causal effects under network interference**
Presenter: **Fei Fang**, Duke University, United States
Co-authors: Alexandre Belloni, Alexander Volfovsky

The estimation of causal effects is increasingly relevant in different applied fields. We consider a causal inference problem in the presence of interference. The focus is on observational studies where interference across units is governed by a known network interference. However, the radius (and intensity) of interference is unknown and can be dependent on the observed treatment assignments in the relevant subnetwork. We study causal estimators for the average direct treatment effect given the network interference. The proposed estimators build upon a Lepski-like procedure that searches over the possible relevant radius/assignment patterns. In the process, we also obtain estimators for the radius of the interference that can be dependent on the treatment assignment of neighbors. Thus it creates an adaptive estimation of the network interference structure. We establish oracle inequalities and corresponding adaptive rates for the direct treatment effect estimators. The adaptive network interference can be defined over the labelled subgraphs themselves or on features of these, which recover many assumptions previously used in the literature. We present theoretical examples and numerical simulations that illustrate the performance of the proposed estimators.

E0930: **Testing for the independence of long-range dependent time series based on distance correlation**
Presenter: **Annika Betken**, University of Twente, Netherlands

The concept of distance correlation is applied for testing the independence of long-range dependent time series. For this, we establish a non-central limit theorem for stochastic processes with values in an L2-Hilbert space. This limit theorem is of general theoretical interest that goes beyond the considered context. For the purpose of testing the independence of time series, it provides the basis for deriving the asymptotic distribution of the distance covariance of subordinated Gaussian processes. Depending on the dependence on the data, the standardization and the limit of distance correlation vary. In any case, the limit is not feasible, such that test decisions are based on a subsampling procedure. We prove the validity of the subsampling procedure and assess the finite sample performance of a hypothesis test based on the distance covariance.

E0964: **High quantile regression for tail dependent time series**
Presenter: **Ting Zhang**, University of Georgia, United States

Quantile regression serves as a popular and powerful approach for studying the effect of regressors on quantiles of a response distribution. However, existing results on quantile regression were mainly developed when the quantile level is fixed, and the data are often assumed to be independent. Motivated by recent applications, we consider the situation where (i) the quantile level is not fixed and can grow with the sample size to capture the tail phenomena; and (ii) the data are no longer independent but collected as a time series that can exhibit serial dependence in both tail and non-tail regions. To study the asymptotic theory for high quantile regression estimators in the time series setting, we introduce a previously undescribed tail adversarial stability condition, and show that it leads to an interpretable and convenient framework for obtaining limit theorems for time series that exhibit serial dependence in the tail region but are not necessarily strong mixing. Numerical experiments are provided to illustrate the effect of tail dependence on high quantile regression estimators, where simply ignoring the tail dependence may lead to misleading p-values.

EO235 Room 107 (Hybrid 7) BAYESIAN METHODS AND THEIR APPLICATIONS Chair: Kuo-Jung Lee

E0626: **Sequential forecasting for bursty count data**
Presenter: **Kaoru Irie**, University of Tokyo, Japan
Co-authors: Aktekin Tevfik, Chris Glynn

Existing methods for sequentially analyzing count data typically utilize a discounting strategy, where the contribution of past observations to parameter updates diminishes with elapsed time. Discount factor techniques offer an intuitive approach to sequentially updating parameters with weighted contributions from all previously observed data; however, when the time series undergoes a sudden change and the observed count significantly jumps, parameter estimates are slow to adapt, as they are heavily informed by data observed prior to the structural break. We introduce an augmented Poisson-gamma state-space (PGSS) model whose state evolution structure is flexible and responsive to sudden changes in the level of counts, focusing on consumer demand settings where sequential and online learning/forecasting are of great interest. Such adaptability is achieved by augmenting the state vector of the PGSS model with an additional state variable for a time-varying discount factor. We develop an efficient particle-based estimation procedure that is suitable for sequential analysis, allowing us to estimate dynamic state variables and static parameters via closed-form conditional sufficient statistics. To illustrate how the state-augmented PGSS model performs with data that exhibit bursts, we present results from a case study to monitor and forecast web traffic data.

E0692: **A comparison of whole brain connectivity between depressed and non-depressed using a Bayesian spatio-temporal model**
Presenter: **Hakmook Kang**, Vanderbilt University, United States
Co-authors: Ilwoo Lyu, Kim Albert, Brian Boyd, Bennett Landman, Warren Taylor

A Bayesian double-fusion technique is introduced for enhancing the estimation of resting-state functional connectivity (FC) based on functional magnetic resonance imaging (fMRI) data between brain regions by using structural connectivity (SC) based on diffusion tensor imaging (DTI) data. Our previous work has been expanded to accommodate estimating the whole-brain functional connectivity matrix instead of focusing on a small number of regions of interest. Concurrently acquired two imaging data will be simultaneously used for FC estimation, which allows us to precisely investigate the relationship between FC and SC, or alterations in white matter microstructural integrity. The method is applied to multi-subject data ($n = 45$) with depression ($n = 20$) and without depression ($n = 25$) to examine how SC differences are related to differences in function (i.e., FC) and in turn related to cognitive task performance in depression.

E0774: **Variational Bayesian inference for network autoregression models**
Presenter: **Lai Wei-Ting**, National Cheng Kung University, Taiwan

A variational Bayesian (VB) method is developed for estimating large-scale dynamic network models in a network autoregressive framework. The proposed VB method allows automatic identification of the dynamic structure of such a model and obtains a direct approximation of the posterior density. Compared to Markov Chain Monte Carlo (MCMC) based sampling methods, the VB method improves computational efficiency without losing estimation accuracy. In real-world data analysis, we apply the proposed VB algorithm to day-ahead natural gas flow forecasting for the German gas transmission network with 51 nodes from October 2013 to September 2015. The VB method provides promising prediction accuracy as well as detected structural dynamic dependencies.

EO039 Room Virtual R1 MODELING TAIL EVENTS Chair: Abdelaati Daouia

E0787: **Composite bias-reduced Lp-quantile-based estimators of extreme quantiles and expectiles**
Presenter: **Antoine Usseglio-Carleve**, Avignon Universita, France
Co-authors: Gilles Stupfler

Quantiles are a fundamental concept in extreme value theory. They can be obtained from a minimization framework using an absolute error loss criterion. The companion notion of expectiles, based on squared rather than absolute error loss minimization, has received substantial attention from the fields of actuarial science, finance and econometrics over the last decade. Quantiles and expectiles can be embedded in a common framework of Lp-quantiles, whose extreme value properties have been explored very recently. Although this generalized notion of quantiles has shown potential for the estimation of extreme quantiles and expectiles, available estimators remain quite difficult to use: they suffer from substantial bias and the question of the choice of the tuning parameter p remains open. We work in a context of heavy tails, and we construct composite bias-reduced estimators of extreme quantiles and expectiles based on Lp-quantiles. We provide a discussion of the data-driven choice of p and of the anchor Lp-quantile level in practice. The proposed methodology is compared to existing approaches on simulated data and real data.

E0624: **Optimal pooling and distributed inference for the tail index and extreme quantiles**
Presenter: **Gilles Stupfler**, ENSAI - CREST, France
Co-authors: Abdelaati Daouia, Simone Padoan

Pooling strategies are investigated for tail index and extreme quantile estimation from heavy-tailed data. To fully exploit the information contained in several samples, we present general weighted pooled Hill estimators of the tail index and weighted pooled Weissman estimators of extreme quantiles calculated through a nonstandard geometric averaging scheme. We develop their large-sample asymptotic theory across a fixed number of samples, covering the general framework of heterogeneous sample sizes with different and asymptotically dependent distributions. Our results include optimal choices of pooling weights based on asymptotic variance and MSE minimization. In the important application of distributed inference, we prove that the variance-optimal distributed estimators are asymptotically equivalent to the benchmark Hill and Weissman estimators based on the unfeasible combination of subsamples, while the AMSE-optimal distributed estimators enjoy a smaller AMSE than the benchmarks in the case of large bias. We consider additional scenarios where the number of subsamples grows with the total sample size and effective subsample sizes can be low. An application to insurance data across several US states is presented.

E0328: **EV-GAN: Simulation of extreme events with ReLU neural networks**
Presenter: **Michael Allouche**, Ecole Polytechnique, France
Co-authors: Stephane Girard, Emmanuel Gobet

Feedforward neural networks based on Rectified linear units (ReLU) cannot efficiently approximate quantile functions which are not bounded, especially in the case of heavy-tailed distributions. We thus propose a new parametrization for the generator of a Generative adversarial network (GAN) adapted to this framework, based on extreme-value theory. We provide an analysis of the uniform error between the extreme quantile and its GAN approximation. It appears that the rate of convergence of the error is mainly driven by the second-order parameter of the data distribution. The above results are illustrated on simulated data and real financial data.

E0458: **Conditional expectile-based risk measures**
Presenter: **Cecile Adam**, KU Leuven, Belgium
Co-authors: Irene Gijbels

Among the main interests in regression analysis is to explore the influence that covariates have on a variable of interest, the response. There is extensive literature on flexible mean regression, in which the targeted quantity is the conditional mean of the response given the covariates. Quantile regression is another method that aims at estimating the conditional median or other quantiles of the response variable given the covariates. An alternative to quantiles are expectiles. Expectile regression estimates the conditional expectiles of the response variable given realized values of the predictor variables. After a brief introduction to expectiles and to univariate nonparametric expectile regression, we discuss the application of expectiles in risk management. Some risk measures are defined using the expectile regression framework and the estimators of these measures are established. The performance of these conditional risk measure estimators is investigated via simulations, and is illustrated on real data.

EO117 Room Virtual R2 STATISTICAL NETWORK DATA ANALYSIS Chair: Binyan Jiang

E0254: **Supervised centrality via sparse spatial autoregression, with an application to 2021 Henan floods social network**
Presenter: **LI Ting**, Hong Kong Polytechnic University, Hong Kong

The social opinions, behaviours and sentiments of the players in a social network are closely associated with their network positions. Identifying the influential players in a network is of importance as it helps understand how ties are formed, how information is propagated, and in turn, can guide the dissemination of new information by focusing on important players. Many notions of centrality have been proposed, most of which are based on the topology of the network. Motivated by a Weibo social network on 2021 Henan Floods where response variables on each node are available, we propose a novel notion of supervised centrality to account for the fact that the centrality of a node is task-specific. To estimate the supervised centrality and identify important players, we develop a novel sparse spatial autoregression model by introducing individual heterogeneity to each user. To overcome the computational difficulties with fitting the model for large social networks, we further develop a forward-addition algorithm and show that it can consistently identify a superset of the influential nodes. We apply our model to analyze three responses in the Henan Floods data: the number of comments, the number of reposts and the number of likes, and obtain interesting results. Simulation study further corroborates the developed theory.

E0592: **Quasi-score matching estimation for spatial autoregressive models with random weights matrix and regressors**
Presenter: **Tao Zou**, The Australian National University, Australia

Due to the rapid development of social networking sites, the spatial autoregressive (SAR) model has played an important role in social network studies. However, the commonly used quasi-maximum likelihood estimation (QMLE) for the SAR model is not computationally scalable as the network size is large. In addition, when establishing the asymptotic distribution of the parameter estimators of the SAR model, both weights matrix and regressors are assumed to be nonstochastic in classical spatial econometrics, which is perhaps not realistic in real applications. Motivated by the machine learning literature, quasi-score matching estimation for the SAR model is proposed. This new estimation approach is still likelihood-based, but significantly reduces the computational complexity of the QMLE. The asymptotic properties of parameter estimators under the random weights matrix and regressors are established, which provides a new theoretical framework for the asymptotic inference of the SAR type models. The usefulness of the quasi-score matching estimation and its asymptotic inference are illustrated via extensive simulation studies.

E0846: **Linear regression and its inference on noisy network-linked data**
Presenter: **Tianxi Li**, University of Virginia, United States
Co-authors: Can Minh Le

Linear regression on network-linked observations has been an essential tool in modeling the relationship between response and covariates with additional network structures. Previous methods either lack inference tools or rely on restrictive assumptions on social effects and usually assume that networks are observed without errors. A regression model with nonparametric network effects is proposed. The model does not assume that the relational data or network structure is exactly observed and can be provably robust to network perturbations. The asymptotic inference framework is established under a general requirement of the network observational errors, and the robustness of this method is studied in a specific setting when the errors come from random network models. We discover a phase-transition phenomenon of the inference validity concerning the network density when no prior knowledge of the network model is available while also showing a significant improvement achieved by knowing the network model. Simulation studies are conducted to verify these theoretical results and demonstrate the advantage of the proposed method over existing work in terms of accuracy and computational efficiency under different data-generating models. The method is then applied to middle school students' network data to study the effectiveness of educational workshops in reducing school conflicts.

E0849: **Extended stochastic block models via Gibbs-type priors**
Presenter: **Daniele Durante**, Bocconi University, Italy

Reliably learning group structures in network data is challenging in several applications. The focus is on converting networks that encode relationships among criminals. Such data exhibit a complex combination of an unknown number of core-periphery, assortative and disassortative structures that may unveil the architectures of the criminal organization. The coexistence of these noisy block patterns limits the reliability of routine community detection algorithms and requires extensions of model-based solutions to realistically characterize the node partition process, incorporate node attributes, and provide improved inference strategies. To cover these gaps, we will present a new class of extended stochastic block models (ESBM) that infer groups of nodes having common connectivity patterns via Gibbs-type priors on the partition process. This choice encompasses many realistic priors for criminal networks, covering solutions with fixed, random and infinite number of groups, and facilitates the inclusion of node attributes in a principled manner. Among the new alternatives in this class, we will focus on the Gnedin process as a realistic prior that allows the number of groups to be finite, random and subject to a reinforcement process coherent with criminal networks. A collapsed Gibbs sampler is proposed for the whole ESBM class, and improved inference strategies are outlined. The ESBM performance is illustrated in simulations and in an application to an Italian mafia network.

EO185 Room Virtual R3 ON THE SECOND-ORDER DYNAMICS OF INTRICATE FUNCTIONAL DATA Chair: Alessia Caponera

E0296: **Spectral density estimation of function-valued spatial processes**
Presenter: **Tailen Hsing**, University of Michigan, United States
Co-authors: Stilian Stoev, Rafail Kartsioukas

The spectral density of a stationary process fully characterizes the second- order properties of the process. We are interested in the estimation of the spectral density of a continuous-parameter stochastic process taking values in an infinite-dimensional Hilbert space. We assume that the process is observed at irregularly-spaced points, which is common in spatial statistics. We consider a lag-window estimator and explore the asymptotic properties of the estimator.

E0969: **Pivotal tests for relevant differences in the second order dynamics of functional time series**
Presenter: **Anne van Delft**, Columbia University, United States
Co-authors: Holger Dette

Motivated by the need to statistically quantify differences between modern (complex) data sets which commonly result in high-resolution measurements of stochastic processes varying over a continuum, we propose novel testing procedures to detect relevant differences between the second-order dynamics of two functional time series. In order to take into account the between-function dynamics that characterize this type of functional data, a frequency domain approach is taken. Test statistics are developed to compare differences in the spectral density operators and in the primary modes of variation as encoded in the associated eigenelements. Under mild moment conditions, we show convergence of the underlying statistics to Brownian motions and construct pivotal test statistics. The latter is essential because the nuisance parameters can be unwieldy and their robust estimation infeasible, especially if the two functional time series are dependent. In addition to these novel features, the properties of the tests are robust to any choice of frequency band enabling also to compare energy contents at a single frequency. The finite sample performance of the tests is verified through a simulation study and is illustrated with an application to fMRI data.

E0265: **Factor models for high-dimensional functional time series**
Presenter: **Shahin Tavakoli**, University of Geneva, Switzerland
Co-authors: Marc Hallin, Gilles Nisol

Theoretical foundations are set up for a high-dimensional functional factor model approach in the analysis of large cross-sections (panels) of functional time series (FTS). We first establish a representation result stating that, under mild assumptions on the covariance operator of the cross-section, we can represent each FTS as the sum of a common component driven by scalar factors loaded via functional loadings, and a mildly cross-correlated idiosyncratic component. The model and theory are developed in a general Hilbert space setting that allows for mixed panels of functional and scalar time series. We then turn to the identification of the number of factors, and the estimation of the factors, their loadings, and the common components. We provide a family of information criteria for identifying the number of factors and proving their consistency. We provide average error bounds for the estimators of the factors, loadings, and common components; the results encompass the scalar case, for which they reproduce and extend, under weaker conditions, well-established similar results. We provide numerical illustrations that corroborate the convergence rates predicted by the theory and provide a finer understanding of the interplay between N and T for estimation purposes. We conclude with an application to forecasting mortality curves, where we demonstrate that our approach outperforms existing methods.

E0780: **Nonparametric statistical inference for i.i.d. sparsely observed diffusions: An FDA perspective**
Presenter: **Neda Mohammadi Jouzdani**, EPFL, Switzerland
Co-authors: Leonardo Santoro, Victor Panaretos

Functional data analysis (FDA) covers an undeniably central role in studying different statistical inference problems, allowing to consider functional datasets on possibly complex domains, with trajectories observed discretely or continuously. Concerning discrete observations, this approach basically imposes some smoothness conditions on the sample paths and/or their covariance function to apply well-developed approximating methods. However, the usual regularity assumptions seriously limit the appropriateness of FDA in many commonly encountered settings, most notably stochastic differential equations. We introduce a careful modification of existing methods, dubbed the reflected triangle estimator and make inferences for the global (integral) behavior of the diffusion processes. We show that this allows for the FDA of processes with nowhere differentiable sample paths, even when these are discretely and noisily observed, including under irregular and sparse designs. We then proceed to relate the global behavior of the processes to their local (differential) behavior by means of an apparently novel PDE. We establish almost sure uniform asymptotic convergence rates of the proposed estimators as the number of observed curves grows to infinity. Our rates are non-asymptotic in the number of measurements per path, explicitly reflecting how different sampling frequencies might affect the speed of convergence.

EO151 Room Virtual R4 NON-LINEAR DEPENDENCE IN MULTIVARIATE TIME SERIES Chair: Hernando Ombao

E0236: **Estimation and inference for networks of multi-experiment point processes**
Presenter: **Ali Shojaie**, University of Washington, United States

Modern high-dimensional point process data, especially those from neuroscience experiments, often involve observations from multiple conditions and/or experiments. Networks of interactions corresponding to these conditions are expected to share many edges, but also exhibit unique, condition-specific ones. However, the degree of similarity among the networks from different conditions is generally unknown. To address these needs, we propose a joint estimation procedure for networks of high-dimensional point processes that incorporates easy-to-compute weights in order to data-adaptively encourage similarity between the estimated networks. We also propose a powerful hierarchical multiple testing procedure for edges of all estimated networks, which takes into account the data-driven similarity structure of the multi-experiment networks. Compared to conventional multiple testing procedures, our proposed procedure greatly reduces the number of tests and results in improved power, while tightly controlling the family-wise error rate. Unlike existing procedures, our method is also free of assumptions on dependency between tests, offers flexibility on p-values calculated along the hierarchy, and is robust to misspecification of the hierarchical structure.

E0563: **Dynamic topological data analysis on time varying trees and cycles**
Presenter: **Moo K Chung**, University of Wisconsin-Madison, United States
Co-authors: Hernando Ombao, Sixtus Dakurah

Persistent homology has been successfully applied to various static graphs and becoming a standard analysis tool. However, it is still not obvious how the method can be used in dynamically changing graphs over time. The challenge is obtaining continuous topological features over time, which might be contradictory since topological features are discrete and expected to be discontinued. We propose a coherent dynamic topological data analysis based on the newly discovered birth-death decomposition of graphs. By ignoring higher-dimensional topological features, it is possible to develop a mathematically coherent dynamic-TDA framework for time-varying graphs. The method is applied to quantify how the maximum spanning trees of the functional brain network of humans are topologically changing over time. We address various statistical challenges in tree and cycle modeling.

E0617: **Adaptive functional principal component analysis**
Presenter: **Jeff Goldsmith**, Columbia University, United States
Co-authors: Angel Garcia de la Garza, Britton Sauerbrei, Adam Hantman

Recent advances have allowed high-resolution observations of firing rates for a collection of individual neurons; these observations can provide insights into patterns of brain activation during the execution of tasks. Our data come from an experiment in which mice performed a reaching motion following an auditory cue, and contain measurements of firing rates from neuron activation in the motor cortex before and after the cue. In this setting, steep increases in firing rates after the cue are expected. Our dimension reduction technique adequately models these sharp changes over time and correctly captures these activation patterns. Initial results suggest different patterns of activation, representing the involvement of different motor cortex functions at different times in the reaching motion.

E0998: **Conex-connect: Learning patterns in extremal brain connectivity from multi-channel EEG data**
Presenter: **Raphael Huser**, King Abdullah University of Science and Technology, Saudi Arabia
Co-authors: Matheus Guerrero, Hernando Ombao

Epilepsy is a chronic neurological disorder affecting more than 50 million people globally. An epileptic seizure acts like a temporary shock to the neuronal system, disrupting normal electrical activity in the brain. Epilepsy is frequently diagnosed with electroencephalograms (EEGs). Current methods study only the time-varying spectra and coherence but do not directly model changes in extreme behavior, neglecting the fact that neuronal oscillations exhibit non-Gaussian heavy-tailed probability distributions. To overcome this limitation, we propose a new approach to characterize brain connectivity based on the joint tail behavior of the EEGs. Our proposed method, the conditional extremal dependence for brain connectivity (Conex-Connect), is a new approach that links the association between extreme values of higher oscillations at a reference channel with the other brain network channels. Using the Conex-Connect method, we discover changes in the extremal dependence driven by the activity at the foci of the epileptic seizure. Our model-based approach reveals that pre-seizure, the dependence is quite stable for all channels when conditioning on extreme values of the focal seizure area. By contrast, the dependence between channels is weaker during the seizure, and dependence patterns are more chaotic. Using the Conex-Connect method, we identified the high-frequency oscillations as the most relevant features explaining the conditional extremal dependence of brain connectivity.

EO375 **Room Virtual R5** **RECENT ADVANCES ON ACTUARIAL THEORY AND STATISTICS** **Chair: Yiying Zhang**

E0634: **Distortion risk contribution ratio measures: Definitions and comparisons**
Presenter: **Yiying Zhang**, Southern University of Science and Technology, China

Relative spillover effects play a key role in analyzing and comparing systemic risks. We introduce the so-called distortion risk contribution ratio measures. Various types of contribution ratio measures are defined and their useful integral-based representations are provided. We establish comparison results between the proposed risk contribution ratio measures of two different bivariate random vectors with the same or different copulas. Sufficient conditions are established in terms of stochastic orders, dependence structures, distortion functions and stress levels. We also study the ordering behavior of these measures on the interaction between paired risks. Numerical examples are also presented as illustrations.

E0704: **Double boosting of mean and dispersion in Tweedie's compound Poisson model with pre-defined base learners**
Presenter: **Guangyuan Gao**, Renmin University of China, China

Tweedie's compound Poisson model is a widely used method for predicting insurance loss. It is often necessary to model both mean and dispersion of insurance loss in Tweedie's compound Poisson model under the framework of double generalized linear models. However, the double generalized linear model is restricted to the linearity of covariates, which requires deliberate feature engineering. We propose a double boosting for joint modelling both mean and dispersion. Most boosting algorithms cannot facilitate random effects or spatial effects which often appear in insurance loss prediction. Thus, in the double boosting, we pre-define suitable base learners for different types of covariates. We conduct simulated data analysis and a real data analysis to illustrate the proposed method.

E0711: **Credibility theory for mean-variance premium principles**
Presenter: **Yaodi Yong**, The University of Hong Kong, Hong Kong
Co-authors: Yiying Zhang

In the credibility theory, hypothetical mean and process variance are two quantities that convey crucial information to decision-makers when determining premiums. Enlightened by the mean-variance premium principle, we propose a credibility approach to estimate both hypothetical and process variance at one time. The proposed estimator consists of linear observations and their quadratic terms. Several numerical illustrations are carried out to show the performance of the estimator. Meanwhile, a spin-off result is found and utilized to compare with the Buhlmann model and the q-credibility model.

E0800: **Optimal reinsurance for multivariate risks**
Presenter: **Yinzhi Wang**, Southwestern University of Finance and Economics, China

Optimizing reinsurance contracts is a big topic of study in the field of actuarial science from both theoretical and practical perspectives. Actuarial literature contains countless formulations and analytical results of what optimal reinsurance should mean for a single risk, but there is limited research on the optimal solution when the cedent runs many lines of business and asks to manage the risk effectively. We extend the problem of optimal reinsurance to a multivariate framework where the cedent has multiple risks which cannot be bundled together into one. More specifically, we have chosen to solve the problem by using layer contracts and a more industrial-based criterion, which is to balance risk and profit through a ratio where a risk measure is divided by expected surplus. Analytical results regarding the solution for the optimal parameters of multivariate risks are given in the paper. They suggest that in the bivariate case, with the expected premium principle, the solution is either balanced, with equal upper limits of the layers, or completely unbalanced, with one finite upper limit and one infinite one, corresponding to a stop-loss contract, depending on the marginal loss distributions. An extensive simulation study is also performed to confirm the analytical results, and extend them, in particular for a more general and realistic premium principle.

EO029 Room Virtual R6 Spatial data modeling: Theory and applications Chair: Pavel Krupskiy

E0371: **Mixed domain asymptotics for geostatistical processes**
Presenter: **Tingjin Chu**, University of Melbourne, Australia

Geostatistics is one of the three main branches of spatial statistics, with the maximum likelihood method being widely used for parameter estimation. The asymptotic properties of maximum likelihood estimators are often considered under the increasing domain asymptotic framework or the infill asymptotic framework. A third framework, the mixed domain asymptotic framework, has the advantage of incorporating both local and global properties of the covariance structure. We establish the asymptotic properties of maximum likelihood estimators under the mixed domain asymptotic framework. In addition to the asymptotic framework, the sampling design and the form of the covariance functions are also important factors for the asymptotic properties of maximum likelihood estimators. General conditions are imposed to ensure the consistency and asymptotic normality of these estimators. The imposed conditions are verified for some commonly used covariance functions. The resulting asymptotics provides novel insights into the convergence rates of parameter estimators under mixed domain asymptotics, as well as some useful guidelines for data analysis in practice.

E0484: **Landslide forecast by time series modelling and analytics of high-dimensional and non-stationary ground motion data**
Presenter: **Guoqi Qian**, The University of Melbourne, Australia
Co-authors: Antoinette Tordesillas, Hangfei Zheng

High-dimensional, non-stationary vector time-series data are often seen in ground motion monitoring of geo-hazard events. For timely and reliable forecasts from them, we developed a new statistical approach based on error-correction cointegration (ECC), vector autoregression (VAR), and a newly developed dimension reduction technique named empirical dynamic quantiles (EDQ). Our ECC-VAR-EDQ method was born by analyzing a big landslide dataset, comprising interferometric synthetic-aperture radar (InSAR) measurements of ground displacement that were observed at 5000+ time states and several thousand locations on a slope. The aim was to develop an early warning system for reliably forecasting any impending slope failure whenever a precursory slope deformation is on the horizon. Specifically, we first reduced the spatial dimension of the observed landslide data by representing them as a small set of EDQ series with negligible loss of information. We then used the ECC-VAR model to optimally fit these EDQ series, from which forecasts of future ground motion can be efficiently computed. Moreover, our method is able to assess the future landslide risk by computing the relevant probability of ground motion to exceed a red-alert threshold level at each future time state and location. Applying the ECC-VAR-EDQ method to the motivating landslide data gives a prediction of the incoming slope failure more than 8 days in advance.

E0557: **Parallel approximations of the Tukey g-and-h likelihoods and predictions for non-Gaussian geostatistics**
Presenter: **Sagnik Mondal**, King Abdullah University of Science and Technology, Saudi Arabia
Co-authors: Sameh Abdulah, Hatem Ltaief, Ying Sun, Marc Genton, David Keyes

Gaussian random fields are among the most popular models to describe spatial data. However, the assumption of Gaussianity in real data is unrealistic since data may show signs of skewness and heavy tails. We consider the Tukey g-and-h (TGH) non-Gaussian random field that shows more robustness in modeling spatial data by including two parameters to incorporate skewness and heavy tail features. This modeling process involves generating a dense symmetric positive definite matrix with $O(n^2)$ space complexity and $O(n^3)$ operational complexity, where n represents the number of spatial locations. On a large scale, this modeling process becomes prohibitive with standard methods. This work provides a parallel high-performance implementation of the TGH random field's inference on state-of-the-art hardware architectures. The implementation permits running the exact non-Gaussian modeling process for a large number of geospatial locations. We also provide a Tile Low-Rank approximation implementation that can accelerate the execution compared to the exact solution by around 7.29X and 2.96X on shared memory and distributed memory systems, respectively, using up to 810K spatial locations.

E0731: **Spatio-temporal Markov regime-switching models based on copulas**
Presenter: **Bouchra Nasri**, University of Montreal, Canada

The aim is to develop copula-based Spatio-temporal Markov regime-switching models with covariates when the variables of interest are continuous, discrete, or zero-inflated. A simulation study to assess the performance of the models is presented. A case study using infectious diseases data and environmental data is used to show the usefulness of the methodology.

EO181 Room Virtual R7 Recent advances in time series econometrics Chair: Jihyun Kim

E0196: **Dynamic factor model for functional time series: Identification, estimation, and prediction**
Presenter: **Nazarii Salish**, University Carlos III de Madrid, Spain
Co-authors: Sven Otto

A functional dynamic factor model for time-dependent functional data is proposed. We decompose a functional time series into a predictive low-dimensional common component consisting of a finite number of factors and an infinite-dimensional idiosyncratic component that has no predictive power. The conditions under which all model parameters, including the number of factors, become identifiable are discussed in detail. The identification results lead to a simple-to-use two-stage estimation procedure based on functional principal components. As part of our estimation procedure, we solve the separation problem between the common functional component and the idiosyncratic functional component. In particular, we obtain a consistent information criterion that provides joint estimates of the number of factors and dynamic lags of the common component. Finally, we illustrate the applicability of our method in a simulation study and to the problem of modeling and predicting yield curves. In an out-of-sample experiment, we demonstrate that our model performs well compared to the widely used term structure Nelson-Siegel model for yield curves.

E0289: **Cointegration in large VARs**
Presenter: **Anna Bykhovskaya**, University of Wisconsin - Madison, United States
Co-authors: Vadim Gorin

The purpose is to analyse cointegration in vector autoregressive processes (VARs) for the cases when both the number of coordinates N and the number of time periods T are large and of the same order. We propose a way to examine a VAR for the presence of cointegration based on a modification of the Johansen likelihood ratio test. The advantage of our procedure over the original Johansen test and its finite sample corrections is that our test does not suffer from over-rejection. This is achieved through novel asymptotic theorems for eigenvalues of matrices in the test statistic in the regime of proportionally growing N and T. The theoretical findings are supported by Monte Carlo simulations and an empirical illustration.

E0438: **Option-implied forecasts with robust change of measure**
Presenter: **Mamiko Yamashita**, Toulouse School of Economics, France

Option prices can be useful in forecasting since it has forward-looking information, yet the challenge lies in how to transform the risk-neutral measure implied by options to the physical one. Whereas existing papers overcome this issue by assuming an equilibrium model, we introduce a novel approach that does not require any model assumptions. Following the robustness literature, we take the risk-neutral measure as the reference model, and consider a set of measures that are "close" to the risk-neutral measure in the relative entropy sense. Then the robust forecast involves minimizing the maximum risk that corresponds to the nature of the forecast in question over this set. We also provide three indicators on the relative entropy bound, which is needed to be determined by a forecaster in our approach, by exploiting the theoretical connection between the risk-neutral and physical measures. In an empirical application, we compute 5% value-at-risk and expected shortfall on S&P500 Index returns. In our sample, we document that the robust value-at-risk is about 4%/12%/17% (for 1-day/10-day/22-day horizons) higher, and the robust expected shortfall is about 5%/15%/24% (for 1-day/10-day/22-day horizons) higher than the risk-neutral counterparts.

E0473: **Asymptotics of functional principal component analysis with weakly dependent data**
Presenter: **Bo Hu**, Peking University, China
The asymptotic theory is developed for principal component analysis for weakly dependent functional data in a separable Hilbert space. We establish the weak convergences of the sample variance operators for various types of weakly dependent functional data. With a version of the functional central limit theorem on Banach spaces we develop, the limiting distributions of the principal values and principal factors are shown to be normal.

EO269 Room Virtual R8 RECENT ADVANCES IN TIME SERIES ANALYSIS Chair: Kaiji Motegi

E0695: **Multifrequency-band tests for white noise under heteroscedasticity**
Presenter: **Ke Zhu**, University of Hong Kong, Hong Kong
A new family of multifrequency-band tests is proposed for the white noise hypothesis by using the maximum overlap discrete wavelet packet transform. At each scale, the proposed multifrequency-band test has the chi-square asymptotic null distribution under mild conditions, which allows the data to be heteroscedastic. Moreover, an automatic multifrequency-band test is further proposed by using a data-driven method to select the scale, and its asymptotic null distribution is chi-square with one degree of freedom. Both multifrequency-band and automatic multifrequency-band tests are shown to have the desired size and power performance by simulation studies, and their usefulness is further illustrated by two applications. As an extension, similar tests are given to check the adequacy of linear time series regression models, based on the unobserved model residuals.

E0913: **State-space method for the quadratic estimator of integrated variance in the presence of market microstructure noise**
Presenter: **Daisuke Nagakura**, Keio University, Japan
Co-authors: Toshiaki Watanabe
Recently, a class of integrated variance (IV) estimators, called the quadratic estimator (QE), has been considered which takes a quadratic form in observed returns. The QE includes several existing IV estimators as special cases, such as the realized variance, two-time scale estimator, and realized kernels. Even in the presence of market microstructure noises (MMNs) in observed prices, some special cases of the QE are consistent. However, they still have finite sample biases due to MMNs. We derive a multivariate state-space representation of QEs, where the IV is a common component in these QEs. Applying the Kalman filter to the state-space representation provides the linear projection of IV on these QEs which has the smallest MSE of any QEs employed in the state space representation. We conduct some simulation studies to check the performance of our method, and then do some empirical analysis, applying our method to the actual data.

E0186: **Regional interdependence of the Japan REIT market: A heteroscedasticity-robust time series approach**
Presenter: **Yoshitaka Iitsuka**, Kobe University, Japan
Co-authors: Kaiji Motegi
The aim is to investigate the dynamic interdependence between the stock returns of regionally disjoint Japanese real estate investment trusts (REITs), where the property type and a market return are controlled. We take a multivariate time series approach with the error term being allowed to have conditional heteroscedasticity of unknown form. We find significant spillover effects from central to local areas in conditional mean, a potential signal of arbitrage opportunities. The spillover effects have become stronger after the COVID-19 crisis for the office and hotel sectors, but not for the residential sector. This contrast suggests that a geographic diversification strategy within residential REIT securities should be more effective than that within an office or hotel, especially during a period of turmoil.

E0183: **Conditional threshold autoregression (CoTAR)**
Presenter: **Kaiji Motegi**, Kobe University, Japan
Co-authors: John Dennis, Shigeyuki Hamori
A new time series model is proposed where the threshold is specified as an empirical quantile of recent observations of a threshold variable. The resulting conditional threshold traces the fluctuation of the threshold variable, which can enhance the fit and interpretation of the model. In the proposed conditional threshold autoregressive (CoTAR) model, the existence of threshold effects can be tested by wild-bootstrap tests which incorporate all possible values of nuisance parameters. The estimation and hypothesis testing of the CoTAR model satisfy desired statistical properties in both large and small samples. We fit the CoTAR model to new confirmed COVID-19 cases in the U.S. and Japan. Significant conditional threshold effects are detected for both countries, and the implied persistence structures are consistent with the fact that the number of new confirmed cases in the U.S. is larger than in Japan.

EO257 Room Virtual R9 ESTIMATION FOR SEMI-PARAMETRIC MIXTURE MODEL Chair: Marie du Roy de Chaumaray

E0475: **Testing the order of multivariate normal mixture models**
Presenter: **Katsumi Shimotsu**, University of Tokyo, Japan
Co-authors: Hiroyuki Kasahara
Finite mixtures of multivariate normal distributions have been widely used in empirical applications in diverse fields such as statistical genetics and statistical finance. Testing the number of components in multivariate normal mixture models is a long-standing challenge even in the most important case of testing homogeneity. A likelihood-based test is developed for the null hypothesis of M components against the alternative hypothesis of $M+1$ components for a general $M \geq 1$. We derive the asymptotic distribution of the proposed EM test statistic under the null hypothesis and local alternatives and show the validity of the parametric bootstrap. The simulations show that the proposed test has a good finite sample size and power properties.

E0193: **Full model estimation for non-parametric multivariate finite mixture models**
Presenter: **Matthieu Marbac**, CREST - ENSAI, France
Co-authors: Marie du Roy de Chaumaray
The problem of full model estimation for non-parametric finite mixture models is addressed. An approach is presented for selecting the number of components and the subset of discriminative variables (i.e., the subset of variables having different distributions among the mixture components). The proposed approach considers a discretization of each variable into B bins and a penalization of the resulting log-likelihood. Considering that the number of bins tends to infinity as the sample size tends to infinity, we prove that our estimator of the model (number of components and subset

of relevant variables for clustering) is consistent under a suitable choice of the penalty term. The interest of our proposal is illustrated on simulated and benchmark data.

E0301: **Adaptive estimation of the nonparametric component under a two-class mixture model**
Presenter: **Gaelle Chagny**, CNRS, Universita de Rouen Normandie, France
Co-authors: Antoine Channarond, Van Ha Hoang, Angelina Roche

A two-class mixture model is considered, where the density of one of the components is known (equal to the uniform density on the interval $[0;1]$). This problem appears in many statistical settings, robust estimation and multiple testing among others. We address the issue of the nonparametric adaptive estimation of the unknown probability density of the second component. We propose a randomly weighted kernel estimator with a fully data-driven bandwidth selection method. Its definition involves empirical counterparts both for the mixture density and the mixing proportion: preliminary estimators for these quantities are also proposed. An oracle-type inequality for the pointwise quadratic risk is derived as well as convergence rates over Holder smoothness classes. The theoretical results are illustrated by numerical simulations.

E0180: **Semiparametric mixture of regression with unspecied error distributions**
Presenter: **Weixin Yao**, UC Riverside, United States

In the fitting of a mixture of linear regression models, the normal assumption has been traditionally used for the error term and then the regression parameters are estimated by the maximum likelihood estimate (MLE). Unlike the least squares estimate (LSE) for the linear regression model, the validity of the MLE for mixtures of regression depends on the normal assumption. In order to relax the strong parametric assumption about the error density, we propose a mixture of linear regression models with unknown error density. We prove the identifiability of our proposed model and provide the asymptotic properties of the proposed estimates. In addition, we will propose an EM-type algorithm that uses a kernel density estimator for the unknown error when calculating the classification probabilities in the E step. Using a Monte Carlo simulation study, we demonstrate that our method works comparably to the traditional MLE when the error is normal. In addition, we demonstrate the success of our new estimation procedure when the error is not normal. An empirical analysis of tone perception data is illustrated for the proposed methodology.

Saturday 04.06.2022 16:10 - 18:15 Parallel Session E – EcoSta2022

EV453 Room Virtual R5 CONTRIBUTIONS IN METHODOLOGICAL ECONOMETRICS Chair: Wai-keung Li

E0794: **The bias of the modified limited information maximum likelihood estimator in static simultaneous equation models**
Presenter: **Gareth Liu-Evans**, University of Liverpool, United Kingdom
Co-authors: Garry Phillips

The Modified LIML (MLIML) estimator has received a resurgence of interest recently in the context of weak instruments and many instruments. Like the original LIML estimator MLIML is consistent, and in the case of many instruments, it has been found asymptotically optimal. The MLIML estimator has all necessary moments and is unbiased to order $O(1/T)$, making it an important alternative to the 2SLS estimator. We find the bias of the MLIML estimator to order $O(1/T^2)$, and similarly, find the LIML estimator pseudo-bias to this order. The MLIML (and LIML) bias can be substantial, and different ways of correcting this are considered in Monte Carlo experiments. As an application of bias-corrected MLIML estimation, we re-estimate the effect of shifting the relative supply of young college workers on the US college graduate wage premium

E0799: **Proximal estimation and inference**
Presenter: **Alberto Quaini**, University of Geneva, Switzerland
Co-authors: Fabio Trojani

A unifying convex analysis framework characterizing the statistical properties of a large class of penalized estimators is built, both under a regular or irregular design. The framework interprets penalized estimators as proximal estimators, defined by a proximal operator applied to a corresponding initial estimator. We obtain new characterizations of the asymptotic properties of proximal estimators, showing that their asymptotic distribution follows a closed-form formula depending only on (i) the asymptotic distribution of the initial estimator, (ii) the estimator's limit penalty subgradient and (iii) the inner product defining the associated proximal operator. In parallel, we characterize the Oracle features of proximal estimators from the properties of their penalty subgradients. We exploit our approach to systematically cover linear regression settings with a regular, singular or nearly singular design. For these settings, we build new root-n consistent, asymptotically normal Ridgeless-type proximal estimators, which feature the Oracle property and are shown to perform satisfactorily in practically relevant Monte Carlo settings.

E0434: **The role of history in measurement**
Presenter: **Ioannis Paraskevopoulos**, Universidad Pontificia Comillas, Spain

The dependence and independence alternatives for a general evolution process in Banach and reflexive Banach spaces are investigated. We want to extend the limits of computation beyond Hilbert spaces. In particular, we examine whether the evolution process depends on its known history. We argue that multidimensional integration exists in reflexive Banach space if the solution obeys the functional central limit theorem and its error has limited variation, above from dilation and below from erosion boundaries. There exists a two-way mapping from reflexive Banach to Banach and back. The evolution process would depend on its history in 3/4 of all existing scenarios and this would necessarily imply that the integrable curves are quasi-periodic depending only on x_0, past initial conditions. One possible scenario will be that no model exists and reversibility will be unattainable as it would map to infinite possible initial points of the past. We offer two frameworks to evaluate these scenarios one is with a stochastic chain that builds on memory and another with a Deep learning Artificial Intelligent system armed with a non-linear operator in Banach $\mathcal{B}$. Either can capture all possibilities of dependence and independence in both Banach and the reflexive Banach spaces.

E1038: **Detecting many weak instruments**
Presenter: **Zhenhong Huang**, The University of Hong Kong, Hong Kong
Co-authors: Chen Wang, Jianfeng Yao

A new specification test is developed for the instrument weakness when the number of instruments and sample size goes to infinity proportionally. We proposed the test based on the fact that the asymptotic difference between the two-stage least squares (2SLS) estimator and OLS estimator disappears under the many weak instruments asymptotics, but converges to a non-zero limit under the alternative asymptotics. We establish the limiting distribution of the difference within two specifications and introduce a delete-d Jackknife procedure to consistently estimate the asymptotic variance/covariance. Monte Carlo experiments demonstrate the performance of the test procedure for both single and multiple endogenous variables. Additionally, we reexamine an analysis of returns to education by using our proposed test. Both the simulation results and empirical analysis indicate the reliability of the test.

EI005 Room 101 (Hybrid 1) RECENT DEVELOPMENTS IN ECONOMETRIC TIME SERIES Chair: Masayuki Hirukawa

E0157: **A robust approach to slope heterogeneity in linear models with interactive effects for large panel data**
Presenter: **Takashi Yamagata**, University of York, United Kingdom
Co-authors: Kazuhiko Hayakawa, Guowei Cui, Shuichi Nagata

A robust approach is proposed against heteroskedasticity, error serial correlation and slope heterogeneity in linear models with interactive effects for large panel data. First, consistency and asymptotic normality of the pooled iterated principal component (IPC) estimator for random coefficient and homogeneous slope models are established. Then, we prove the asymptotic validity of the associated Wald test for slope parameter restrictions based on the panel heteroskedasticity and autocorrelation consistent (PHAC) variance matrix estimator for both random coefficient and homogeneous slope models, which does not require the Newey-West type time-series parameter truncation. These results asymptotically justify the use of the same pooled IPC estimator and the PHAC standard error for both homogeneous-slope and heterogeneous-slope models. This robust approach can significantly reduce the model selection uncertainty for applied researchers. In addition, we propose a Lagrange Multiplier (LM) test for correlated random coefficients with covariates. This test has non-trivial power against correlated random coefficients, but not for random coefficients and homogeneous slopes. The LM test is important because the IPC estimator becomes inconsistent with correlated random coefficients. The finite sample evidence and an empirical application support the reliability and the usefulness of our robust approach.

E0158: **A machine learning attack on illegal trading**
Presenter: **Artem Prokhorov**, University of Sydney, Australia
Co-authors: Robert James, Henry Leung

An adaptive framework is designed for the detection of illegal trading behavior. Its key component is an extension of a pattern recognition tool, originating from the field of signal processing and adapted to modern electronic systems of securities trading. The new method combines the flexibility of dynamic time warping with contemporary approaches from extreme value theory to explore large-scale transaction data and accurately identify illegal trading patterns. Importantly, our method does not need access to any confirmed illegal transactions for training. We use a high-frequency order book dataset provided by an international investment firm to show that the method achieves remarkable improvements over alternative approaches in the identification of suspected illegal insider trading cases.

E1041: **Johansen test with Fourier-type smooth non-linear trends in cointegrating relations**
Presenter: **Mototsugu Shintani**, University of Tokyo, Japan
Co-authors: Takamitsu Kurita

The objective is to develop a methodology for testing cointegrating rank in vector autoregressive (VAR) models subject to Fourier-type smooth non-linear deterministic trends. A class of trigonometric functions is incorporated into VAR models in such a way that one can simultaneously examine non-linear and non-stationary characteristics of various types of time series data. Then, log-likelihood ratio test statistics for the selection of cointegrating rank are investigated, leading to the approximation of limit quantiles of the statistics by using simulation. A Monte Carlo analysis is also conducted, along with an empirical application to economic data, in order to demonstrate the usefulness of the proposed methodology in a practical context.

EO137 **Room 102 (Hybrid 2)** FINANCE AND MACROECONOMETRICS **Chair: Etsuro Shioji**

E0452: **Revisiting output convergence and economic growth determinants in OECD and some emerging countries**
Presenter: **Takashi Matsuki**, Osaka Gakuin University, Japan

The aim is to investigate the long-run convergence of per capita output across OECD members and some emerging countries from 1953 to 2019. To confirm the existence of output convergence toward reference countries, several unit root testing methods are employed. Moreover, to find possible growth determinants to promote convergence, some stationary covariates for the tests are used. In addition, the approach allows for the presence of endogenous structural breaks in the time series under investigation, to capture sharp drops in per capita outputs, which may be brought about by influential economic events such as serious economic slumps in domestic economies or the global financial crises. We also examine whether some institutional factors help to hold the convergence hypothesis.

E0424: **Term premiums and regime-switching prices of macro risks**
Presenter: **Sun Ho Lee**, Korea University, Korea, South
Co-authors: Kyu Ho Kang

A time-varying effect of macro factors on term premiums is investigated. We consider an arbitrage-free Nelson-Siegel term structure model of interest rates with possibly unspanned macro factors. In the model, the parameters in the prices of macro risks are assumed to change over time according to a first-order Markov regime-switching process. For regime identification and model choice, we classify two macro factors (real activity and inflation) into three categories: (i) ones with unspanned macro risk, (ii) ones without unspanned macro-risk, or (iii) ones with regime-switching unspanned macro risk. Given the three categories for the two variables, there are a total of 3^2 combinations. The models with different combinations are estimated and compared using a Bayesian approach. Based on the US monthly data from 1987 to 2008, the model with one regime-switching unspanned macro risk is most supported by the data. Specifically, the real activity contains unspanned information and it has a stronger impact on term premiums during recessions. We also report the results using the zero lower bound period data.

E0233: **Exchange rate pass-through under the unconventional monetary policy regime**
Presenter: **Yushi Yoshida**, Shiga University, Japan
Co-authors: Yuri Sasaki, Siyu Zhang, Weiyang Zhai

The structural VAR model is applied to Japan under the unconventional monetary policy regime, 2000Q1 and 2019Q4. In addition to the traditional sign restrictions, we impose narrative sign restrictions based on five phenomenal economic episodes. Estimated exchange rate pass-through induced by monetary policy shock or exogenous exchange rate shock is consistent with the conventional view, i.e., a Japanese yen depreciation induces inflation at the consumer level. On the other hand, we found evidence of perverse exchange rate pass-through induced by demand shock. A ten percent exchange rate depreciation driven by weak domestic demand is associated with a one percent deflation at the consumer level. The magnitude of the latter effect is greater than the former. This demand-shock-induced exchange rate pass-through effect may have undermined the continuous efforts of the Bank of Japan to achieve the target of a two percent inflation rate.

E0216: **When aggregate stock returns are negatively-skewed: International evidence**
Presenter: **Kyu Ho Kang**, Korea University, Korea, South
Co-authors: Kitak Kim

A novel stochastic volatility model with time-varying skewness is proposed. The skewness is modeled by a split-normal return error, and the asymmetric error variance is assumed to follow a first-order Markov-switching process. We show that this modeling approach enables us to simulate the SV via one-block Gibbs sampling and demonstrate that our posterior sampling algorithm is reliable and efficient in simulation studies. According to our empirical applications to several aggregate stock return data, the aggregate returns exhibit negative skewness during normal periods. Meanwhile, the stock returns are symmetric during market crash episodes.

E0300: **The pandemic and government bonds: Evidence from volatility smiles in Japan**
Presenter: **Etsuro Shioji**, Hitotsubashi University, Japan

The purpose is to study how the financial market has reacted to the aggressive fiscal and monetary policies that have been implemented since the outbreak of the COVID-19 in Japan. Even before the pandemic, the country's debt to GDP ratio was well over 200%. The situation has drastically worsened since February 2020. Prices of the JGB futures options are analyzed to measure the policies' influences on private-sector perceptions about the future course of the JGB market. To that end, we derive volatility smiles from those option prices on a daily basis. We study how the location and the shape of the smile curve have responded to the introduction of the new policy measures. Results show that the Bank of Japan's Yield Curve Control (YCC) policy has played a decisive role in stabilizing the JGB market.

EO225 **Room 103 (Hybrid 3)** RECENT ADVANCES IN SURVIVAL ANALYSIS **Chair: Byungtae Seo**

E0342: **Semiparametric accelerated failure time model with interval-censored data under outcome-dependent sampling design**
Presenter: **Tsui-Shan Lu**, National Taiwan Normal University, Taiwan

In epidemiological studies or clinical trials, it is always of interest to discuss the relationship between the failure time of a certain disease and other possible exposure measurements, while in the meantime the researchers are in favor of cost-effective study designs. We consider complete interval-censored data collected under an outcome-dependent sampling (ODS) design and the semiparametric accelerated failure time (AFT) model for parameter estimation, which directly links the failure time to the covariates through a log-linear model without specifying the error distribution. We conducted extensive simulation studies to assess the performance of the proposed estimators under various settings and applied our methods to a real data set derived from a coronary heart disease study from National Taiwan Hospital.

E0584: **Quantile residual life regression analysis of HIV/AIDS patients in Korea**
Presenter: **Soomin Kim**, Yonsei University, Korea, South

An HIV patient's residual lifetime is a major point of interest for both the patient and their physicians. While existing analyses on patient survival make forecasts based on data collected at the start of the study, residual lifetime analysis allows for a dynamic analysis based on added data collected up to a certain point in time. Since data on patient survival time shows a long-tailed distribution to the right, the median rather than the mean provides a more useful summary statistic of distribution. This study utilizes modeling of the quantile, including the median as a special case. Using data from the HIV/AIDS prospective cohort study in Korea, we propose statistical inference procedures that model the residual lifetime of HIV patients until they develop dyslipidemia. In this model, we model the quantiles of HIV patients remaining lifetime based on longitudinal biomarkers such as CD4 cells count, which is an important biomarker for HIV patients. To increase the computational efficiency in variance estimation, we propose an induced smoothing approach for the non-smooth estimating functions based on a check function. The proposed estimators are shown to have desirable asymptotic properties. Simulation experiments demonstrated that they perform reasonably well under finite samples.

E0653: **Imputation with verifiable identification condition for nonignorable missing outcomes**
Presenter: **Kenji Beppu**, Osaka University, Japan
Co-authors: Kosuke Morikawa, Jongho Im

Missing data often cause undesirable results such as bias and loss of efficiency. These results become more substantial problems when the response mechanism is nonignorable such that the response model depends on the unobserved variable. It is often required to estimate the joint distribution of the unobserved variable and response indicator to handle nonignorable nonresponse. However, model misspecification and identification issues prevent obtaining robust estimates even if we carefully estimate the target joint distribution. We model the distribution for the observed parts and derive sufficient conditions for the model identifiability, assuming a logistic distribution on the response mechanism and a generalized linear model as the main outcome model of interest. More importantly, the derived sufficient conditions are testable with the observed data and do not require any instrumental variables, which have been often assumed to guarantee the model identifiability but cannot be practically determined in advance. To analyze missing data, we propose a new fractional imputation method which incorporates the verifiable identifiability using observed data only. Furthermore, we present the performance of the proposed estimators in numerical studies and apply the proposed method to two sets of real data, the Opinion Poll for 2022 South Korean Presidential Election and public data collected from the US National Supported Work Evaluation Study.

E0775: **Accelerated failure time modelling via nonparametric mixtures**
Presenter: **Byungtae Seo**, Sungkyunkwan University, Korea, South
Co-authors: Sangwook Kang

An accelerated failure time (AFT) model assuming a log-linear relationship between failure time and a set of covariates can be either parametric or semiparametric depending on the distributional assumption for the error term. Both classes of AFT models have been popular in the analysis of censored failure time data. The semiparametric AFT model is more flexible and robust to departures from the distributional assumption than its parametric counterpart. The semiparametric AFT model, however, is subject to producing biased results for estimating any quantities involving an intercept. Meanwhile, parametric AFT models can be severely impaired by misspecifications. We propose a new type of AFT model using a nonparametric Gaussian scale mixture distribution. The proposed method can provide a consistent and robust estimator for the structural parameters in AFT models. The finite sample properties of the proposed estimators will also be presented via an extensive simulation study and a read data set.

E0906: **Fractional imputation approach for Cox regression with missing covariate**
Presenter: **Jongho Im**, Yonsei University, Korea, South
Co-authors: Taesuk Park, Sangwook Kang

In a case-cohort study, the main exposure variable is often only available for some subjects, while other covariates are available for the whole cohort. This incomplete data can be viewed as a special case of missing covariate by design. Previous works have used a popular multiple imputation approach to efficiently handle this missing data. Instead of multiple imputation, we can use fractional imputation as another repeated imputation approach. Fractional imputation is yet to be widely used in practice because it is relatively new and there is more complexity in variance estimation. However, fractional imputation has its own advantages, for example, it provides consistent variance estimation for the method-of-moment type estimators and creates a singly completed dataset rather than multiply completed datasets. A limited simulation study is implemented to confirm the performance of the proposed approach. In addition, misspecification in the imputation model is investigated to check the robustness of the proposed imputation method.

EO271 Room 104 (Hybrid 4) LEARNING AND MODELLING OF COMPLEX TIME SERIES AND SPATIAL PROCESSES Chair: Zudi Lu

E0423: **Robust inference on infinite and growing dimensional time series regression**
Presenter: **Abhimanyu Gupta**, University of Essex, United Kingdom
Co-authors: Myung Hwan Seo

A class of tests is developed for a growing number of restrictions in infinite and increasing order time series models such as multiple regression with growing dimension, infinite-order autoregression and nonparametric sieve regression. Examples include the Chow test, exponential tests, and testing of general linear restrictions of growing rank p. Notably, our tests introduce a new scale correction to the conventional quadratic forms that are recentered and normalized to account for diverging p. This correction accounts for a high-order long-run variance that emerges as p grows with sample size. We propose a bias correction via a null-imposed bootstrap to control finite sample bias without sacrificing power unduly. A simulation study stresses the importance of robustifying testing procedures against the high-order long-run variance even when p is moderate. The tests are illustrated with an application to oil regressions.

E0485: **Nonparametric maximum likelihood estimation for GINAR(p) models**
Presenter: **Taito Kihara**, Keio University, Japan

The nonparametric maximum likelihood estimator (NPMLE for short) for GINAR(p) models, which is based on generalized thinning operator, is defined and its asymptotic properties are discussed. An estimation method of semiparametric INAR(p) models has been previously proposed, however, estimation methods of semiparametric GINAR(p) models have not been addressed before. Previous work is extended to GINAR(p). Furthermore, semiparametric GINAR(p) models are more flexible than parametric GINAR(p) models when one analyzes actual integer-valued data. Particularly, it would be a useful tool to analyze data when underlying innovation distribution is multimodal or hard to detect the shape of the distribution function. We will show a numerical experiment with a GINAR(p) model which has multimodal innovation distribution as the data generating model to show the consistency of NPMLE which we can prove theoretically.

E0582: **Weighted estimation procedures for time-varying heavy-tailed processes**
Presenter: **Fumiya Akashi**, University of Tokyo, Japan
Co-authors: Junichi Hirukawa, Konstantinos Fokianos

A family of locally stationary processes is often useful when we model real data. Results for the finite variance case are extended to those of infinite variance cases. We consider a parameter estimation problem of autoregressive models with time-varying coefficients and propose a self-weighted local least absolute deviation regression estimator. The model is possibly finite or infinite variance and heteroscedastic one, and under mild conditions for the moment of the model, we show asymptotic normality of the proposed estimator. Some simulation experiments illustrate the finite sample performance of the proposed estimator.

E0632: **Adaptive group fused Lasso for panel threshold model with cross-sectional dependence**
Presenter: **Lulu Wang**, University of Southampton, United Kingdom
Co-authors: Zudi Lu

Panel threshold regression has been one of the most popular methods in nonlinear panel time series analysis. The most common method to determine the number of threshold parameters is using a bootstrap procedure to approximate the sampling distribution under the assumption of cross-sectional independence, which may work poor in dealing with the strong dependence on most climate and finance data. We consider estimation of panel threshold model where both regressors and residuals are allowed to be cross-sectional dependent via adaptive group fused Lasso. We show that with probability approaching one, the proposed method can correctly determine the unknown number of threshold parameters and estimate regression parameters consistently. We establish the asymptotic theories of the Lasso estimators of regression coefficients. Simulation studies demonstrated that the proposed estimation method works well in finite samples under cross-sectional dependent conditions. We finally apply our Lasso estimation method to study the effect of precipitation on the stocks in FTSE 100.

E0884: **On dynamic functional-coefficient autoregressive spatio-temporal models with irregular location wide nonstationarity**
Presenter: **Zudi Lu**, University of Southampton, United Kingdom
Co-authors: Xiaohang Ren, Rongmao Zhang

Nonlinear modelling of spatio-temporal data is often a challenge due to irregularly observed locations and location-wide non-stationarity. We propose a semiparametric family of Dynamic Functional-coefficient Autoregressive Spatio-Temporal (DyFAST) models to address the difficulties. First, we specify the dynamic autoregressive smooth coefficients depending on both a concerned regime and location so that the models can characterise not only the dynamic regime-switching nature but also the location-wide non-stationarity in real spatio-temporal data. Second, two semiparametric smoothing schemes are proposed to model the dynamic neighbouring-time interaction effects with irregular locations incorporated by (spatial) weight matrices. The first scheme popular in econometrics supposes that the weight matrix is pre-specified. In practice, many weight matrices can be generated differently by data location features. Model selection for an optimal one is popular but may suffer from a loss of features of different candidates. Our second scheme is thus to suggest a weight matrix fusion to let data combine or select the candidates. Accordingly, different semiparametric smoothing procedures are developed. Both theoretical properties and Monte Carlo simulations are investigated. The empirical application to an EU energy market dataset further demonstrates the usefulness of our DyFAST models.

EO377 **Room 105 (Hybrid 5)** **DATA ANALYTICS IN STATISTICS AND ECONOMETRICS** **Chair: Cy Sin**

E0258: **A greedy active learning algorithm in multinomial logistic regression**
Presenter: **Hsiang-Ling Hsu**, National University of Kaohsiung, Taiwan

For building a proper classification system to predict the class type of data, large amounts of labeled training samples are needed, which might result in lots of resources to obtain the effective labeled information. For this issue, we can adopt active learning to recruit the crucial data from a massive unlabeled data set, and then obtain its labeled information, finally put it into a labeled data set, which is utilized to construct the classifier. For a binary data analysis with the logistic regression models, the GATE algorithm not only considers the subject selection scheme but also integrates the variable extraction step to build the classifier more efficiently. An active learning procedure of multiple-class classification data has been previously proposed via individualized binary models for both categorical and ordinal labeled data. Moreover, for the active learning, it has been shown that selecting the initial samples effectively assists in building the classification model. We construct an active learning procedure that integrates the concepts of the initial samples determination, subject screening and variable selection simultaneously for applying multiple-class classification problems. Simulation studies and the analyzed results of open data sets demonstrate the classification performances for the proposed algorithm.

E0299: **Shrinkage estimations for social interactions models**
Presenter: **Hon Ho Kwok**, National Taiwan University, Taiwan

The aim is to give an in-depth analysis of shrinkage estimations for social interactions models. Identification and consistent estimations for social interactions models have been well developed. The statistical properties of the instrumental generalized method of moments (GMM) and maximum likelihood (ML) estimators were thoroughly analyzed in the past few decades. However, the properties of shrinkage estimators have yet been fully understood. A general class of linear and partially nonlinear social interactions models, and its shrinkage estimators, which are relatively computationally simple, are considered. The connection between consistency and shrinkage is developed. The estimation quality is rigorously investigated, for example, in terms of mean squared errors (MSE) and prediction errors. The mathematical conditions for acceptable quality are derived. Lastly, the theory developed would be extended to a decent model selection theory for social interactions models.

E0791: **No-regret forecasting with egalitarian committees**
Presenter: **Jiun-Hua Su**, Academia Sinica, Taiwan

The forecast combination puzzle is often found in literature: The equal-weight scheme tends to outperform sophisticated methods of combining individual forecasts. Exploiting this finding, we propose a Hedge Egalitarian Committees Algorithm (HECA), which can be implemented via mixed-integer quadratic programming. Specifically, egalitarian committees are formed by the ridge regression with shrinkage toward equal weights; subsequently, the forecasts provided by these committees are averaged by the hedge algorithm. We establish the no-regret property of HECA. Using data collected from the ECB Survey of Professional Forecasters, we find the superiority of HECA relative to the equal-weight scheme during the COVID-19 recession.

E0857: **Regularized estimation in dynamic panel with a multifactor error structure**
Presenter: **Shou-Yung Yin**, National Taipei University, Taiwan

The limiting behavior (consistency) of the regularized estimation in the dynamic panel with a multifactor structure is derived. Compared with the OLStype common correlated effects estimator (CCE), the regularized CCE approach does not need to assume that the sample size should be larger than the number of the variables of the approximated model. Therefore, consistency can be established under regularized CCE approach. The simulation results confirm that regularized CCE approach can outperform the OLS CCE approach and is robust even when p is relatively small. We also apply the proposed approach to investigate the determinants of economic growth in 157 countries covered from 1970 to 2019. The results show that after considering the regularization, the patterns of estimated coefficients are different in these countries compared to the results of OLS type estimation.

E0911: **Parameter estimation in a biomechanical model with multiplicative errors**
Presenter: **Wei-Ying Wu**, National Dong Hwa University, Taiwan

A biomechanical model often requires parameter estimation and selection in a known but complicated nonlinear function. Since for data from a head-neck position tracking system, one of the biomechanical models show multiplicative time-dependent errors, we develop a modified penalized least squares estimator that can handle such error structure. The method can be also applied to a model with non-zero mean time-dependent additive errors. Asymptotic properties of the proposed estimator are investigated. A simulation study demonstrates that the proposed estimation performs well in both parameter estimation and selection with temporally correlated error. The comparison with an existing method for head-neck position tracking data shows better performance of the proposed method in terms of the variance accounted for VAF.

EO391 Room 107 (Hybrid 7) RECENT ADVANCES IN SHRINKAGE ESTIMATION Chair: Yuzo Maruyama

E0338: **Bayesian shrinkage estimation for stratified count data**
Presenter: **Yasuyuki Hamura**, Kyoto University, Japan

The problem of simultaneously estimating Poisson parameters under the standardized squared error loss is considered in situations where side information in aggregated data is available. Bayesian shrinkage estimators are constructed using conjugate priors. The risk functions of estimators are compared, conditions for domination are obtained, and minimaxity and admissibility of a proposed estimator are proved. Several related problems are also discussed.

E0524: **An empirical Bayes approach to shrinkage estimation on the manifold of symmetric positive-definite matrices**
Presenter: **Chun-Hao Yang**, National Taiwan University, Taiwan
Co-authors: Hani Doss, Baba Vemuri

The James-Stein estimator is an estimator of the multivariate normal mean and dominates the maximum likelihood estimator (MLE) under squared error loss. The original work inspired great interest in developing shrinkage estimators for a variety of problems. Nonetheless, research on shrinkage estimation for manifold-valued data is scarce. We propose shrinkage estimators for the parameters of the Log-Normal distribution defined on the manifold of $N \times N$ symmetric positive-definite matrices. For this manifold, we choose the Log-Euclidean metric as its Riemannian metric since it is easy to compute and has been widely used in a variety of applications. By using the Log-Euclidean distance in the loss function, we derive a shrinkage estimator in an analytic form and show that it is asymptotically optimal within a large class of estimators that includes the MLE, which is the sample Frechet mean of the data. We demonstrate the performance of the proposed shrinkage estimator via several simulated data experiments. Additionally, we apply the shrinkage estimator to perform statistical inference in both diffusion and functional magnetic resonance imaging problems.

E0717: **Adapting to arbitrary quadratic loss via singular value shrinkage**
Presenter: **Takeru Matsuda**, RIKEN Center for Brain Science, Japan

The Gaussian sequence model is a canonical model in nonparametric estimation. We introduce a multivariate version of the Gaussian sequence model and investigate adaptive estimation over the multivariate Sobolev ellipsoids, where adaptation is not only to unknown smoothness but also to arbitrary quadratic loss. First, we derive an oracle inequality for the Efron-Morris singular value shrinkage estimator, which is a matrix generalization of the James-Stein estimator. Next, we develop an asymptotically minimax estimator on the multivariate Sobolev ellipsoid for each quadratic loss, which can be viewed as a generalization of Pinsker's theorem. Then, we show that the blockwise Efron-Morris estimator is exactly adaptive minimax over the multivariate Sobolev ellipsoids under any quadratic loss. It attains sharp adaptive estimation of any linear combination of the mean sequences.

E0841: **Weighted shrinkage estimators of normal mean matrices**
Presenter: **Ryota Yuasa**, The Institute of Statistical Mathematics, Japan
Co-authors: Tatsuya Kubokawa

In the estimation of the mean matrix in a multivariate normal distribution, the Efron-Morris estimator and the James-Stein estimator are two well-known minimax procedures, where the former is matricial shrinkage and the latter is scalar shrinkage. The methods for combining the two estimators with random weight functions are addressed. For deriving weight functions, we suggest the two methods. One is the minimization of a part of the unbiased estimator of the risk function, and the other is the empirical Bayes approach. The resulting weights are related to statistics for testing the sphericity of a covariance matrix. The resulting weighted shrinkage estimators are minimax. We also consider the case of an unknown covariance matrix. Numerical experiments are conducted to confirm the theoretical findings.

E0742: **Bayes extended estimators with shrinkage priors for multivariate normal models**
Presenter: **Michiko Okudo**, The University of Tokyo, Japan
Co-authors: Fumiyasu Komaki

The focus is on constructing predictive densities for multivariate normal models with unknown mean vectors. Bayesian predictive densities based on shrinkage priors often have complex representations and its computation requires approximation by taking the average of plugin densities. We approximate Bayesian predictive densities to reduce computational time and space by projecting them onto normal models with unknown mean and unknown covariance matrices, which include the original model as a subspace. We evaluate the Kullback-Leibler risk performance of the proposed methods, and compare them with those of Bayesian predictive densities with the uniform prior. Pythagorean relation of Bayesian predictive densities and its projection is also shown.

EO215 Room Virtual R1 BAYESIAN COMPUTATION FOR COMPLEX MODELS Chair: David Nott

E0164: **Efficient data augmentation techniques for state space models**
Presenter: **Siew Li Linda Tan**, National University of Singapore, Singapore

Data augmentation improves the convergence of iterative algorithms, such as the EM algorithm and Gibbs sampler by introducing carefully designed latent variables. We first propose a data augmentation scheme for the first-order autoregression plus noise model, where optimal values of working parameters introduced for recentering and rescaling of the latent states, can be derived analytically by minimizing the fraction of missing information in the EM algorithm. The proposed data augmentation scheme is then utilized to design efficient Markov chain Monte Carlo (MCMC) algorithms for Bayesian inference of some non-Gaussian and nonlinear state-space models, via a mixture of normals approximation coupled with a block-specific reparametrization strategy. Applications on simulated and benchmark real datasets indicate that the proposed MCMC sampler can yield improvements in simulation efficiency compared with centering, noncentering and even the ancillarity-sufficiency interweaving strategy.

E0240: **Variational inference for cutting feedback with misspecified models**
Presenter: **David Nott**, National University of Singapore, Singapore
Co-authors: Xuejun Yu, Michael Stanley Smith

Bayesian analyses combine information represented by different terms in a joint Bayesian model. When one or more of the terms is misspecified, it can be helpful to restrict the use of information from suspect model components to modify posterior inference. This is called "cutting feedback", and both the specification and computation of the posterior for such cut models is challenging. We formulate the cut posterior distributions as the solution to a constrained optimization problem, which naturally leads to optimisation-based variational computation methods. The proposed variational methods are faster than existing Markov chain Monte Carlo (MCMC) approaches for computing cut posterior distributions by an order of magnitude. It is also shown that variational methods allow for the evaluation of computationally intensive conflict checks that can be used to decide whether or not feedback should be cut. The methods are illustrated in a number of simulated and real examples, including an application where recent methodological advances that combine variational inference and MCMC within the variational optimization are used.

E0252: **Fast and accurate variational inference for models with many latent variables**

Presenter: **Michael Smith**, University of Melbourne, Australia

Co-authors: Ruben Loaiza-Maya, David Nott, Peter Danaher

Models with a large number of latent variables are often used to fully utilize the information in big or complex data. However, they can be difficult to estimate using standard approaches, and variational inference methods are a popular alternative. Key to the success of these is the selection of an approximation to the target density that is accurate, tractable and fast to calibrate using optimization methods. Most existing choices can be inaccurate or slow to calibrate when there are many latent variables. We propose a family of tractable variational approximations that are more accurate and faster to calibrate for this case. We derive a simplified expression for the re-parameterization gradient of the variational lower bound, which is the main ingredient of efficient optimization algorithms used to implement variational estimation. We illustrate using a random coefficients Tobit model applied to two million sales by 20,000 individuals in a large consumer panel from a marketing study. Last, we show how to implement data sub-sampling in variational inference for our approximation, which can lead to a further reduction in computation time.

E0448: **Population calibration using likelihood-free Bayesian inference**

Presenter: **Christopher Drovandi**, Queensland University of Technology, Australia

Co-authors: Brodie Lawson, Alexander Browning, Adrianne Jenner

A likelihood-free approach is developed for population calibration, which involves finding distributions of model parameters when fed through the model produces a set of outputs that matches available population data. Unlike most other approaches to population calibration, our method produces uncertainty quantification on the estimated distribution. Furthermore, the method can be applied to any population calibration problem, regardless of whether the model of interest is deterministic or stochastic, or whether the population data is observed with or without measurement error. We demonstrate the method on several examples, including one with real data. We also discuss the computational limitations of the approach. Immediate applications for the methodology developed here exist in many areas of medical research including cancer, COVID-19, drug development and cardiology.

E0847: **Flexible variational Bayes based on a copula of a mixture of normals**

Presenter: **Robert Kohn**, University of New South Wales, Australia

Co-authors: David Nott, David Gunawan

Variational Bayes methods approximate the posterior density by a family of tractable distributions and use optimisation to estimate the unknown parameters of the approximation. The variational approximation is useful when exact inference is intractable or very costly. A flexible variational approximation based on a copula of a mixture of normals is developed, which is implemented using the natural gradient and a variance reduction method. The efficacy of the approach is illustrated by using simulated and real datasets to approximate multimodal, skewed and heavy-tailed posterior distributions, including an application to Bayesian deep feedforward neural network regression models. Each example shows that the proposed variational approximation is much more accurate than the corresponding Gaussian copula and a mixture of normal variational approximations.

EO438 Room Virtual R2 RECENT ADVANCES IN MACHINE LEARNING Chair: Bharath Sriperumbudur

E0316: **When shape constraints meet kernel machines**

Presenter: **Zoltan Szabo**, LSE, United Kingdom

Co-authors: Pierre-Cyril Aubin-Frankowski

Shape constraints enable one to incorporate prior knowledge into predictive models in a principled way with various successful applications. Including this side information in a hard fashion (e.g, at every point of an interval) for rich function classes, however, is a quite challenging task. We will present a convex optimization framework to encode hard affine constraints on function values and derivatives in the flexible family of kernel machines. The efficiency of the approach is illustrated in joint quantile regression (analysis of aircraft departures).

E0317: **Fractal Gaussian networks: A sparse random graph model based on Gaussian multiplicative chaos**

Presenter: **Krishnakumar Balasubramanian**, University of California, Davis, United States

A novel stochastic network model, called Fractal Gaussian Network (FGN), is introduced that embodies well-defined and analytically tractable fractal structures. FGNs are driven by the latent spatial geometry of Gaussian Multiplicative Chaos (GMC), a canonical model of fractality in its own right from probability theory. FGNs interpolate continuously between the popular purely random geometric graphs (aka the Poisson Boolean network), and random graphs with increasingly fractal behavior. After introducing and motivating the model, we will discuss some probabilistic (e.g., expected motif counts, spectral properties) and statistical questions (e.g., detecting the presence of fractality and parameter estimation based on observed network data) related to FGNs, and present some preliminary real-world network data analysis.

E0318: **Measuring generalization with optimal transport**

Presenter: **Youssef Mroueh**, IBM Research, United States

New generalization bounds are described that reliably predict the generalization of deep neural networks. Although much progress has been made, theoretical error bounds still often behave disparately from empirical observations. We develop margin-based generalization bounds, where the margins are normalized with optimal transport costs between independent random subsets sampled from the training distribution. In particular, the optimal transport cost can be interpreted as a generalization of variance which captures the structural properties of the learned feature space. The bounds robustly predict the generalization error, given training data and network parameters, on large scale datasets. Theoretically, we demonstrate that the concentration and separation of features play crucial roles in generalization, supporting empirical results in the literature.

E0314: **Annealed flow transport Monte Carlo**

Presenter: **Michael Arbel**, INRIA Grenoble Rhone-Alpes, France

Annealed Importance Sampling (AIS) and its Sequential Monte Carlo (SMC) extensions are state-of-the-art methods for estimating normalizing constants of probability distributions. We propose here a novel Monte Carlo algorithm, Annealed Flow Transport (AFT), that builds upon AIS and SMC and combines them with normalizing flows (NFs) for improved performance. This method transports a set of particles using not only importance sampling (IS), Markov chain Monte Carlo (MCMC) and resampling steps - as in SMC, but also relies on NFs which are learned sequentially to push particles towards the successive annealed targets. We provide limit theorems for the resulting Monte Carlo estimates of the normalizing constant and expectations with respect to the target distribution. Additionally, we show that a continuous-time scaling limit of the population version of AFT is given by a Feynman–Kac measure which simplifies to the law of a controlled diffusion for expressive NFs. We demonstrate experimentally the benefits and limitations of our methodology on a variety of applications.

E0545: **Modern kernel methods for econometrics**

Presenter: **Krikamol Muandet**, Max Planck Institute for Intelligent Systems, Germany

While recent developments of kernel methods have led to numerous applications in machine learning and statistics, they have not been fully utilized to solve problems in econometrics. We will provide examples of how modern kernel methods can be employed to solve unique econometric problems ranging from conditional moment (CM) test and distributional treatment effect (DTE) estimation to an instrumental variable (IV) regression.

In addition, we will highlight the potential research directions that lie at the intersection of machine learning and economics.

EO169 Room Virtual R3 EXTREME VALUE ANALYSIS IN TIME AND SPACE Chair: Gilles Stupfler

E0477: **Long range dependence in the tails**
Presenter: **Marco Oesting**, University of Stuttgart, Germany

The presence or absence of long memory in time series is known to have major effects on the asymptotic properties of statistical estimators. While classical notions of long-range dependence typically rely on characteristics of the bulk of the distribution such as covariances, many common estimators in extreme value statistics are based on observations in the tails only. Motivated by this fact, we propose to separately study long memory in the extremes as given by the tail process of a regularly varying times series and the corresponding max-stable analogue. Based on a recent definition of long-range dependence that is invariant under marginal transformations, we revisit a necessary and sufficient criterion for long-range dependence of max-stable time series in terms of the pairwise extremal coefficient function. We present statistical applications of this characterization and show its effect on limit theorems which turns out to be similar to classical definitions. Furthermore, we discuss the extension of the results to processes in the max-domain of attraction.

E0636: **Heavy-tailed extremile regression in risky seismic areas**
Presenter: **Abdelaati Daouia**, Fondation Jean-Jacques Laffont, France
Co-authors: Thibault Laurent, Gilles Stupfler

Extremile regression defines a least-squares analog of quantile regression as is the case in the duality between the conditional mean and median. The use of extremiles appears naturally in risk mitigation where they enjoy various intuitive meanings in terms of weighted moments rather than tail probabilities. They account for the magnitude of infrequent events and not only for their relative frequency. They belong to both classes of concave distortion risk measures and coherent spectral risk measures of law–invariant type. We study their implications for estimating and inferring tail risk, focusing on heavy-tailed seismic distributions in risky areas. Based on a localized Hill estimator of the conditional tail index, we present an extrapolated estimator for high conditional extremiles and derive its asymptotic normality under mild conditions. This extremile estimator shows an excellent performance in simulations compared with the existing competitors. On an earthquake dataset, it suggests a more reasonable and prudent differentiation of the severity of massive earthquakes geographically compared to the traditional Value at Risk and Tail Conditional Mean.

E0675: **Asymptotic expansions for blocks estimators of cluster indices**
Presenter: **Rafal Kulik**, University of Ottawa, Canada

Cluster indices describe extremal behaviour of stationary time series. We consider their disjoint and sliding blocks estimators. Using a modern theory of multivariate, regularly varying time series, we obtain a sharp asymptotic expansion on the difference between these two types of estimators. As a consequence, we show that in the Peaks-Over-Threshold framework, sliding and disjoint blocks estimators have the same limiting variance.

E0798: **Tail risk inference via expectiles in heavy-tailed time series**
Presenter: **Simone Padoan**, Bocconi University, Italy
Co-authors: Gilles Stupfler, Anthony Davison

Expectiles define the only law-invariant, coherent and elicitable risk measure apart from the expectation. The popularity of expectile-based risk measures is steadily growing and their properties have been studied for independent data, but further results are needed to establish that extreme expectiles can be applied with the kind of dependent time series relevant to financial data modelling. We provide a basis for inference on extreme expectiles and expectile-based marginal expected shortfall in a general β-mixing context that encompasses ARMA and GARCH models with heavy-tailed innovations. Our methods allow the estimation of marginal (pertaining to the stationary distribution) and dynamic (conditional on the past) extreme expectile-based risk measures. Simulations and applications to financial returns show that the new estimators and confidence intervals greatly improve on existing ones when the data are dependent.

EO189 Room Virtual R4 BAYESIAN METHODS IN ECONOMICS Chair: Veronica Ballerini

E0605: **Forecasting cryptocurrencies log-returns: A Bayesian approach using social media sentiment indexes**
Presenter: **Federico DAmario**, Sapienza University of Rome, Italy
Co-authors: Milos Ciganovic

Academics are increasingly acknowledging the contribution of social media information to make predictions in many areas, particularly in financial markets and economics. We leverage the predictive power of Twitter and Reddit sentiment together with Google Trends indexes to forecast the log-returns of ten cryptocurrencies divided into three tiers according to their market capitalization. We evaluate the performance of three Bayesian VARs specified with hierarchical shrinkage priors using daily data from November 2017 to January 2022. We perform a four-step-ahead forecast and we find a significant improvement in mean directional accuracy compared with some state-of-the-art forecasting models.

E0651: **Bayesian principal stratification with longitudinal data and truncation by death**
Presenter: **Giulio Grossi**, University of Florence, Italy
Co-authors: Marco Mariani, Alessandra Mattei, Fabrizia Mealli

In many causal studies, outcomes are censored by death, in the sense that they are neither observed nor defined for units who die. In such studies, the focus is usually on the stratum of 'always survivors' up to a single fixed time s. Building on a recent strand of the literature, we propose an extended framework for the analysis of longitudinal studies, where units can die at different time points, and the main endpoints are observed and well-defined only up to the death time. We develop a Bayesian longitudinal principal stratification framework, where units are cross-classified according to the longitudinal death status. Under this framework, the focus is on causal effects for the principal strata of units that would be alive up to a time point s irrespective of their treatment assignment, where these strata may vary as a function of s. We can get precious insights into the effects of treatment by inspecting the distribution of baseline characteristics within each longitudinal principal stratum, and by investigating the time trend of both principal stratum membership and survivor-average causal effects. We illustrate our approach for the analysis of a longitudinal observational study aimed to assess, under the assumption of strong ignorability of treatment assignment, the causal effects of a policy promoting start-ups on firms' survival and hiring policy, where firms' hiring status is censored by death.

E0667: **Estimating causal effects of community health financing via principal stratification**
Presenter: **Silvia Noirjean**, University of Florence, Italy
Co-authors: Mario Biggeri, Laura Forastiere, Fabrizia Mealli, Maria Nannini

In applied economics, the common empirical strategy for analyzing experimental data with noncompliance is to use the Instrumental Variables method. When the effects are heterogeneous, this method allows, under certain assumptions, to identify the causal effect for Compliers, i.e., the subset of units whose treatment is affected by the assignment. One of these assumptions is the Exclusion Restriction (ER), which precludes the possibility of a causal effect for Never Takers, i.e., those whose treatment is unaffected by the assignment. We show the consequences of violations of this assumption in the impact evaluation of intervention of Community Health Financing (CHF), where households were randomly assigned to attend sensitization sessions and then receive the offer to join a CHF scheme. The analyses are performed using Bayesian Principal Stratification by first assuming and then relaxing the ER for Never Takers. This allows showing the positive impact of the intervention on the health costs of both Compliers and Never Takers. The causal effects for the former could be due to the sensitization but also to the actual participation in the scheme; those for the latter are attributable to the sensitization only and, by stating the ER, would have been assumed not to exist.

E0858: **Modelling preferences via Wallenius process**
Presenter: **Rosario Barone**, University of Rome Tor Vergata, Italy
Co-authors: Veronica Ballerini, Brunero Liseo

The volume of sales in a local real estate market is well known to be subject to fluctuations that depend on quotations, that depend on socioeconomic macro variables in turn. However, there exist micro and less volatile determinants of such volume of sales, that are based on the aggregate individuals' preferences. To disregard such determinants would imply returning biased predictions on the relative volume of sales for different municipality areas, conditioned on the observed quotations. In fact, buyers' preferences with respect to different zones of the local real estate market can be estimated assuming that the periodical sales volumes follow a Wallenius noncentral hypergeometric distribution (WNC), and that a sequence of WNC generates a newly defined stochastic process, i.e., a Wallenius process (WP). WNC describes a biased urn problem in which the probability to sample a certain number of colored balls depends not only on the number of balls of that color in the urn, but also on the weight associated with each color. Given the intractability of the likelihood function, the inference is performed via Approximate Bayesian computation (ABC) methods.

EO329 Room Virtual R7 STATISTICS TO IMPROVE THE DEVELOPMENT OF CULTIVARS Chair: Reka Howard

E0837: **Future-oriented strategy via simulations optimizes breeding schemes with selection indices**
Presenter: **Kosuke Hamazaki**, The University of Tokyo, Japan
Co-authors: Hiroyoshi Iwata

In recent years, genomic selection using prediction values based on genomic prediction models has been contributing to more efficient and rapid breeding. Although various models have been proposed to improve the selection accuracy, in breeding programs, it is known that the decision for selection and crossing based on the genomic prediction has a greater impact on the final genetic gain than the accuracy of the models themselves. This study proposes a framework to optimize decision-making in breeding programs by utilizing numerical optimization approaches. We focused on the optimal mating combination of parental candidates in each generation, including the allocation of progenies for crosses, and parameterized it based on a soft-max function that combines multiple selection indices. To improve genetic gain while maintaining the genetic diversity of the breeding population, predicted breeding values and genetic diversity of the progenies in a subsequent generation were used as indices. We then proceeded with simulation-based breeding, giving a parameter to weight these indices, and optimized the parameters using a numerical optimization algorithm called StoSOO. The results showed that the breeding conducted under the scheme optimized based on the proposed method showed a higher genetic gain in the final generation compared to the non-optimized breeding.

E0840: **Modelling soybean growth: A nonlinear mixed model approach**
Presenter: **Maud Delattre**, INRAE, France
Co-authors: Hiroyoshi Iwata, Jessica Tressou

Field experiments on soybean were conducted in Arid Land Research Center, Tottori, Japan, under several experimental conditions. The growth was monitored by a drone measuring each day the plant height of about 200 soybean varieties for which whole-genome sequence data are also available. Based on these data, the objective is to propose an original statistical approach to refine the understanding of the determinants of soybean growth and improve the prediction of phenotypic traits of interest. The problem is formalized through a nonlinear mixed-effects model in which random effects allow modeling of genetic and environmental effects and their variability. Parameter estimation in nonlinear mixed models is however not straightforward, especially due to the model's nonlinearity and the random effects. SAEM (Stochastic Approximation of the Expectation-Maximization algorithm) is widely used in this context, but it is rarely used in plant biology. The originality compared to standard mixed-effects models is that the soybean model integrates the relationships between varieties through the kinship matrix, which requires an adaptation of the algorithm. SAEM is implemented and predictions of expected growth curves per variety can then be deduced by a maximum a posteriori approach. The methodology is applied to the experimental data from Tottori.

E0873: **Genome-enabled analysis of time-series high-throughput phenotyping data**
Presenter: **Gota Morota**, Virginia Polytechnic Institute and State University, United States

The advent of plant phenomics, coupled with the wealth of genotypic data generated by next-generation sequencing technologies, provides exciting new resources for studies of complex traits. However, these new technologies also bring new challenges to quantitative genetics, namely, a need to develop robust frameworks that can accommodate these high-dimensional data for genomic prediction and genome-wide association studies. One unique aspect of high-throughput phenotyping data is that phenomics platforms often produce large-scale data with high temporal resolution. We developed a random regression model framework for modeling trait trajectories by accounting for covariances across timepoints to accommodate time-series measurements in genome-enabled analysis. The random regression model has recently been extended to a Bayesian random regression marker effect model that can incorporate mixture priors to marker effects to introduce more meaningful biological assumptions for longitudinal trait analysis. We demonstrate the utility of the random regression model and random regression marker effect model using both simulated and real rice data.

E0891: **Prediction of flowering and maturity time of soybean using stacking**
Presenter: **Akio Onogi**, Ryukoku University, Japan

In crop breeding, it is important to control phenological traits according to target regions because phenological traits are related to local adaptation. Typical phenological traits, flowering and maturity times, are known to be affected by genetic and environmental factors such as temperature. To predict the flowering and maturity time of new cultivars accurately, an ensemble learning, stacking, was applied to soybean data and compared with other methods. The explanatory variables included daily mean, maximum, and minimum temperature, precipitation, hours of sunshine, and day length as environmental factors, and genotypes of five genes relevant to flowering. The response variables were days from sowing to flowering for flowering time, and days from flowering to maturity for maturity time. A total of 41 learners including random forests, cubist, gradient boosting, support vector machine, and elastic net were used as base models of stacking. Random forests and linear regression were compared as the meta-model. Besides stacking, each base model and method based on an eco-physiological model of crop phenology were also compared. The evaluation using independent data shows the superiority of stacking that used random forests as the meta-model among the methods compared.

E0803: **Sparse classification with multi-type data**
Presenter: **Reka Howard**, University of Nebraska - Lincoln, United States
Co-authors: Vamsi Manthena, Diego Jarquin

Genomic selection is a technique in plant breeding that implements a model for predicting phenotypes using marker information without the need of testing the individuals in fields thus saving resources. Since genomic selection was first introduced many statistical methods have been proposed, and not only marker information was used for prediction but also other data types (high-throughput phenotyping, pedigree, weather, etc.). Integrating different data types becomes a complex challenge when they have very different dimensions. A key challenge is to build models that are able to access the unique information present in each data type in order to improve the prediction capabilities. Breeders are often interested in categorical phenotypic traits such as resistance to drought or salinity, susceptibility to disease, and days to maturity or flowering. While there is extensive literature covering the prediction of continuous traits, there is limited literature developing genomic prediction models for classification. We present a classification method where we integrate three data types - secondary traits, weather, and genomic information for classification. We compared our method to two standard classifiers such as random forests and SVMs. The proposed three-stage method allows us to access the information present in each data type to improve prediction.

EO251 Room Virtual R8 ADVANCED STATISTICAL METHODS IN ECONOMICS AND FINANCE Chair: Huei-Wen Teng

E0291: Testing monotonicity of mean potential outcomes in a continuous treatment

Presenter: **Yu-Chin Hsu**, Academia Sinica, Taiwan

While most treatment evaluations focus on binary interventions, a growing literature also considers continuously distributed treatments, e.g. hours spent in a training program to assess its effect on labor market outcomes. We propose a Cramer-von Mises-type test for testing whether the mean potential outcome given a specific treatment has a weakly monotonic relationship with the treatment dose under a weak unconfoundedness assumption. This appears interesting for testing shape restrictions, e.g. whether increasing the treatment dose always has a non-negative effect, no matter what the baseline level of treatment is. We formally show that the proposed test controls the asymptotic size and is consistent against any fixed alternative. These theoretical findings are supported by the methods of finite sample behavior in our Monte-Carlo simulations. As an empirical illustration, we apply our test to the Job Corps study and reject a weakly monotonic relationship between the treatment (hours in academic and vocational training) and labor market outcomes like earnings or employment.

E0305: Estimations of the conditional tail average treatment effect

Presenter: **Yu-Min Yen**, National Chengchi University, Taiwan

The aim is to study estimation of the conditional tail average treatment effect (CTATE), defined as a difference between conditional tail expectations of potential outcomes. The CTATE can capture heterogeneity and deliver aggregated local information of treatment effects over different quantile levels and is closely related to the notion of second-order stochastic dominance and the Lorenz curve. These properties render it a valuable tool for policy evaluations. We consider a semiparametric treatment effect framework under endogeneity for the CTATE estimation using a newly introduced class of consistent loss functions jointly for the conditional tail expectation and quantile. We establish the asymptotic theory of our proposed CTATE estimator and provide an efficient algorithm for its implementation. We then apply the method to the evaluation of effects from participating in programs of the Job Training Partnership Act in the US.

E0923: Monte Carlo simulation and its applications

Presenter: **Yu-Ying Tzeng**, National Chengchi University, Taiwan

Monte Carlo simulation methods are widely used in finance such as pricing an option on equity, valuing interest rate derivatives and evaluating portfolios because this method can work well as the dimension increases. However, Monte Carlo simulation can be improved in two ways: quasi-Monte Carlo can improve the accuracy; Markov chain Monte Carlo can improve the applicability from a lower dimension to a higher dimension or complicated distribution. Hence, we can combine these two methods for risk measures.

E0972: Analysis of value-at-risk and expected shortfall under a jump-diffusion model with left-skewed jump sizes

Presenter: **Wei-Chung Miao**, National Taiwan University of Science and Technology, Taiwan

Co-authors: Xenos Chang-Shuo Lin, Having Yi-Ju Chien

A jump-diffusion model is proposed which incorporates the left skewedjump size distribution and discuss the effects of the left skewness in jump sizes on the two major risk measures: Value-at-Risk (VaR) and Expected Shortfall (ES). The jump size distribution is described by a shifted gamma (SG) distribution and our proposed jump-diffusion model (termed SGJD model) can be seen as an extended version of the classical jump-diffusion model. We provide mathematical analysis of the proposed model and derive analytical formulas for the two risk measures under our model. Since a new parameter is introduced to capture jump size skewness and Merton's classical model is actually a limiting case with skewness parameter approaching 0, our numerical analysis examines how the return distributions and VaR/ES vary as the skewness parameter deviates from 0. Our results show that the skewness parameter plays a significant role in VaR and ES, particularly when the time interval is small and confidence level is high. These observations justify the incorporation of the left skewed jump size and provide supports for the proposed SGJD model when the far left end of the return distribution is concerned.

E1009: Quantitative trading of vertical spread option strategies with stop-loss by machine learning

Presenter: **Min-Kuan Chen**, National Taipei University of Technology, Taiwan

Co-authors: Mu-En Wu, Wen-Shuen Wu

In recent years, quantitative trading with AI techniques has been developed in finance research and applications. In prior works, quantitative trading studies forecasted the underlying asset dynamics and evaluated the expected value. However, the win rate is hard to explore and predict. We leave the odds by vertical spread option strategies to address the challenge, which could pre-lock profit and loss. Furthermore, we proposed a method to estimate the probability with the stop-loss mechanism as a win rate via a statistical approach and machine learning to improve the performance. The spread strategy position will be closed when the stop-loss is triggered. Otherwise, it will remain open until the option expires. Subsequently, we visualize the win rate by heat-map and select the profitable spread strategy at multiple moments and spreads. The results show that the accumulated profit and loss curve monotonically increases and enhances performance. This suggests that our method generates a promising approach and applies to practical program trading. Other AI techniques, such as neural networks and deep learning, may also predict win rate, as we have demonstrated here using machine learning.

EO395 Room Virtual R9 STATISTICS ON SHAPES AND MANIFOLDS Chair: Joern Schulz

E0221: Stochastic shape analysis and probabilistic geometric statistics

Presenter: **Stefan Sommer**, University of Copenhagen, Denmark

Analysis and statistics of shape variation can be formulated in geometric settings with geodesics modelling transitions between shapes. Extensions of these smooth geodesic models will be considered to account for noise and uncertainty: Stochastic shape processes and stochastic shape matching algorithms. In the stochastic setting, matching algorithms take the form of bridge simulation schemes which also provide approximations of the transition density of the stochastic shape processes. Examples of stochastic shape processes and connected bridge simulation algorithms will be covered. We will connect these ideas to statistics for data on general manifolds, particularly to the diffusion mean.

E0347: **Most probable paths for anisotropic Brownian motions on manifolds**
Presenter: **Erlend Grong**, University of Bergen, Norway
Co-authors: Stefan Sommer

The diffusion mean is an intrinsic way of considering the mean of a dataset located on a Riemannian manifold. It also allows considering general types of covariances through a Brownian motion that is not anisotropic. We describe this approach and how computing this means and covariance is related to the most probable paths on manifolds, which are different from geodesics.

E0627: **Statistical analysis of locally parameterized shapes**
Presenter: **Mohsen Taheri Shalmani**, University i Stavanger, Norway
Co-authors: Joern Schulz

The establishment of correspondence and defining shape representation are crucial steps in statistical shape analysis for detecting local dissimilarities between two groups of objects. Most shape representations are based on either noninvariant spatial properties to rigid transformation or extrinsic geometric properties. Shape analysis based on extrinsic geometric properties could be misleading, and based on noninvariant properties is biased because the act of alignment is necessary. Also, mathematical interpretation of the type of dissimilarity, e.g., bending, elongation, twisting, protrusion, etc., is desirable. By defining local coordinate systems on object skeletal, a novel shape representation based on intrinsic and invariant object properties will be discussed. The proposed shape representation is also superior for simulation and skeletal deformation. The power of the hypothesis testing based on the introduced shape representation is demonstrated in the simulated data as well as the left hippocampi of patients with Parkinson's disease versus a healthy control group.

E0697: **Geometric and statistical models of analyzing two-object complexes**
Presenter: **Zhiyuan Liu**, University of North Carolina at Chapel Hill, United States

The shape correlation of multi-object complexes in the human body is important for understanding the development of disease. The development of autism, for example, often changes the shapes of multiple brain structures. While there exist many statistical methods that can extract correlation from multi-block data, very little research can effectively extract intrinsic shape correlation. It is especially difficult to extract shape correlation when the involved objects have different variability in separate non-Euclidean spaces. Moreover, it is difficult to capture intrinsic shape information within and between objects for a joint analysis of multi-object complexes. Geometric and statistical models are presented that can extract shape correlation from two-object complexes. These models are designed to be insensitive to different variability of objects. Also, the results can be straightforwardly interpreted by researchers and clinical users.

EC428 **Room 106 (Hybrid 6)** **CONTRIBUTIONS IN FINANCIAL ECONOMETRICS (IN-PERSON)** **Chair: Yuta Koike**

E0722: **Shrinking in COMFORT**
Presenter: **Simon Hediger**, University of Zurich, Switzerland
Co-authors: Jeffrey Naef

Nonlinear shrinkage is combined with the Multivariate Generalized Hyperbolic (MGHyp) distribution to account for heavy tails in estimating the first and second moments in high dimensions. An Expectation-Maximization (EM) algorithm is developed that is fast, stable, and applicable in high dimensions. Theoretical arguments for the monotonicity of the proposed algorithm are provided and it is shown in simulations that it is able to accurately retrieve parameter estimates. Finally, in an extensive Markowitz portfolio optimization analysis, the approach is compared to state-of-the-art benchmark models. The proposed model excels with a strong out-of-sample portfolio performance combined with a comparably low turnover.

E0786: **Duality in optimal consumption-investment problems with alternative data**
Presenter: **Kexin Chen**, The Hong Kong Polytechnic University, Hong Kong
Co-authors: Hoi Ying Wong

An optimal consumption-investment problem is investigated when the expected return of a risky asset is modulated by a hidden Markov chain, representing different unobserved economic regimes. While the classical approach estimates the hidden state from historical asset prices, the technology nowadays enables investors to make decisions using alternative data as a complementary source of observations. Social media commentary, expert opinion, pandemic data, and GPS data are a part of "alternative data", that is, data that originate outside of the standard repertoire of market data but are considered useful for predicting stock trends. We model the asset price and alternative data series as a diffusion process and a marked point process, respectively. We, therefore, incorporate the alternative data into the filtering process that extends the Wonham filter to a degenerate jump diffusion with Lévy type jumps. This introduces a remarkable analytical challenge to the corresponding stochastic control problem. We resolve the difficulties with a novel duality approach. We link the dual problem to an optimization problem over a set of equivalent local martingale measures and devise a methodology to obtain the optimal solution with the alternative data filtering technique. We show that the dual problem admits a unique smooth solution for hyperbolic absolute risk aversion (HARA) utility functions. In addition, we obtain an explicit feedback on optimal consumption-investment strategy.

E0970: **On the construction of neural networks for value-at-risk forecasting**
Presenter: **Ye Chen**, The University of Sydney, Australia
Co-authors: Yi Jiang, Richard Gerlach

In the context of forecasting Value-at-Risk (VaR) using quantile regression models, much research effort has been devoted to specifying the functional forms of quantile dynamics that relate the present period VaR to a set of explanatory variables available at the previous period. Typical choices of predictors include past returns and summaries of past returns observed at a higher frequency. Neural networks aimed at minimising the quantile loss have been proposed as generalisations of quantile regression models and applied to produce VaR forecasts. However, little attention has been paid to analysing the architecture of quantile regression neural networks and their impact on VaR forecasts, especially when higher-frequency returns are provided as inputs. We empirically assess performance in forecasting daily VaR for various designs of feedforward and recurrent networks of different input types. We also compare the performance of neural network models with that of the more traditional CAViaR- and GARCH-type models.

E0948: **Factor analysis for heavy-tailed, heteroscedastic data**
Presenter: **Chang Yuan Li**, University of California, Santa Barbara, United States
Co-authors: Alexander Shkolnik

Factor analysis of financial asset returns aims to decompose a return covariance matrix into systematic and specific components. One or both of these components are believed to have considerably heavier tails than the Gaussian distribution. Traditional statistical approaches like PCA and MLE suffer from drawbacks: sensitivity to outliers, and strict assumptions on the underlying distributions. And so, these are often not suitable for the purpose of decomposing financial asset returns. We propose a convex optimization procedure to decompose a security return covariance into its low rank and diagonal parts. The diagonal parts, corrupted by outliers, can be improved by weighted ridge regression and outlier correction methods. By doing so, the low-rank estimate can be improved as well. We illustrate the results with some analytical examples as well as simulated and empirical models.

E0671: **On the uncertainty of a combined forecast: The critical role of correlation**

Presenter: **Jan Magnus**, Vrije Universiteit Amsterdam, Netherlands

The purpose is to show that the effect of the zero-correlation assumption in combining forecasts can be huge and that ignoring (positive) correlation can lead to confidence bands around the forecast combination that are much too narrow. In the typical case where three or more forecasts are combined, the estimated variance increases without a bound when correlation increases. Intuitively, this is because similar forecasts provide little information if we know that they are highly correlated. Although we concentrate on forecast combinations and confidence bands, our theory applies to any statistic where the observations are linearly combined. We apply our theoretical results to explain why forecasts by Central Banks (in our case, the Bank of Japan) are so frequently misleadingly precise. In most cases, a correlation above 0.7 is required to produce reasonable confidence bands.

Sunday 05.06.2022 08:00 - 10:05 Parallel Session F – EcoSta2022

EO417 Room 101 (Hybrid 1) STATISTICAL APPLICATIONS IN NEUROSCIENCE Chair: Elizabeth Sweeney

E0343: **Longitudinal ComBat: A method for harmonizing longitudinal multi-scanner imaging data**
Presenter: **Kristin Linn**, University of Pennsylvania, United States
Co-authors: Russell Shinohara, Joanne Beer

Neuroimaging is a major underpinning of modern neuroscience research and the study of brain development, abnormality, and disease. Combining neuroimaging datasets from multiple sites and scanners can increase statistical power for detecting biological effects of interest. However, technical variation due to differences in scanner manufacturer, model, and acquisition protocols may bias estimation of these effects. Originally proposed to address batch effects in genomic data sets, ComBat has been shown to be effective at removing unwanted variation due to scanners in cross-sectional neuroimaging data. We propose an extension of the ComBat model for longitudinal data and demonstrate its performance using simulations as well as longitudinal cortical thickness data from the Alzheimers Disease Neuroimaging Initiative (ADNI) study. We demonstrate that longitudinal ComBat controls type I error and have higher power for detecting changes in thickness over time compared to alternative methods such as applying cross-sectional ComBat to the longitudinal thickness trajectories.

E0758: **Quantitative susceptibility maps in multiple sclerosis lesions**
Presenter: **Elizabeth Sweeney**, University of Pennsylvania, United States

Multiple sclerosis (MS) is an inflammatory disease of the central nervous system characterized by lesions in the brain and spinal cord. Magnetic resonance images (MRI) are sensitive to these lesions. A particular type of lesion called a chronic active lesion, is characterized by a hyperintense rim of iron-enriched, activated microglia and macrophages, and has been linked to greater tissue damage. An MRI technique called quantitative susceptibility mapping (QSM) provides efficient in vivo quantification of susceptibility changes related to iron deposition and identifies these chronic active lesions, called QSM rim positive (rim+) lesions. QSM rim+ MS lesions and their longitudinal behavior have the potential to serve as a biomarker of chronic inflammation and to be utilized to monitor disease progression and evaluate disease-modifying therapies in MS. We will discuss the challenges of estimating treatment effects using the longitudinal behavior of QSM rim+ lesions. One of the major limitations of this model is that the inflammatory stage or age of the lesion is unknown, causing misregistration of the lesion-level data. We will also introduce a methodology to estimate the age of MS lesions using cross-sectional MRI information.

E0880: **Modeling trajectories using functional linear first-order differential equations**
Presenter: **Julia Wrobel**, Colorado School of Public Health, United States

A novel regression method is introduced that fuses concepts from functional linear regression and ordinary differential equations. The method models i.i.d. trajectories in a dynamical systems framework which captures in the influence of forces and derivatives on the path of an object. The motivation comes from novel data from an experiment exploring the relationship between neural firing rates and hand trajectories of mice performing a reaching task while under neurological assessment. This is an example from the increasingly common class of problems where outcome and responses are measured densely in parallel. For these data streams, we want to understand the relationship between inputs and outputs that are both functions measured on the same domain. Recent work using these data suggests that the dynamics of the arm during dexterous, voluntary movements are tightly coupled to neural control signals from the motor cortex. To better quantify how brain activity affects current and future paw position, our model incorporates initial position and has parameters that treat the relationship between the paw trajectory and the brain as a dynamical system of inputs and outputs, that state of which evolve over time. We compare our method to historical functional linear regression in simulations and on the mouse kinematic data.

E1019: **Nonparametric functional data modeling of pharmacokinetic processes with applications in dynamic PET imaging**
Presenter: **Todd Ogden**, Columbia University, United States
Co-authors: Baoyi Shi

Modeling a pharmacokinetic process typically involves solving a system of linear differential equations and estimating the parameters upon which the functions depend. In order for this approach to be valid, it is necessary that a number of fairly strong assumptions hold, assumptions involving various aspects of the kinetic behavior of the substance being studied. In many situations, such models are understood to be simplifications of the “true” kinetic process. While in some circumstances such a simplified model may be a useful (and close) approximation to the truth, in some cases, important aspects of the kinetic behavior cannot be represented. We present a nonparametric approach, based on principles of functional data analysis, to modeling pharmacokinetic data. We illustrate its use through application to data from a dynamic PET imaging study of the human brain.

EO015 Room 102 (Hybrid 2) RECENT DEVELOPMENTS ON NETWORK AND TENSOR DATA ANALYSIS Chair: Yuan Zhang

E0292: **Rank and factor loadings estimation in time series tensor factor model by pre-averaging**
Presenter: **Clifford Lam**, London School of Economics and Political Science, United Kingdom

A pre-averaging method is introduced for tensor time series data to estimate factor loadings matrices and the rank of the core tensor for a time series tensor factor model. Without the knowledge of either the rank or the factor loading matrices, we pre-average the fibres of an unfolded tensor and systematically search for one that maximizes the signals from the factors. Projection directions corresponding to the “strongest” factors for each mode of the tensor are then obtained. These directions are then used to re-estimate the rank of the core tensor and all factor loadings matrices. Rates of convergence of these estimators are spelt out, all under a set of econometrics assumptions for the factors and the noise of the tensor data, allowing serial correlations in the noise as well as cross-correlations among noise fibres. Our proposed method bypass the difficulty of proving that the estimated factor loadings matrices produced from the usual HOOI converge to the true underlying ones even under the i.i.d. noise setting. Simulation results show the effectiveness of our method compared to other state-of-the-art ones. A set of real data is also analyzed.

E0219: **Community detection in general hypergraph via garph embedding**
Presenter: **Yaoming Zhen**, City University of Hong Kong, Hong Kong
Co-authors: Junhui Wang

Conventional network data has primarily focused on pairwise interactions, yet multi-way interactions among multiple entities have been frequently observed in real-life hypergraph networks. A novel method is proposed for detecting community structure in general hypergraph networks, uniform or non-uniform. The proposed approach introduces a null vertex to augment a non-uniform hypergraph into a uniform multi-hypergraph and then embeds the multi-hypergraph in a low-dimensional vector space such that vertices within the same community are close to each other. The resultant optimization task can be efficiently tackled by an alternative updating scheme. The asymptotic consistencies of the proposed method are established in terms of both community detection and hypergraph estimation, which are also supported by numerical experiments on some synthetic and real-life hypergraph networks.

E0332: **Smooth tensor estimation with unknown permutations**
Presenter: **Chanwoo Lee**, University of Wisconsin - Madison, United States
Co-authors: Miaoyan Wang

The problem of structured tensor denoising in the presence of unknown permutations is considered. Such data problems arise commonly in recommendation systems, neuroimaging, community detection, and multiway comparison applications. We develop a general family of smooth tensor models up to arbitrary index permutations; the model incorporates the popular tensor block models and Lipschitz hypergraphon models as special cases. We show that a constrained least-squares estimator in the block-wise polynomial family achieves the minimax error bound. A phase transition phenomenon is revealed with respect to the smoothness threshold needed for optimal recovery. In particular, we find that a polynomial of degree up to $(m-2)(m+1)/2$ is sufficient for accurate recovery of order-m tensors, whereas a higher degree exhibits no further benefits. This phenomenon reveals the intrinsic distinction for smooth tensor estimation problems with and without unknown permutations. Furthermore, we provide an efficient polynomial-time Borda count algorithm that provably achieves optimal rate under monotonicity assumptions. The efficacy of our procedure is demonstrated through both simulations and Chicago crime data analysis.

E0597: **Randomization inference in experiments on networks**
Presenter: **David Choi**, Carnegie Mellon University, United States

In experiments that study social phenomena, such as peer influence or herd immunity, the treatment of one unit may influence the outcomes of others. Such interference between units violates traditional approaches for causal inference so that additional assumptions are often imposed to model or limit the underlying social mechanism. For binary outcomes, we propose an approach that does not require such assumptions, allowing for interference that is both unmodeled and arbitrarily strong, with confidence intervals derived using only the randomization of treatment. However, the estimates will have wider confidence intervals and weaker causal implications than those attainable under stronger assumptions, essentially showing only that effects exist and are associated with specified measures of treatment exposure, such as the number of treated friends or neighborhood treatment rate. The approach allows for the usage of regression, matching, or weighting, as may best fit the application at hand. The inference is done by bounding the distribution of the estimation error over all possible values of the unknown counterfactual, using an integer program. Examples are shown using a vaccination trial and two experiments investigating the effects of social influence.

E0831: **Individual-centered partial information in social networks**
Presenter: **Xiao Han**, University of Science and Technology of China, China

In statistical network modeling and inference, we often assume either the full network is available or multiple subgraphs can be sampled to estimate various global properties of the full network. However, in a real social network, people frequently make decisions based on their local view of the network alone. We consider a partial information framework that characterizes the local network centered at a given individual by path length (or knowledge depth L) and gives rise to a partial adjacency matrix. Under $L=2$, we focus on the problem of (global) community detection using the popular stochastic block model (SBM) and its degree-corrected variant (DCSBM). We derive general properties of the eigenvalues and eigenvectors from the major term of the partial adjacency matrix and propose new spectral-based community detection algorithms for these two types of models, which can achieve almost exact recovery under appropriate conditions. Our settings in the DCSBMalso allow us to interpret the efficiency of clustering using neighborhood features of the central node. Using simulated and real networks, we demonstrate the performance of our algorithms in inferring global community memberships using a partial network. In particular, we show that the clustering accuracy indicates the different global structure is visible to different individuals.

EO045 **Room 103 (Hybrid 3)** PRECISION MEDICINE WITH COMPLETE DATA **Chair: Yifan Cui**

E0889: **A reinforcement learning framework for A/B testing**
Presenter: **Chengchun Shi**, LSE, United Kingdom

A/B testing, or online experiment, is a standard business strategy to compare a new product with an old one in pharmaceutical, technological, and traditional industries. Major challenges arise in online experiments of two-sided marketplace platforms (e.g., Uber) where there is only one unit that receives a sequence of treatments over time. In those experiments, the treatment at a given time impacts the current outcome as well as future outcomes. We introduce a reinforcement learning framework for carrying out A/B testing in these experiments while characterizing the long-term treatment effects. The proposed testing procedure allows for sequential monitoring and online updating. It is generally applicable to a variety of treatment designs in different industries. In addition, we systematically investigate the theoretical properties of our testing procedure. Finally, we apply our framework to both simulated data and a real-world data example obtained from a ridesharing company to illustrate its advantage over the current practice.

E0951: **Variable selection for individualized treatment rules with discrete outcomes**
Presenter: **Zeyu Bian**, McGill University, Canada
Co-authors: Erica Moodie, Sahir Bhatnagar

An individualized treatment rule (ITR) is a decision rule that aims to improve individual patients' health outcomes by recommending optimal treatments according to patients' specific information. In observational studies, collected data may contain many variables which are irrelevant to making treatment decisions. Including all available variables in the statistical model for the ITR could yield a loss of efficiency and an unnecessarily complicated treatment rule, which is difficult for physicians to interpret or implement. Thus, a data-driven approach to selecting important covariates with the aim of improving the estimated decision rules is crucial. While there is a growing body of literature on selecting variables in ITRs with continuous outcomes, relatively few methods exist for discrete outcomes, which pose additional computational challenges even in the absence of variable selection. We propose a variable selection method for ITRs with discrete outcomes. We show theoretically and empirically that our approach has the double robustness property, and that it compares favorably with other competing approaches. We illustrate the proposed method on data from a study of an adaptive web-based stress management tool to identify which variables are relevant for tailoring treatment.

E0904: **Model-assisted uniformly honest inference for optimal treatment regimes in high dimension**
Presenter: **Yunan Wu**, The University of Texas at Dallas, United States
Co-authors: Lan Wang, Haoda Fu

New tools are developed to quantify uncertainty in optimal decision making and to gain insight into which variables one should collect information about given the potential cost of measuring a large number of variables. We investigate simultaneous inference to determine if a group of variables is relevant for estimating an optimal decision rule in a high-dimensional semiparametric framework. The unknown link function permits flexible modelling of the interactions between the treatment and the covariates, but leads to nonconvex estimation in high dimensions and imposes significant challenges for inference. We first establish that a local restricted strong convexity condition holds with high probability and that any feasible local sparse solution of the estimation problem can achieve the near-oracle estimation error bound. We further rigorously verify that a wild bootstrap procedure based on a debiased version of the local solution can provide asymptotically honest uniform inference for the effect of a group of variables on optimal decision making. We also propose an efficient algorithm for estimation. Our simulations and real data example suggest satisfactory performance.

E0900: **Reinforcement learning in possibly nonstationary environments**
Presenter: **Mengbing Li**, University of Michigan, United States
Co-authors: Chengchun Shi, Zhenke Wu, Piotr Fryzlewicz

Reinforcement learning (RL) methods are considered in offline nonstationary environments. Many existing RL algorithms in the literature rely on the stationarity assumption that requires the system transition and the reward function to be constant over time. However, the stationarity assumption is restrictive in practice and is likely to be violated in a number of applications, including traffic signal control, robotics and mobile health. We develop a consistent procedure to test the nonstationarity of the optimal policy based on pre-collected historical data, without additional online data collection. Based on the proposed test, we further develop a sequential change-point detection method that can be naturally coupled with existing state-of-the-art RL methods for policy optimisation in nonstationary environments. The usefulness of our method is illustrated by theoretical results, simulation studies, and a real data example from the 2018 Intern Health Study.

E0898: **Dynamic treatment effects: High-dimensional inference under model misspecification**
Presenter: **Yuqian Zhang**, Renmin University of China, China
Co-authors: Jelena Bradic, Weijie Ji

The estimation and inference of average treatment effects in dynamic settings are considered, where covariates and treatments are longitudinal. We focus on high-dimensional cases when the sample size N is potentially much smaller than the covariate vectors dimension d. The marginal structural mean models are considered. We identify a new, broad doubly (multiply) robust estimator, which we name a “sequential model doubly robust estimator”. We achieve root-N inference even when model misspecification occurs. For that purpose, new loss functions and new nuisance parameters are introduced, named "moment targeted", aimed to reduce the bias of model misspecification. New loss functions resolve a long-standing open problem of dynamic double robustness. We identify the weakest conditions up to date that match naive intuition. Multiple time model double robustness is achieved whenever each time exposure is model doubly-robust itself. This significantly extends the literature even in low-dimensions, where the doubly robust property requires a number of complex conditions to hold.

EO097 Room 104 (Hybrid 4) NEW FRONTIERS IN NETWORK DATA ANALYSIS Chair: Emma Jingfei Zhang

E0462: **Using maximum entry-wise deviation to test the goodness-of-fit for stochastic block models**
Presenter: **Emma Jingfei Zhang**, University of Miami, United States

The stochastic block model is widely used for detecting community structures in network data. How to test the goodness-of-fit of the model is one of the fundamental problems and has gained growing interest in recent years. We propose a novel goodness-of-fit test based on the maximum entry of the centered and re-scaled adjacency matrix for the stochastic block model. One noticeable advantage of the proposed test is that the number of communities can be allowed to grow linearly with the number of nodes ignoring a logarithmic factor. We prove that the null distribution of the test statistic converges in distribution to a Gumbel distribution, and we show that both the number of communities and the membership vector can be tested via the proposed method. Furthermore, we show that the proposed test has an asymptotic power guarantee against a class of alternatives. We also demonstrate that the proposed method can be extended to the degree-corrected stochastic block model. Both simulation studies and real-world data examples indicate that the proposed method works well.

E0499: **Learning cross-layer dependence structure for multilayer networks**
Presenter: **Jonathan Stewart**, Florida State University, United States

Multilayer networks are a network data structure in which a set of nodes in a population of interest have multiple modes of interaction or relation (e.g., persons in a social network can have familial ties, friendships, professional relationships, and more). Each layer of the network corresponds to an individual network or graph defined through one mode of interactions or relations. We propose a class of models for cross-layer dependence in multilayer networks, aiming to learn how interactions in one or more layers may influence interactions in other layers of the multilayer network. On the methodological side, we develop a class of models for both sparse and dense multilayer networks that focus on modeling cross-layer dependence and can incorporate node covariates. We elaborate algorithms for parameter estimation and model selection. On the theoretical side, our contributions include establishing non-asymptotic theoretical guarantees which establish rates of convergence in high-dimensional settings for both maximum likelihood estimators and maximum pseudo-likelihood estimators, as well as a non-asymptotic bound on the error of the multivariate normal approximation for our estimators. We demonstrate a method for model selection that controls the false discovery rate and show through simulations that our methods provide an accurate learning platform for learning the cross-layer dependence structure of multilayer networks.

E0751: **Modeling continuous-time networks of relational events**
Presenter: **Subhadeep Paul**, The Ohio State University, United States

Spatiotemporal data with complex network dependencies are increasingly available in many application problems involving human mobility, social media, disease transmission, and international relationships. The observed data consist of timestamped relational events in many such application settings. For example, in social media, users interact with each other through events that occur at specific time instances such as liking, mentioning, commenting, or sharing another user’s content. In international relations and conflicts, nations commit acts of hostility or disputes through discrete time-stamped events. We will introduce statistical models and methods for analyzing such datasets combining tools from network analysis and multivariate point processes. We will also describe scalable estimation methods and study the asymptotic properties of the estimators. Finally, we will demonstrate the models are able to fit several real datasets well and predict temporal motif structures in those datasets.

E1008: **Scalable community detection in massive networks via predictive inference**
Presenter: **Srijan Sengupta**, North Carolina State University, United States
Co-authors: Marianna Pensky, Subhankar Bhadra

Identification of community structure in networks has been of particular interest in the statistics literature. In recent years, we have witnessed massive network datasets being generated in many fields. Community detection is challenging for such massive networks since existing standard community detection algorithms require high runtime and storage. We propose a novel algorithm using so-called predictive inference, where we use any statistically sound community detection algorithm to cluster a subset of nodes and use the estimated communities to classify the rest of the nodes. Decomposing the clustering problem into a small clustering sub-problem and a classification problem leads to excellent savings in runtime and memory with little loss of accuracy. We establish the theoretical properties of the proposed method and demonstrate its numerical performance in synthetic and real-world networks.

E0957: **Identification and estimation of causal peer effects using double negative controls for unmeasured network confounding**
Presenter: **Naoki Egami**, Columbia University, United States

Scientists have been interested in estimating causal peer effects to understand how peoples behaviors are affected by their network peers. However, it is well known that identification and estimation of causal peer effects are challenging in observational studies for two reasons. The first is the identification challenge due to unmeasured network confounding, for example, homophily bias and contextual confounding. The second issue is network dependence of observations, which one must take into account for valid statistical inference. Negative control variables, also known as placebo variables, have been widely used in observational studies including peer effect analysis over networks, although they have been used primarily for bias detection. We establish a formal framework which leverages a pair of negative control outcome and exposure variables (double negative controls) to nonparametrically identify causal peer effects in the presence of unmeasured network confounding. We then propose a generalized method of moments estimator for causal peer effects, and establish its consistency and asymptotic normality under an assumption about psi-network dependence. Finally, we provide a network heteroskedasticity and autocorrelation consistent variance estimator. Our methods are illustrated with an application to peer effects in education.

EO195 Room 105 (Hybrid 5) SEMI-PARAMETRIC INFERENCE AND MODELING WITH SHAPE-CONSTRAINTS Chair: Hyebin Song

E0815: **Joint estimation of monotone curves via functional principal component analysis**

Presenter: **Yei Eun Shin**, National Cancer Institute, United States

A functional data approach is developed to jointly estimate a collection of monotone curves that are irregularly and possibly sparsely observed with noise. In this approach, the unconstrained relative curvature curves instead of the monotone-constrained functions are directly modeled. Functional principal components are used to describe the major modes of variations of curves and allow borrowing strength across curves for improved estimation. A two-step approach and an integrated approach are considered for model fitting. The simulation study shows that the integrated approach is more efficient than separate curve estimation and the two-step approach. The integrated approach also provides more interpretable principal component functions in an application of estimating weekly wind power curves of a wind turbine.

E0823: **Nonparametric inference under a monotone hazard ratio order**

Presenter: **Ted Westling**, University of Massachusetts Amherst, United States
Co-authors: Yujian Wu

The ratio of the hazard functions of two populations or two strata of a single population plays an important role in a time-to-event analysis. Cox regression is commonly used to estimate the hazard ratio under the assumption that it is constant in time, which is known as the proportional hazards assumption. However, this assumption is often violated in practice, and when it is violated, the parameter estimated by Cox regression is difficult to interpret. The hazard ratio can be estimated in a nonparametric manner using smoothing, but smoothing-based estimators are sensitive to the selection of tuning parameters, and it is often difficult to perform valid inferences with such estimators. In some cases, it is known that the hazard ratio function is monotone. We demonstrate that the monotonicity of the hazard ratio function defines an invariant stochastic order, and we study the properties of this order. Furthermore, we introduce an estimator of the hazard ratio function under a monotonicity constraint. We demonstrate that our estimator converges in distribution to a mean-zero limit, and we use this result to construct asymptotically valid confidence intervals. Finally, we conduct numerical studies to assess the finite-sample behavior of our estimator, and we use our methods to estimate the hazard ratio of progression-free survival in pulmonary adenocarcinoma patients treated with Gefitinib or carboplatin-paclitaxel.

E0835: **Piecewise monotone estimation in one-parameter exponential families**

Presenter: **Yuto Miyatake**, Osaka University, Japan
Co-authors: Takeru Matsuda

The problem of estimating a piecewise monotone sequence of normal means is called the nearly isotonic regression. An efficient algorithm has been devised for this problem by modifying the pool adjacent violators algorithm (PAVA). We are concerned with estimating a piecewise monotone sequence for general one-parameter exponential families such as binomial, Poisson, and chi-square. We propose an efficient algorithm based on the modified PAVA, which utilizes the duality between the natural and expectation parameters. We also provide a method for selecting the regularization parameter using an information criterion. Simulation results demonstrate that the proposed method detects change-points in piecewise monotone parameter sequences in a data-driven manner. We present several applications such as spectrum estimation, causal inference and discretization error quantification of ODE solvers.

E0907: **A monotone single index model for missing-at-random longitudinal proportion data**

Presenter: **Satwik Acharyya**, University of Michigan, United States
Co-authors: Debdeep Pati, Dipankar Bandyopadhyay

Beta distributions are commonly used to model proportion valued response variables, commonly encountered in longitudinal studies. We develop semi-parametric Beta regression models for proportion valued responses, where the aggregate covariate effect is summarized and flexibly modeled, using an interpretable monotone time-varying single index transform of a linear combination of the potential covariates. We utilize the potential of single-index models, which are effective dimension reduction tools and accommodate link function misspecification in generalized linear mixed models. Our Bayesian methodology incorporates the missing-at-random feature of the proportion response and utilizes Hamiltonian Monte Carlo sampling to conduct inference. We explore finite-sample frequentist properties of our estimates and assess the robustness via detailed simulation studies. Finally, we illustrate our methodology via application to a motivating longitudinal dataset on obesity research recording proportion body fat.

EO049 Room 106 (Hybrid 6) ADAPTIVE CLINICAL TRIAL DESIGN Chair: Yisheng Li

E0896: **Platform designs with added new arms**

Presenter: **Haitao Pan**, St. Jude Children's Research Hospital, United States

Platform trials defined as multi-arm multi-stage trials with adding and removing experimental arms have gained attention recently. Statistical issues with respect to "adding new arms", however, have not been well discussed. In the setting of pre-planned deferred arms, with goals of controlling the familywise error rate (FWER) and achieving a targeted experimental-wise power, we provide a principled approach, including, how to modify the critical boundaries to control the FWER when new arms are added, how to re-estimate the sample size and provide the optimal allocation ratio for the control to experimental arm to maintain the experimental-wise power. The influence of timing of adding new arms to the design's operating characteristics has also been examined. We developed an R package for practitioners to implement this method.

E0932: **A semi-mechanistic dose-finding design in oncology using pharmacokinetic/pharmacodynamic modeling**

Presenter: **Yisheng Li**, The University of Texas MD Anderson Cancer Center, United States

While a number of phase I dose-finding designs in oncology exist, the commonly used ones are either algorithmic or empirical model-based. We propose a new framework for modeling the dose-response relationship, by systematically incorporating the pharmacokinetic (PK) data collected in the trial and the hypothesized mechanism of the drug effects, via dynamic PK/PD modeling, as well as modeling of the relationship between a latent cumulative pharmacologic effect and a binary toxicity outcome. The resulting design is an extension of the existing designs that make use of pre-specified summary PK information (such as the area under the concentration-time curve [AUC] or maximum serum concentration). The simulation studies show, with moderate departure from the hypothesized mechanisms of the drug action, that the performance of the proposed design on average improves upon those of the common designs, including the continual reassessment method, Bayesian optimal interval design, modified toxicity probability interval method, and a design called PKLOGIT that models the effect of AUC on toxicity. In case of considerable departure from the underlying drug effect mechanism, the performance of the design is shown to be comparable with that of the other designs. We illustrate the proposed design by applying it to the setting of a phase I trial of a γ-secretase inhibitor in metastatic or locally advanced solid tumors. We also provide an R package to implement the proposed design.

E0954: **A Bayesian adaptive design for pediatric basket trials**
Presenter: **Yimei Li**, University of Pennsylvania, United States

The basket trial is a novel type of trial that treats patients with the same genetic aberration regardless of their cancer types. Pediatric basket trials have unique features that require additional considerations for incorporating the adult information. We propose a Bayesian basket design for pediatric trials with adult data (BPAD) that performs dual information borrowing: borrow information from the adult data to the pediatric trial and between the cancer types within the pediatric trial. The BPAD design also accommodates potential heterogeneous treatment effect across cancer types, by allowing each cancer type belonging to the sensitive or insensitive latent subgroups. To make a go/no-go decision for each cancer type in the interim analyses, the design adaptively update the members of the subgroups based on the accumulated pediatric and adult data and borrow information among cancer types within the same subgroup. The simulation study shows that the BPAD design has better performance than a few existing designs, yielding high power to detect the treatment effect for the sensitive cancer types and maintaining a desirable type I error rate for the insensitive cancer types.

E1004: **Biomarker-based Bayesian randomized clinical trial design for identifying a target population**
Presenter: **Akiyoshi Nakakura**, Kyoto University Graduate School of Medicine, Japan
Co-authors: Satoshi Morita, Yasuo Sugitani, Hideharu Yamamoto

The challenges and potential benefits of incorporating biomarkers into clinical trial designs have been increasingly discussed, in particular, to develop new agents for immuno-oncology or targeted cancer therapies. To more accurately identify a sensitive subpopulation of patients, in many cases, a larger sample size - and consequently higher development costs and a longer study period - might be required. A biomarker-based Bayesian (BM-Bay) randomized clinical trial design is discussed whcih incorporates a predictive biomarker measured on a continuous scale with pre-determined cutoff points or a graded scale to define multiple patient subpopulations. We consider designing interim analyses with suitable decision criteria to achieve correct and efficient identification of a target patient population for developing a new treatment. The proposed decision criteria allow not only the take-in of sensitive subpopulations but also the ruling-out of insensitive ones on the basis of the efficacy evaluation of a time-to-event outcome. Extensive simulation studies are conducted to evaluate the operating characteristics of the proposed method, including the probability of correct identification of the desired subpopulation and the expected number of patients, under a wide range of clinical scenarios.

EO109 Room 107 (Hybrid 7) CAUSAL INFERENCE AND REINFORCEMENT LEARNING Chair: Peter Song

E0167: **Theory for identification and inference with synthetic controls: A proximal causal inference framework**
Presenter: **Xu Shi**, University of Michigan, United States

Synthetic control methods are commonly used to estimate the treatment effect on a single treated unit in panel data settings. A synthetic control (SC) is a weighted average of control units built to match the treated unit's pre-treatment outcome trajectory, with weights typically estimated by regressing pre-treatment outcomes of the treated unit to those of the control units. However, it has been established that such regression estimators can fail to be consistent. We introduce a proximal causal inference framework to formalize identification and inference for both the SC weights and the treatment effect on the treated. We show that control units previously perceived as unusable can be repurposed to consistently estimate the SC weights. We also propose to view the difference in the post-treatment outcomes between the treated unit and the SC as a time series, which opens the door to a rich literature on time-series analysis for treatment effect estimation. We further extend the traditional linear model to accommodate general nonlinear models allowing for binary and count outcomes which are understudied in the SC literature. We illustrate our proposed methods with simulation studies and an application to the evaluation of the 1990 German Reunification.

E0346: **Learning causal directionality via a directed entropy-information metric**
Presenter: **Peter Song**, University of Michigan, United States
Co-authors: Soumik Purkayastha

Interest in direction of dependence in statistical literature has been rising, as such knowledge is critical for directed graphs used to study causality and mediation. Studying asymmetric dependence between two variables may help confirm a hypothesized causal relationship. While most standard statistical models can estimate the magnitude of associations, they cannot distinguish between effect and direction of effect as they implicitly assume a symmetric dependence between two variables. We posit a new measure called the directed mutual information (DMI) in which mutual information and conditional entropy are used to study both association and directed dependence. We establish key large-sample properties for the DMI and develop algorithms to test for independence as well as quantify directed dependence. The proposed method is evaluated by simulation studies and applied to a real-world example.

E0508: **Directed acyclic graphs with unmeasured confounders using instrumental variables**
Presenter: **Kean Ming Tan**, University of Michigan, United States

Directly acyclic graphs play an important role to describe causal relationships among a set of random variables. Existing work has mainly focused on estimating directed acyclic graphs under the assumption that all of the relevant variables are observed. However, in many scientific settings, there may be unobserved variables that are associated with the observed variables. Without adjusting the unobserved variables, the estimated causal relationships among the set of observed variables may be spurious. To address the aforementioned issue, we propose to estimate a directly acyclic graph under the presence of unmeasured confounders using an instrumental variable approach. The approach is motivated by an application in neuroscience: using optogenetics, neuroscientists can design and activate specific neurons with optical stimulation, which can be treated as instrumental variables. Specifically, we show that the ancestral relationships can be recovered with a computationally efficient screening procedure. The screening procedure is then combined with the generalized two-stage least squares method to ensure computationally efficient recovery of the model parameters subjected to the acyclic constraint of a directed acyclic graph. We illustrate the performance of the proposed method via extensive numerical studies and an application to the aforementioned neuroscience data.

E0990: **Opportunities for causal inference and reinforcement learning in real-world interventional mobile health studies**
Presenter: **Zhenke Wu**, University of Michigan at Ann Arbor, United States

Twin revolutions in wearable technologies and smartphone-delivered digital health interventions have significantly expanded the accessibility and uptake of personalized interventions in multiple domains of health sciences. For example, push notifications to promote healthy behaviors can be sent via mobile devices that are adapted to continuously collect information on an individual's current context. These time-varying adaptive interventions are hypothesized to lead to meaningful short- and long-term behaviour change. Key scientific questions in statistical terms will be formulated. However, standard assumptions such as non-interference and stationarity might be violated in real-world mobile health studies due to peer influence and long monitoring periods. We will present two methodological solutions, the first for estimating a new type of peer effects and the second for optimal policy learning under possibly non-stationary environments. We will use a multi-institution cohort of first-year medical interns in the United States to illustrate the ideas. We will also highlight that teams of engineers, and clinical and data scientists can collaborate to build statistical models that extract scientific insights from noisy and longitudinal interventional mobile health data.

EO023 **Room Virtual R1** **MODERN MULTIVARIATE METHODS FOR MULTIFACETED DATA** **Chair: Anuradha Roy**

E0204: **A nonparametric mixed-effects mixture model for patterns of clinical measurements associated with COVID-19**
Presenter: **Yuedong Wang**, University of California - Santa Barbara, United States

Some, but not all, COVID-19 patients had changes in biological/clinical variables such as temperature and oxygen saturation days before symptoms occur. We propose a flexible nonparametric mixed-effects mixture model (NMEM) that simultaneously identifies risk factors and classifies patients with biological change. We model the latent biological change probability using a logistic regression model with L1 penalty and trajectories in each latent class using splines. We apply the EM algorithm and penalized likelihood to estimate all parameters and mean functions. Simulation studies indicate the proposed method performs well. We apply the NMEM model to investigate changes in temperature in COVID-19 patients receiving hemodialysis.

E0266: **Multivariate-t linear mixed models for longitudinal data with censored and intermittent missing responses**
Presenter: **Wan-Lun Wang**, National Cheng Kung University, Taiwan

Multivariate longitudinal data arising in clinical trials and medical studies often exhibit complex features such as censored responses, intermittent missing values, and atypical or outlying observations. The multivariate-t linear mixed model (MtLMM) has been recognized as a powerful tool for robust modeling of multivariate longitudinal data in the presence of potential outliers or fat-tailed noises. A generalization of MtLMM, called the MtLMM-CM, is presented to properly adjust for censorship due to detection limits of the assay and missingness embodied within multiple outcome variables recorded at irregular occasions. An expectation conditional maximization either (ECME) algorithm is developed to compute parameter estimates using the maximum likelihood (ML) approach. The asymptotic standard errors of the ML estimators of fixed effects are obtained by inverting the empirical information matrix according to Louis' method. The proposed methodology is illustrated on a real dataset from HIV-AIDS studies and a simulation study under a variety of scenarios.

E0280: **Mixtures of unrestricted skew normal factor analyzers with incomplete data**
Presenter: **Tsung-I Lin**, National Chung Hsing University, Taiwan

Mixtures of factor analyzers (MFA) based on the restricted skew-normal distribution (rMSN) have been shown to be a flexible tool to handle asymmetrical high-dimensional data with heterogeneity. However, the rMSN distribution is oft-criticized a lack of sufficient ability to accommodate potential skewness arising from more than one feature space. An alternative extension of MFA is presented by assuming the unrestricted skew-normal (uMSN) distribution for the component factors. In particular, the proposed mixtures of unrestricted skew-normal factor analyzers (MuSNFA) can simultaneously capture multiple directions of skewness and deal with the occurrence of missing values or nonresponses. Under the missing at random (MAR) mechanism, we develop a computationally feasible expectation conditional maximization (ECM) algorithm for computing the maximum likelihood estimates of model parameters. Practical aspects related to model-based clustering, prediction of factor scores and missing values are also discussed. The utility of the proposed methodology is illustrated with the analysis of simulated data and the Pima Indian women's diabetes data containing genuine missing values.

E0311: **Quantum enhanced feature subset selection**
Presenter: **Basabi Chakraborty**, Iwate Prefectural University, Japan

Optimal feature subset selection is an important prerequisite for any pattern classification or machine learning problem. Redundant and irrelevant features degrade performance as well as increase the computational cost of the classifier. An efficient feature evaluation metric and an optimal search process are the two basic requirements for optimal feature subset selection. A lot of feature subset selection algorithms have been developed so far based on statistical and mathematical tools. The recent rapid increase of high dimensional data has created high computational challenges in the area of machine learning and data mining and the need for stable and scalable feature selection algorithms with reduced computational cost is ever increasing. Quantum computing is known to possess enormous processing ability by exploiting inherent parallelism and potentially provides significant speedup compared to classical computing. We would like to describe work on the development of quantum-enhanced optimal feature subset selection algorithms. We have proposed a quantum-based optimization approach to classical feature evaluation metrics and simulated it on classical machines to examine their effectiveness. We have also proposed a novel quantum-inspired metaheuristic based feature selection algorithm and examined its performance with simulation experiments for benchmark data sets.

E0642: **Penalized likelihood approach in multivariate regression with missing values and its application to materials science**
Presenter: **Kei Hirose**, Kyushu University, Japan
Co-authors: Keisuke Teramoto

In the field of materials science and engineering, statistical analysis has recently been used to predict multiple material properties from an experimental design. These material properties correspond to response variables in the multivariate regression model. We conduct a penalized maximum likelihood procedure to estimate model parameters, including the regression coefficients and covariance matrix of response variables. In particular, we employ L1 regularization to achieve a sparse estimation of regression coefficients and inverse covariance matrix of response variables. In some cases, there may be a relatively large number of missing values in the response variables, owing to the difficulty of collecting data on material properties. We, therefore, propose a method that incorporates a correlation structure among the response variables into a statistical model to improve the prediction accuracy. The expectation-maximization (EM) algorithm is constructed, which allows the application to a dataset with missing values in the responses. We apply our proposed procedure to real data consisting of 22 material properties.

EO293 **Room Virtual R10** **INFERENCE OF HIERARCHICAL AND NONPARAMETRIC STRUCTURES** **Chair: Minwoo Chae**

E0272: **Deep generative models for nonparametric estimation of singular distributions**
Presenter: **Minwoo Chae**, Pohang University of Science and Technology, Korea, South

While deep generative models are popularly used to model high-dimensional data, theoretical understanding of it is largely unexplored. We investigate the statistical properties of deep generative models from a nonparametric distribution estimation viewpoint. In the considered model, data are assumed to concentrate around some low-dimensional structure. Estimating the distribution supported on this low-dimensional structure is challenging due to its singularity. In particular, a likelihood approach can fail to estimate the target distribution consistently. We obtain convergence rates with respect to the Wasserstein metric for two methods: a sieve MLE based on the perturbed data and a GAN type estimator. Our analysis gives some insights into i) how deep generative models can avoid the curse of dimensionality, ii) how likelihood approaches work for singular distribution estimation, and iii) why GAN performs better than likelihood approaches.

E0362: **Bayesian data selection**
Presenter: **Jeff Miller**, Harvard University, United States
Co-authors: Eli Weinstein

Insights into complex, high-dimensional data can be obtained by discovering features of the data that match or do not match a model of interest. To formalize this task, we introduce the "data selection" problem: finding a lower-dimensional statistic - such as a subset of variables - that is well fit by a given parametric model of interest. A fully Bayesian approach to data selection would be to parametrically model the value of the statistic, nonparametrically model the remaining "background" components of the data, and perform standard Bayesian model selection for the choice of statistic. However, fitting a nonparametric model to high-dimensional data tends to be highly inefficient, statistically and computationally. We propose a novel score for performing both data selection and model selection, the "Stein volume criterion", that takes the form of a generalized marginal likelihood with a kernelized Stein discrepancy in place of the Kullback-Leibler divergence. The Stein volume criterion does not require one to fit or even specify a nonparametric background model, making it straightforward to compute - in many cases it is as simple as fitting the parametric model of interest with an alternative objective function.

E0547: **A unified framework for parameters estimation in finite mixture models**
Presenter: **Yun Wei**, Duke University, United States
Co-authors: Sayan Mukherjee, Long Nguyen

For parameters estimation in finite mixture models, the minimum distance estimators and the denoised method of moments estimator are known to be minimax optimal. We provide a unified framework, including both estimators as special cases, and the unified framework could be applied to produce new estimators by choosing different classes of test functions. Theories are obtained under the unified framework and they extend the existing theories when specializing our framework to their cases.

E0535: **Max-infinitely divisible processes with exchangeability and their inference**
Presenter: **Shuhei Mano**, The Institute of Statistical Mathematics, Japan

Infinitely divisible processes play important roles in Bayesian nonparametric inference, as we see in constructions of the Dirichlet process prior. It is known that there is a close relationship between infinitely divisible processes and max-infinitely divisible processes. We will focus on max-infinitely divisible processes with exchangeability and discuss their inference, where we can see some similarities to that in Bayesian nonparametric models.

E0698: **On the multivariate Fourier integral theorem: Statistical and methodological perspectives**
Presenter: **Nhat Ho**, University of Texas, Austin, United States

The focus is on overcoming the inferential and interpretability challenges of deep neural networks by use of the Fourier integral theorem, a remarkable result from mathematics. We first demonstrate that the Fourier integral theorem provides natural Monte Carlo estimators in many machine learning and data science problems, such as multivariate density estimation. Then, leveraging our insight from these estimators, we propose a novel generative model based on sequentially sampling each feature of new data from a martingale sequence of conditional distribution estimators. The proposed generative model paves the way for developing uncertainty quantification and predictive inference procedures. Our key finding and idea on which this proposal is dependent are that the combination of the Fourier integral theorem, Monte Carlo methods for direct estimation of quantities of interest, and sampling-based approaches to statistical and machine learning, such as generative models, provide a natural and perfect synergy both from a mathematical and practical perspective. If time permitted, we also briefly discuss applications of the multivariate Fourier integral Theorem to improve Transformer-based language models and Graph Neural Networks.

EO409 Room Virtual R11 MODERN STATISTICAL METHODS FOR COMPLEX DATA ANALYSIS Chair: Pai-Ling Li

E1036: **Sparse Bayesian CNN**
Presenter: **Yongdai Kim**, Seoul National University, Korea, South
Co-authors: Insung Kong, jinwon park

A sparse Bayesian CNN model is proposed where a sparse prior is out on the filters in each layer. We develop an efficient MCMC algorithm and investigate how well the proposed Bayesian CNN selects filters.

E0860: **Semi-supervised learning using elliptical distributions with unknown density generators**
Presenter: **Chin-Tsang Chiang**, National Taiwan University, Taiwan

A more general elliptical distribution model is proposed for the classification of the groups with both labeled and unlabeled data. Different from existing multivariate normal and t distribution models in semi-supervised learning, the forms of the density generators are left unspecified. By incorporating the information of unlabeled data into training, a pseudo maximum likelihood method is developed to estimate the finite-dimensional model parameters. An efficient computational procedure is further presented to perform the maximization of the pseudo-likelihood function. Especially, the proposed estimators of the posterior group probabilities are useful for constructing a prediction rule. In addition, our estimators are shown to be asymptotically more efficient than the corresponding ones using only labeled data. Simulations and applications to empirical data are also used to illustrate the methodology.

E0845: **Truncated estimation for functional linear model and its application to agricultural data**
Presenter: **Hidetoshi Matsui**, Shiga University, Japan

Truncated estimation for the functional linear model is a useful technique for investigating the relationship between a functional predictor and a scalar response. We consider the problem of estimating a varying-coefficient functional linear model, where the predictor is a function of time and the scalar response depends on not only a functional predictor but also an exogenous variable. The aim is to estimate the model so that the functional predictor does not relate to the response after a certain point in time at any value of the exogenous variable. We apply the sparse regularization to shrink the corresponding domain of the coefficient function towards exactly zero. Simulation studies are conducted to investigate the effectiveness of the proposed method. We also apply the method to the analysis of agricultural data to identify when an environmental factor relates to the crop yield.

E0710: **Generalized linear model with functional covariate and its derivatives**
Presenter: **Pai-Ling Li**, Tamkang University, Taiwan
Co-authors: Jeng-Min Chiou

A generalized functional linear regression model is proposed by considering a functional covariate and its derivatives as functional predictors. The unobserved derivatives of a random function may carry useful information and need to be estimated. We apply the notion of functional principal component analysis to modeling functional predictors. The proposed regression model is parameterized in various ways to investigate the effect of each functional predictor. The performance of the proposed method is demonstrated through a traffic data example.

EO061 Room Virtual R12 NEW HORIZONS IN LONGITUDINAL STUDIES Chair: MinJae Lee

E0657: **Statistical inference for streamed longitudinal data**
Presenter: **Emily Hector**, North Carolina State University, United States
Co-authors: Jingshen Wang, Lan Luo

Modern longitudinal data, for example from wearable devices, measure biological signals on a fixed set of participants at a diverging number of time points. Traditional statistical methods are not equipped to handle the computational burden of repeatedly analyzing the cumulatively growing dataset each time new data is collected. We propose a new estimation and inference framework for the streaming updating of point estimates and their standard errors across serially collected dependent datasets. Our streaming framework is used to investigate the relationship between physical activity and several diseases through the analysis of accelerometry data from the National Health and Nutrition Examination Survey.

E0795: **Predictive model for sparse longitudinal data**
Presenter: **Seonjin Kim**, Miami University, United States
Co-authors: Shixuan Wang, Hyunkeun Cho, Won Chang

A multivariate function-on-function kernel-based estimator is proposed to predict the mean response trajectory for sparse and irregularly measured longitudinal data. The kernel function is constructed by weighing in the subject-wise similarity on L_2 metric space between predictor trajectories, where we assume an analogous fashion in predictor trajectories over time would result in a similar trend in the response trajectory among subjects. In order to deal with the curse of dimensionality caused by the multiple predictors, we propose a novel multiplicative model with multivariate Gaussian kernels. This model is capable of achieving dimension reduction as well as selecting functional covariates with predictive significance. The asymptotic properties of the proposed nonparametric estimator are investigated under mild regularity conditions. We illustrate the robustness and the flexibility of our proposed methods via the simulation study and an application to the Framingham Heart Study.

E0968: **Nonparametric estimation of repeated densities with heterogeneous sample sizes**
Presenter: **Xiongtao Dai**, Iowa State University, United States
Co-authors: Jiaming Qiu, Zhengyuan Zhu

Estimating the prevalence age distributions of patients with different diseases is considered . A key challenge comes from the highly varying sample sizes for different conditions, making it difficult to estimate the age profile of a rare condition. To address this challenge, we propose a fully data-driven approach to pool information across conditions and estimate each distribution efficiently, without specifying a parametric form. Our technique draws from functional data analysis, which concerns, for example, a sample of developmental trajectories. We model densities as random trajectories and obtain low-dimensional exponential families for approximation, which is theoretically justified. We will show that the proposed approach yields interpretable results and is numerically efficient for modeling data from electronic health records.

E1037: **Bayesian analysis of longitudinal dyadic/multiple outcome data with informative missing data**
Presenter: **Jaeil Ahn**, Georgetown University, United States

Analysis of longitudinal dyadic/multiple outcomes with missing data is challenging due to the complicated correlations within and between dyads/multiple outcomes, as well as non-ignorable missing data. A Bayesian mixed-effects hybrid model is introduced to analyze longitudinal dyadic data with non-ignorable dropouts/intermittent missingness. To address this, we factorize the joint distribution of the measurement, random effects, and dropout processes into three components. The proposed model accounts for the dyadic interplay using the concept of actor and partner effects as well as dyad-specific random effects. We evaluate the performance of the proposed methods using a simulation study and apply our method to longitudinal dyadic datasets that arose from a prostate cancer trial. Then, we will introduce a Bayesian mixed-effects selection model to analyze the multivariate quality of life data with non-ignorable missing data. Compared to the first model, we first describe the overall/marginal effects of predictors on outcomes and then incorporate a variable selection feature in the missing data mechanism to evaluate the impact of potentially moderate to high dimensional outcomes on missing data mechanisms. We will illustrate how the proposed model works using a longitudinal study of the quality of life in gastric cancer patients who underwent distal gastrectomy

EO289 Room Virtual R13 ESTIMATION AND HYPOTHESIS TESTING Chair: Xuejun Jiang

E0598: **Nonnested model selection based on empirical likelihood ratio**
Presenter: **Jiancheng Jiang**, UNC Charlotte, United States

An empirical likelihood ratio (ELR) test is proposed for nonparametric model selection, where the competing models may be nested, nonnested, overlapping, misspecified, or correctly specified. It compares the prediction performances of models based on the cross-validation and allows for heteroscedasticity of the errors. We develop its asymptotic distributions for comparing any two supervised learning models under a general framework with convex loss functions. However, for general loss functions, the prediction errors from the cross-validation involve repeatedly fitting the models with one observation held out. An easily implemented approximation is then introduced. It is shown that the approximated test shares the same asymptotics as the original one. We apply the proposed tests to compare additive models and varying-coefficient models. Furthermore, a distributed ELR test is proposed to test the importance of a group of variables in possibly misspecified additive models with massive data, and a fast calculation procedure for the test is introduced. Simulations show that the proposed tests work well and have favorable finite sample performance over some existing approaches. The methodology is validated on an empirical application.

E0175: **Multiply robust estimation of quantile treatment effects with missing responses**
Presenter: **Yanlin Tang**, East China Normal University, China

Causal inference and missing data have attracted significant research interests in recent years, while the current literature usually focuses on only one of these two issues. Moreover, compared with the commonly used average treatment effect (ATE), the quantile treatment effect (QTE) is able to provide a complete picture of the difference between the treatment and control groups, as well as robustness to the outliers in the responses. Therefore, we develop a method to estimate the QTE in the context of missing data based on the idea of inverse probability weighting (IPW). The proposed IPW estimator has the property of multiply robustness, that is, as long as the class of candidate models of propensity scores contains the correct model and so does the candidate models for the probability of being observed, the resulting QTE estimator is root-n consistent and asymptotic normal. Simulation studies are conducted to investigate the performance of the proposed method, and real data from CHARLS is analyzed and different treatment effects are observed at various quantile levels of the response.

E0919: **Local influence analysis for the sliced average third-moment estimation**
Presenter: **Fei Chen**, Yunnan University of Finance and Economics, China
Co-authors: Weidong Rao, Xiaofei Liu

Sliced average third-moment estimation (SATME) is a typical method for sufficient dimension reduction (SDR) based on high order conditional moment. It is useful, particularly in the scenarios of regression mixtures. However, as SATME uses the third-order conditional moment of the predictors given the response, it may not as robust as some other SDR methods that use lower-order moments, say, sliced inverse regression (SIR) and slice average variance estimation (SAVE). Based on the space displacement function, a local influence analysis framework of SATME is constructed including a statistic of influence assessment for the observations. Furthermore, a data-trimming strategy is suggested based on the above influence assessment. The proposed methodologies solve a typical issue that also exists in some other SDR methods. A real-data analysis and simulations are presented.

E0973: **Robust estimation and test for Pearson's correlation coefficient**
Presenter: **Pengfei Liu**, Jiangsu Normal University, China

Using the idea of grouping under a moderate data framework, the median-of-means type non-parametric estimator is proposed for Pearson's correlation coefficient which has been used widely in various disciplines. Under certain conditions on the growing rate of the number of subgroups, the consistency and asymptotic normality of the proposed estimator are investigated. Furthermore, we construct a new method to test Pearson's correlation coefficient based on the empirical likelihood method for the median. Extensively numerical simulations are designed to demonstrate the superiorities of our estimator. It is shown that the new proposed estimator is quite robust with respect to outliers. Finally, we use the proposed method to study the Pearson's correlation between the open price and the rate of price spread for the Shanghai Stock Exchange composite index from May 18, 2015, to June 21, 2019.

EO043 Room Virtual R2 MACHINE LEARNING AND STABILITY Chair: Andreas Christmann

E0205: **Simple stochastic and online gradient decent algorithms for pairwise learning**
Presenter: **Yiming Ying**, State University of New York at Albany, United States

Pairwise learning refers to learning tasks where the loss function depends on a pair of instances. It instantiates many important machine learning tasks such as bipartite ranking and metric learning. A popular approach to handle streaming data in pairwise learning is an online gradient descent (OGD) algorithm, where one needs to pair the current instance with a buffering set of previous instances with a sufficiently large size and therefore suffers from a scalability issue. We will present our recent proposal of simple stochastic and online gradient descent methods for pairwise learning. A notable difference from the existing studies is that our proposed method only pairs the current instance with the previous one in building a gradient direction, which is efficient in both the storage and computational complexity. We will present novel stability results, optimization, and generalization error bounds for both convex and nonconvex as well as both smooth and nonsmooth problems. The study resolves an open question on developing meaningful generalization bounds for OGD using a buffering set with a very small fixed size. We also extend our algorithms and stability analysis to develop differentially private SGD algorithms for pairwise learning which significantly improves the existing results.

E0283: **Robust topological inference**
Presenter: **Bharath Sriperumbudur**, Pennsylvania State University, United States
Co-authors: Siddharth Viswanath, Kenji Fukumizu, Satoshi Kuruki

The distance function to a compact set plays a crucial role in the paradigm of topological data analysis. In particular, the sublevel sets of the distance function are used in the computation of persistent homology—a backbone of the topological data analysis pipeline. Despite its stability to perturbations in the Hausdorff sense, persistent homology is highly sensitive to outliers. We develop a framework of statistical inference for persistent homology in the presence of outliers. Drawing inspiration from recent developments in robust statistics, we propose a median-of-means variant of the distance function (MoM Dist) and establish its statistical properties. In particular, we show that, even in the presence of outliers, the sublevel filtrations and weighted filtrations induced by MoM Dist are both consistent estimators of the true underlying population counterpart, and their rates of convergence in the bottleneck metric are controlled by the fraction of outliers in the data. Finally, we demonstrate the advantages of the proposed methodology through simulations and applications using benchmark datasets.

E0330: **Convergence of stochastic gradient descent algorithms for functional data learning**
Presenter: **Xin Guo**, The University of Queensland, Australia
Co-authors: Xiaming Chen, Bohao Tang, Jun Fan, Zheng-Chu Guo, lei shi

Functional linear models are a fruitfully applied general framework for regression problems, including those with intrinsically infinite-dimensional data. Online gradient descent methods, despite their evidenced power of processing online or large-sized data, are not well studied for learning with functional data. We study reproducing kernel-based online learning algorithms for functional data, under both vanishing step-size and finite horizon settings. We derive convergence rates for both the prediction and estimation problems. In particular, our analysis suggests that convergence for the prediction problems requires much weaker regularity assumptions than that of the estimation problems.

E0203: **Qualitative robustness of divide-and-conquer methods for large data sets**
Presenter: **Andreas Christmann**, University of Bayreuth, Germany

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

EO233 Room Virtual R3 NEW CHALLENGES FOR SPARSE METHODS Chair: Qing Mai

E0182: **An efficient greedy search algorithm for high-dimensional linear discriminant analysis**
Presenter: **Quefeng Li**, University of North Carolina - Chapel Hill, United States

High-dimensional classification is an important statistical problem that has applications in many areas. One widely used classifier is the Linear Discriminant Analysis (LDA). In recent years, many regularized LDA classifiers have been proposed to solve the problem of high-dimensional classification. However, these methods rely on inverting a large matrix or solving large-scale optimization problems to render classification rules: methods that are computationally prohibitive when the dimension is ultra-high. With the emergence of big data, it is increasingly important to develop more efficient algorithms to solve the high-dimensional LDA problem. We propose an efficient greedy search algorithm that depends solely on closed-form formulae to learn a high-dimensional LDA rule. We establish a theoretical guarantee of its statistical properties in terms of variable selection and error rate consistency; in addition, we provide an explicit interpretation of the extra information brought by an additional feature in an LDA problem under some mild distributional assumptions. We demonstrate that this new algorithm drastically improves the computational speed compared with other high-dimensional LDA methods, while maintaining comparable or even better classification performance.

E0270: **Tensor t distribution and tensor response regression**
Presenter: **Ning Wang**, Florida State University, United States

In recent years, promising statistical modeling approaches to tensor data analysis have been rapidly developed. Traditional multivariate analysis tools, such as multivariate regression and discriminant analysis, are generalized from modeling random vectors and matrices to higher-order random tensors. Equipped with tensor algebra and high-dimensional computation techniques, concise and interpretable statistical models and estimation procedures prevail in various applications. One of the biggest challenges to statistical tensor models is the non-Gaussian nature of many real-world data. Unfortunately, existing approaches are either restricted to normality or implicitly using least-squares type objective functions that are computationally efficient but sensitive to data contamination. Motivated by this, we adopt a simple tensor t distribution that is, unlike the commonly used matrix t distributions, compatible with tensor operators and reshaping of the data. We study the tensor response regression with tensor t error, and develop penalized likelihood-based estimation and a novel one-step estimation. We study the asymptotic relative efficiency of various estimators and establish the one-step estimator's oracle properties and near-optimal asymptotic efficiency. We further propose a high-dimensional modification to the one-step estimation procedure and show that it attains the minimax optimal rate in estimation.

E0303: **Sparse composite quantile regression with consistent parameter tuning**
Presenter: **Yuwen Gu**, University of Connecticut, United States
Co-authors: Hui Zou

Composite quantile regression (CQR) provides an efficient estimation of the coefficients in linear models, regardless of the error distributions. We consider penalized CQR for both variable selection and efficient coefficient estimation in a linear model under ultrahigh dimensionality and possibly heavy-tailed error distribution. Both lasso and folded concave penalties are discussed. An L_2 risk bound is derived for the lasso estimator to establish its estimation consistency and the strong oracle property of the folded concave penalized CQR is shown for a feasible solution via the LLA algorithm. Information criteria for selecting the regularization parameter in the folded concave penalized CQR are proposed and shown to be selection consistent. The nonsmooth nature of the penalized CQR poses great numerical challenges for high-dimensional data. We provide a unified and effective numerical optimization algorithm for computing penalized CQR via ADMM. We demonstrate the superior efficiency of penalized CQR estimator, as compared to the penalized least squares estimator, through simulated data under various error distributions.

E0306: **A doubly-enhanced EM algorithm for model-based tensor clustering**
Presenter: **Qing Mai**, Florida State University, United States

Modern scientific studies often collect data sets in the form of tensors, which call for innovative statistical analysis methods. In particular, there is a pressing need for tensor clustering methods to understand the heterogeneity in the data. We propose a tensor normal mixture model (TNMM) approach to enable probabilistic interpretation and computational tractability. The statistical model leverages the tensor covariance structure to reduce the number of parameters for parsimonious modeling, and at the same time explicitly exploits the correlations for better variable selection and clustering. We propose a doubly-enhanced expectation-maximization (DEEM) algorithm to perform clustering under this model. Both the E-step and the M-step are carefully tailored for tensor data in order to account for statistical accuracy and computational cost in high dimensions. Theoretical studies confirm that DEEM achieves consistent clustering even when the dimension of each mode of the tensors grows at an exponential rate of the sample size. Numerical studies demonstrate favorable performance of DEEM in comparison to existing methods.

E0700: **Signed network embedding and its applications to detection of communities and anomalies**
Presenter: **Junhui Wang**, City University of Hong Kong, Hong Kong

Signed networks are frequently observed in real life with additional sign information associated with each edge, yet such information has been largely ignored in existing network models. We will introduce a unified embedding model for signed networks to disentangle the intertwined balance structure and anomaly effect, which can greatly facilitate the downstream analysis, including community detection, anomaly detection, and network inference. The proposed model captures both balance structure and anomaly effect through a low rank plus sparse matrix decomposition, which are jointly estimated via a regularized formulation. Its theoretical properties will be discussed in terms of asymptotic consistency and finite-sample probability bounds for network embedding, community detection and anomaly detection. The advantage of the proposed embedding model is also demonstrated through extensive numerical experiments on both synthetic networks and an international relation network.

EO249 Room Virtual R4 TREND AND CHANGE-POINT ANALYSIS IN TIME SERIES Chair: Kin Wai Chan

E0312: **$L_2 - L_\infty$ inference of breaks for high dimensional time series**
Presenter: **Likai Chen**, Washington University in Saint Louis, United States
Co-authors: Jiaqi Li, Weining Wang, Wei Biao Wu

A new method is proposed for multiple change points detection of high-dimensional time series. The proposed approach targets dense or clustered cross-sectional signals. An L_2-aggregated statistics is adopted within each rectangular window and L_∞ aggregation is applied for all the windows. On the theory front, we develop an asymptotic theory concerning the limiting distributions of the change-point test statistics under both the null and alternatives and the consistency of the estimated break dates. The core of our theory is to extend the high-dimensional Gaussian approximation theorem for L_2-based statistics for dependent data. Weakly temporal and cross-sectional dependences can be allowed. Simulations show the power enhancement in the presence of dense or clustered signals relative to the maximum statistics.

E0315: **Statistical inference for change points in high-dimensional data**
Presenter: **Runmin Wang**, Southern Methodist University, United States
Co-authors: Xiaofeng Shao, Stanislav Volgushev, Changbo Zhu

Estimation and testing of change points in high-dimensional data have wide applications in many disciplines, such as biological science, economics and finance. We introduce a new U-statistic based approach to both problems and show its advantage over several existing methods via theory and simulations. The talk consists of two parts. In the first part, we will introduce a new test based on U-statistics for testing a mean shift in high-dimensional data. The test aims to detect dense alternatives and is tuning parameter-free. At the core of our theory, we show weak convergence of a sequential U-statistic based process and derive the limiting distribution under both the null and alternatives. In the second part, we will discuss a change point location estimator which maximizes a new U-statistic based objective function. Under mild and easily interpretable assumptions, we derive its convergence rate and asymptotic distribution after suitable centering and normalization. A comparison with the popular least-squares based approach illustrates our theoretical advantage. A bootstrap-based approach is also proposed to construct a confidence interval with accurate coverage, which is corroborated by simulation results. We shall illustrate our method using a real data example.

E0542: **Inference of signal variance in time series for mean stationarity test**
Presenter: **Hon Kiu To**, The Chinese University of Hong Kong, Hong Kong
Co-authors: Kin Wai Chan

Inference of mean structure is an important problem in time series analysis. Various tests have been developed to test for different mean structures, including but not limited to, the presence of structural break(s), and parametric mean structures. Many of them are designed under specific mean structures, and may potentially lose power upon violation of such structures. We propose a new mean stationarity test built around the signal variance. The proposed test can detect the non-constancy of the mean function under serial dependence. It is shown to have promising power in detecting hardly noticeable periodic structures. The proposal is further generalized to test for smooth mean structures and the relevant structural changes in time series. A real-data application on global land surface temperature data is presented.

E0321: **Segmenting time series via self-normalization**
Presenter: **Feiyu Jiang**, Fudan University, China
Co-authors: Zifeng Zhao, Xiaofeng Shao

A novel and unified framework is proposed for change-point estimation in multivariate time series. The proposed method is fully nonparametric, enjoys effortless tuning and is robust to temporal dependence. One salient and distinct feature of the proposed method is its versatility, where it allows change-point detection for a broad class of parameters (such as mean, variance, correlation and quantile) in a unified fashion. At the core of our method, we couple the self-normalization (SN) based tests with a novel nested local-window segmentation algorithm, which seems new in the growing literature of change-point analysis. Due to the presence of an inconsistent long-run variance estimator in the SN test, non-standard theoretical arguments are further developed to derive the consistency and convergence rate of the proposed SN-based change-point detection method. Extensive numerical experiments and relevant real data analysis are conducted to illustrate the effectiveness and broad applicability of our proposed method in comparison with state-of-the-art approaches in the literature.

E0539: **A general framework for constructing locally self-normalized multiple-change-point tests**
Presenter: **Kin Wai Chan**, The Chinese University of Hong Kong, Hong Kong
Co-authors: Cheuk Hin Cheng

A general framework is proposed for constructing self-normalized multiple-change-point tests with time-series data. The framework is applicable to a wide class of popular change point detection statistics, including cumulative sum process, outlier-robust rank statistics and order statistics. Neither robust and consistent estimation of nuisance parameters, selection of bandwidth parameters, nor pre-specification of the number of change points is required. The finite-sample performance shows that our proposal is size-accurate, robust against misspecification of the alternative hypothesis, and more powerful than existing methods.

EO401 **Room Virtual R5** **MACHINE LEARNING USING EXPERIMENTAL DESIGN IDEAS** **Chair: Lin Wang**

E0619: **A scalable Gaussian process for large-scale periodic data**
Presenter: **Qian Xiao**, University of Georgia, United States

The periodic Gaussian process (PGP) has been increasingly used in various contexts to model periodic data due to its good performance. Yet, it has such a high computational complexity of $O(n^3)$ (n is the data length) that the application of PGP is often obstructed for processing large-scale periodic data, such as speech signals, vibration signals and periodic motions. To address this challenge, we proposed a circulant PGP (CPGP) model which can greatly accelerate the computations of both parameter estimations and model predictions. The proposed CPGP decomposes the full likelihood into the sum of two computationally scalable composite likelihoods, and its computational complexity is $O(p^2)$, even $O(plog(p))$ for some special cases, where p is a candidate period of PGP which is much smaller than n. Numerical examples are included to show the scalability and the computational efficiency of the proposed CPGP compared to some state-of-the-art methods. Simulation and real case studies are discussed to further illustrate the superiority of the CPGP.

E0764: **Proximal learning for individualized treatment regimes under unmeasured confounding**
Presenter: **Xiaoke Zhang**, George Washington University, United States
Co-authors: Zhengling Qi, Rui Miao

Data-driven individualized decision making has recently received increasing research interest. Most existing methods rely on the assumption of no unmeasured confounding, which unfortunately cannot be ensured in practice, especially in observational studies. Motivated by the recently proposed proximal causal inference, we develop several proximal learning approaches to estimating optimal individualized treatment regimes (ITRs) in the presence of unmeasured confounding. In particular, we establish several identification results for different classes of ITRs, exhibiting the trade-off between the risk of making untestable assumptions and the value function improvement in decision making. Based on these results, we propose several classification-based approaches to finding a variety of restricted in-class optimal ITRs and developing their theoretical properties. The appealing numerical performance of our proposed methods is demonstrated via an extensive simulation study and a real data application.

E0635: **scSampler: Fast diversity-preserving subsampling of large-scale single-cell transcriptomic data**
Presenter: **Nan Miles Xi**, Loyola University Chicago, United States

The number of cells measured in single-cell transcriptomic data has grown fast in recent years. For such large-scale data, subsampling is a powerful and often necessary tool for exploratory data analysis. However, the easiest random subsampling is not ideal from the perspective of preserving rare cell types. Therefore, diversity-preserving subsampling is required for the fast exploration of cell types in a large-scale dataset. We propose scSampler, an algorithm for fast diversity-preserving subsampling of single-cell transcriptomic data. Using simulated and real data, we show that scSampler consistently outperforms existing subsampling methods in terms of both the computational time and the Hausdorff distance between the full and subsampled datasets.

E0914: **Model-robust subdata selection for big data**
Presenter: **Chenlu Shi**, University of California, Los Angeles, United States
Co-authors: Boxin Tang

Subdata selection is necessary because of challenges arising from the statistical analysis of big data using limited computing resources. The existing work on subdata selection relies heavily on a specified model, which calls for an approach that is robust to model misspecification. We propose the use of space-filling designs for subdata selection and examine a fast algorithm for its implementation. The algorithm performs surprisingly well when compared to the reference distribution given by complete search. Simulations are conducted to compare our approach with a recently introduced IBOSS method, and the results show that our method is not just robust to model misspecification but also robust to model uncertainty. While robustness to model misspecification and uncertainty may be expected due to the nature of space-filling designs, we discover that our method enjoys an additional property of robustness when there exist substantial correlations among covariates.

E0727: **Balanced subsampling for big data with categorical predictors**
Presenter: **Lin Wang**, George Washington University, United States

The dramatic growth of big datasets presents a new challenge to data storage and analysis. Data reduction, or subsampling, that extracts useful information from datasets is a crucial step in big data analysis. We will introduce a balanced subsampling approach for big data with categorical predictors. The merits of the proposed approach are two-fold: (i) it is easy to implement and fast; (ii) the selected subsample allows robust effect estimation and prediction. Theoretical results and extensive numerical results show that the proposed approaches are superior to simple random subsampling. The advantages of the balanced subsampling approach are also illustrated through the analysis of real-life examples.

EO365 **Room Virtual R6** **ADVANCED METHODS IN LARGE-SCALE BIOMEDICAL DATA ANALYSIS** **Chair: Bingxin Zhao**

E0468: **Integration of imaging and sequencing data in the context of visual cell sorting**
Presenter: **Gang Li**, University of Washington at Seattle, United States

Visual cell sorting (VCS) is a single-cell co-assay that combines microscopy and high-throughput sequencing. The microscopy measures cell morphology and marks cells with phenotypes of interest, which enables sorting of cells based on visual phenotype. The subsequent sequencing step can be used to measure any one of a variety of cell characteristics, such as gene expression, chromatin accessibility, or chromatin 3D architecture. In the current VCS analysis pipeline, the imaging data is used primarily to generate discrete morphology labels. This approach does not fully exploit the rich information from images. VCS can associate single-cell profiles with their associated morphological phenotypes, but the images and the single-cell profiles do not have a direct correspondence. To attempt to recover this correspondence information, we developed a weakly-supervised manifold alignment algorithm, with the goal of embedding the single-cell sequencing measurements and microscopy images into a shared manifold in such a way that two observations derived from the same cell are nearby in the embedded space. Clearly, successfully creating such an embedding would be valuable because it would allow us to explicitly describe how changes in gene expression relate to specific changes in cell morphology. Our approach sheds light on how gene expression profiles interact with cell morphology.

E0491: **BRIDGE: A novel transcriptome-wide association analysis framework for biomarker identification**
Presenter: **Zhaolong Yu**, Yale University, United States

Transcriptome-wide association studies (TWAS) have several advantages over traditional genome-wide association studies (GWAS) because TWAS performs gene-level association tests by which the multiple testing burden is reduced and association results are more interpretable. We will introduce a novel TWAS method based on joint bounded-variable least-squares. For imputation models, we trained the expression imputation models with genotype and RNA-sequencing data from the updated version of the Genotype-Tissue Expression (GTEx) project. We will show that the imputation accuracy of our method outperformed other state-of-the-art methods. Our pipeline also includes non-coding transcripts by performing specific expression adjustments. Using the gene imputation models, we performed TWAS on a number of complex traits based on their respective GWAS summary statistics and identified novel gene-trait associations.

E0622: **The graphical R2D2 estimator for the precision matrices**
Presenter: **Yan Zhang**, The University of Hong Kong, Hong Kong
Co-authors: Dailin Gan, Guosheng Yin

Biological networks are important for the analysis of human diseases, which summarize the regulatory interactions and other relationships between different molecules. Understanding and constructing networks for molecules, such as DNA, RNA and proteins, can help elucidate the mechanisms of complex biological systems. The Gaussian Graphical Models (GGMs) are popular tools for the estimation of gene regulatory networks because of their biological interpretability. Nonetheless, reconstructing GGMs from high-dimensional datasets is still challenging and current methods cannot handle the sparsity and high-dimensionality issues arising from datasets very well. Here we developed a new GGM, called the graphical R2D2 (R2-induced Dirichlet Decomposition), based on the R2D2 priors for linear models. When the true precision matrix is sparse and of high dimension, the graphical R2D2 provides the estimates with the smallest information divergence from the sampling model. Besides, we also provide a full Gibbs sampler for implementing the graphical R2D2 estimator. We also provide breast cancer gene network analysis using the graphical R2D2 estimator and the important genes recognized from the inferred gene regulatory networks are consistent with biological ground truth.

E0652: **SKPD: A general framework of signal region detection in image regression**
Presenter: **Long Feng**, City University of Hong Kong, Hong Kong

The aim is to present a novel Frequentist framework for signal region detection in high-resolution and high-order image regression problems. Image data and scalar-on-image regression are intensively studied in recent years. However, most existing studies on such topics focused on outcome prediction, while the research on image region detection is rather limited, even though the latter is often more important. We introduce a general framework named Sparse Kronecker Product Decomposition (SKPD) to tackle this issue. The SKPD framework is general in the sense that it works for both matrices (e.g., 2D grayscale images) and (high-order) tensors (e.g., 2D colored images, brain MRI/fMRI data) represented image data. Moreover, unlike many Bayesian approaches, our framework is computationally scalable for high-resolution image problems. Specifically, our framework includes: 1) the one-term SKPD; 2) the multi-term SKPD; and 3) the nonlinear SKPD. The nonlinear SKPD is highly connected to shallow convolutional neural networks (CNN), particular to CNN with one convolutional layer and one fully connected layer. The effectiveness of SKPDs is validated by real brain imaging data in the UK Biobank database.

E0683: **Scalable rare variant meta-analysis of sequencing studies using summary statistics and functional annotations**
Presenter: **Xihao Li**, Harvard T.H. Chan School of Public Health, United States
Co-authors: Zilin Li, Xihong Lin

Large-scale whole-genome/exome sequencing (WGS/WES) studies have enabled the analysis of rare variants (RVs) associated with complex human traits and diseases. Existing RV meta-analysis approaches are not scalable when applied to WGS/WES data. We propose MetaSTAAR, a powerful and resource-efficient RV meta-analysis framework, for large-scale WGS association studies. MetaSTAAR accounts for population structure and relatedness for both continuous and dichotomous traits. By storing LD information of RVs in a new sparse matrix format, the proposed framework is highly storage efficient and computationally scalable for analyzing large-scale WGS/WES data without information loss. Furthermore, MetaSTAAR dynamically incorporates multiple functional annotations to empower RV association analysis, and enables conditional analyses to identify RV-set signals independent of nearby common variants. We applied MetaSTAAR to identify RV-sets associated with four quantitative lipid traits in 30,138 related samples from the NHLBI TOPMed Program Freeze 5 data, consisting of 14 ancestrally diverse studies and 255 million variants in total, as well as the UK Biobank WES data of 200,000 related samples.

EO419 **Room Virtual R7** **INTERVAL-CENSORED FAILURE TIME DATA (FOR JIANGUO SUN 60TH BIRTHDAY)** **Chair: Yang-Jin Kim**

E0324: **Semiparametric analysis of multivariate recurrent events with informative censoring**
Presenter: **Yang Li**, Indiana University School of Medicine, United States
Co-authors: Bin Zhang

In healthcare and clinical studies, recurrent events are frequently encountered both during hospitalizations and after hospital discharge. By recurrent events, we mean that one subject can potentially experience the same type of event repeatedly. In practice, it is common that two or more related types of recurrent events are of interest during the follow-up and thus multivariate recurrent events (MREs) arise. A possible complicating factor in many recurrent event studies is informative censoring. Compared to rate or intensity models, MRE mean functions can be clinically more interpretable especially when the event recurrence is likely fatal. We consider a semiparametric regression analysis on the mean functions with informative censoring. For modeling capacity and flexibility, both additive and multiplicative covariate effects are included. Marginal models will be employed to avoid distributional assumptions or specified correlation structures between MREs and informative censoring. An estimating-equation based inference procedure is developed for both the parametric and nonparametric components. The simulation study shows that the proposed inference procedure performs well. The proposed approach is applied to analyze a motivating dataset from the Mother's Gift Study to evaluate the effectiveness of maternal influenza vaccine and infant pneumococcal conjugate vaccine (PCV7) in reducing infant illnesses in Bangladesh.

E0366: **Optimal subsampling for massive survival data**
Presenter: **HaiYing Wang**, University of Connecticut, United States

With increasingly available on massive survival data, researchers need valid and computationally scalable statistical methods for survival modeling. Existing works focus on relative risk models using the online updating and divide-and-conquer strategies. The focus is on using optimal subsampling algorithms to tackle the computational issues in analyzing survival data. We first discuss the results on fast approximation to the maximum likelihood estimator for a parametric Weibull accelerate failure time model, and then present our current state of knowledge and the challenges of optimal subsampling in the context of semiparametric models with censored data.

E0789: **Prediction accuracy for joint model of interval-censored data and longitudinal markers**
Presenter: **Yang-Jin Kim**, Sookmyung Women University, Korea, South

Joint model for longitudinal marker and time to event data has been an attractive modeling that provides the association between the risk of an event and the change of makers. The interest is to provide appropriate measures for quantifying the prediction accuracy of the prediction model of interval-censored event time with longitudinal markers. Dynamic versions of AUC (Area Under the Curve) and Brier score to reflect updated information are suggested. The performance of the proposed methods under the finite sample is evaluated with simulation and the paquid dataset is analyzed as a real data example.

E0812: **Joint modeling of multivariate longitudinal data and recurrent events: Application to the urea cycle disorders study**
Presenter: **DoHwan Park**, Univ. ov Maryland – Baltimore County, United States

A joint modeling method is developed to analyze the bivariate longitudinal outcomes and time to recurrent events data. We combine the bivariate normal mixed effect model and the frailty model by including the multivariate normal random variables, which account for the dependence among the repeated measures and the dependence between two longitudinal outcomes and recurrent events. We use nonparametric maximum likelihood estimation (NPMLE) to estimate the parameters. EM algorithm was used to compute the NPMLEs and their variance estimators. The results from the simulation studies show that the NPMLEs are noticeably unbiased. The standard error estimators well reflect the true variations of the proposed estimators and the performance is better than individual models. Finally, we apply our procedure to analyzing data from the Urea Cycle Disorders study.

E0872: **Bayesian joint analysis of longitudinal data and interval-censored failure time data**
Presenter: **Lianming Wang**, University of South Carolina, United States
Co-authors: Yuchen Mao, xuemei sui

Joint modeling of longitudinal measurements and survival time has gained great attention in statistics literature in the last two decades. Most of the existing works focus on the joint analysis of the longitudinal response and right-censored survival time. We propose a new frailty model for joint analysis of a longitudinal response and interval-censored survival time. Such data commonly arise in real-life studies where participants are examined at periodical or irregular follow-up times. The proposed joint model has the following appealing properties: (1) the regression coefficients can be interpreted as the marginal effects in both the longitudinal model and the survival model components and (2) the statistical association between the longitudinal response and the survival response can be described and quantified using several association measures in simple explicit forms. The adoption of splines allows us to model the unknown baseline functions with only a finite number of unknown coefficients while providing great modeling flexibility. An efficient Gibbs sampler is developed for posterior computation. Simulation results show that the proposed method performs very well in estimating all the regression parameters and the unknown baseline functions. The proposed method is further illustrated by a real-life application to the patient data from the Aerobics Center Longitudinal Study.

EO133 Room Virtual R8 MACHINE LEARNING FOR MODERN DATA Chair: Hongxiao Zhu

E0341: **Dimension reduction with prior information for knowledge discovery**
Presenter: **Anh Bui**, Virginia Commonwealth University, United States

The focus is on the problem of mapping high-dimensional data to a low-dimensional space, in the presence of other known features. This problem is ubiquitous in science and engineering as there are often controllable/measurable features in most applications. Furthermore, the discovered features in previous analyses can become the known features in subsequent analyses, repeatedly. To solve this problem, a broad class of methods, which is referred to as conditional multidimensional scaling, is proposed. An algorithm for optimizing the objective function of conditional multidimensional scaling is also developed. The proposed framework is illustrated with kinship terms, facial expressions, and simulated car-brand perception examples. These examples demonstrate the benefits of the framework for being able to marginalize out the known features to uncover unknown, unanticipated features in the reduced-dimension space and for enabling a repeated, more straightforward knowledge discovery process. Computer codes for this work are available in the open-source cml R package.

E0360: **Diffusion Schrodinger bridge with applications to score-based generative modeling**
Presenter: **Jeremy Heng**, ESSEC Business School, Singapore
Co-authors: Arnaud Doucet, Valentin De Bortoli, James Thorton

Progressively applying Gaussian noise transforms complex data distributions to approximately Gaussian. Reversing this dynamic defines a generative model. When the forward noising process is given by a Stochastic Differential Equation (SDE), it has been recently demonstrated how the time inhomogeneous drift of the associated reverse-time SDE may be estimated using score-matching. A limitation of this approach is that the forward-time SDE must be run for a sufficiently long time for the final distribution to be approximately Gaussian. In contrast, solving the Schrdinger Bridge problem (SB), i.e. an entropy-regularized optimal transport problem on path spaces, yields diffusions that generate samples from the data distribution in finite time. We present Diffusion SB (DSB), an original approximation of the Iterative Proportional Fitting (IPF) procedure to solve the SB problem, and provide theoretical analysis along with generative modelling experiments. The first DSB iteration recovers the existing methodology with the flexibility of using shorter time intervals, as subsequent DSB iterations reduce the discrepancy between the final-time marginal of the forward (resp. backward) SDE with respect to the prior (resp. data) distribution. Beyond generative modeling, DSB offers a widely applicable computational optimal transport tool as the continuous state-space analogue of the popular Sinkhorn algorithm.

E0641: **Model data heterogeneity with Dirichlet diffusion trees**
Presenter: **Hongxiao Zhu**, Virginia Tech, United States
Co-authors: Shuning Huo

A challenge of modern data analysis is the difficulty to model complex data heterogeneity structures caused by sub-populations or latent factors. We propose a Bayesian latent tree model to characterize data heterogeneity and link the heterogeneity structure with covariates. We adopt Dirichlet Diffusion Trees to model the latent hierarchical data structure underlying the observed data, and propose a regression framework by associating covariates with the parameters of the latent trees. To perform posterior inference, we propose a Markov chain Monte Carlo algorithm to alternatively update the latent tree structures and the regression coefficients. We demonstrate the effectiveness of the model through a simulation study and imaging data on brain Glioblastoma Multiforme images.

E0756: **Learning the data manifold for reusable augmentations**
Presenter: **Kion Fallah**, Georgia Institute of Technology, United States
Co-authors: Marissa Connor, Christopher Rozell

The manifold hypothesis suggests that variations in high-dimensional, real-world data lie on or near a low-dimensional manifold. We discuss recent work to learn this manifold from data by incorporating a generative manifold model in the latent space of a deep auto-encoder. This model represents the manifold with a dictionary of Lie group operators, representing the non-linear path between any two data points with a sparse combination of dictionary entries. To speed up training, we propose an inference procedure that can be quickly run on a GPU. After unsupervised training, we demonstrate that the learned Lie group operators are re-usable across a dataset for generating semantically meaningful augmentations.

E0685: **Minimum L1 interpolators: Precise asymptotics and multiple descent**
Presenter: **Yuting Wei**, University of Pennsylvania, United States

An evolving line of machine learning works observes empirical evidence that suggests interpolating estimators — the ones that achieve zero training error — may not necessarily be harmful. We pursue a theoretical understanding of an important type of interpolator: the minimum L1-norm interpolator, which is motivated by the observation that several learning algorithms favor low L1-norm solutions in the over-parameterized regime. Concretely, we consider the noisy sparse regression model under Gaussian design, focusing on linear sparsity and high-dimensional asymptotics (so that both the number of features and the sparsity level scale proportionally with the sample size). We observe, and provide rigorous theoretical justification for, a curious multi-descent phenomenon; that is, the generalization risk of the minimum L1-norm interpolator undergoes multiple (and possibly more than two) phases of descent and ascent as one increases the model capacity. This phenomenon stems from the special structure of the minimum L1-norm interpolator as well as the delicate interplay between the over-parameterized ratio and the sparsity, thus unveiling a fundamental distinction in geometry from the minimum L2-norm interpolator. Our finding is built upon an exact characterization of the risk behavior, which is governed by a system of two non-linear equations with two unknowns.

EO073 Room Virtual R9 ADVANCES IN NONPARAMETRIC AND SEMIPARAMETRIC PANEL DATA MODELS Chair: Alexandra Soberon

E0220: **Testing beta constancy in asset pricing models**
Presenter: **Luis Antonio Arteaga Molina**, Universidad de Cantabria, Spain
Co-authors: Juan Manuel Rodriguez-Poo

A methodology is proposed for testing coefficients constancy in varying coefficient asset pricing models with endogenous regressors. The testing procedure is defined as a generalized likelihood ratio that focus on the comparison of the restricted and unrestricted sum of squared residuals. As a by product, we have developed a nonparametric method that takes into account the endogenous nature of the regressors to estimate the prices of risk. Resembling the instrumental variable literature, we propose to use a three stages estimation procedure to estimate the varying coefficient; besides we establish the asymptotic properties of the estimators. Finally, we investigate the finite sample properties of our test by means of Monte Carlo experiments study and using critical values and p-values estimated by the bootstrap technique.

E0281: **Estimation of a varying coefficient, fixed-effects Cobb-Douglas production function in levels**
Presenter: **Daniel Henderson**, University of Alabama, United States

A semiparametric varying coefficient estimator is proposed for a Cobb-Douglas production function for panel data with several practical features. First, we estimate the model without a log transformation to avoid induced non-negligible estimation bias. Second, we disentangle the impact of traditional inputs from that of environment variables, which impact output indirectly through altering the output elasticity of inputs and the state of technology via unknown functions. We introduce a linear index structure in the unknown functions to circumvent the curse of dimensionality and allow the output elasticity of different inputs to depend on different environment variables. Third, our technology function accounts for latent heterogeneity across individual units, which can be freely correlated with inputs and/or environment variables. Our estimator combines series and kernel methods for both the unknown parameters and functions. We demonstrate that the proposed estimator exhibits promising finite-sample performance

E0349: **Nonparametric modeling of environmental time series distributions**
Presenter: **Joachim Schnurbus**, University of Passau, Germany
Co-authors: Harry Haupt

A nonparametric kernel-based approach is proposed for modeling the distribution of stochastic processes which may exhibit nonlinearities and non-stationarities driven by trend and/or seasonal patterns. Particular emphasis is placed on providing an approach that is computationally cheap, easy to interpret, allows for the inclusion of multiple seasonality, and is suitable for estimation and forecasting. The approach is demonstrated for a panel of environmental time series.

E0500: **Practical aspects of using quadratic moment conditions in linear dynamic panel data models**
Presenter: **Andrew Adrian Yu Pua**, Xiamen University, China
Co-authors: Markus Fritsch, Joachim Schnurbus

The focus is on the estimation of the lag parameter of linear dynamic panel data models with first-order dynamics based on the quadratic moment conditions. We first show that extending the standard assumptions to allow for mean stationarity and time series homoscedasticity and employing these assumptions in estimation restores standard asymptotics and mitigates the non-standard asymptotic distributions found in the literature. Because using these additional assumptions for estimation purposes may be too restrictive for practical usage, we analyze theoretically an IV estimator based on the quadratic moment conditions and provide a practical data-based approach to detect whether one would face a data generating process that does not suffer from non-standard behavior, while maintaining the default no serial correlation assumption.

E0264: **Measurement errors in panel data regression: A direct estimation approach**
Presenter: **Alexandra Soberon**, Universidad de Cantabria, Spain
Co-authors: Winfried Stute

The estimation of a multiple mismeasured regressor errors-in-variables model with panel data is considered. Using the dependence structure of this data as an additional source of information, we are able to provide a correction for measurement error. More precisely, closed-form two-step estimators are obtained as solutions to estimating equations that exploit the information contained in the second-order moments of the residuals and quasi-residuals obtained by partialling out perfectly measured regressors. Then, the resulting estimators are valid even when distributional assumptions such as the non-normality of the error variables cannot be justified. The asymptotic properties of these estimators are analyzed for both random effects and fixed effects and the finite sample properties are shown via Monte Carlo simulations. Also, the methodology is used in a corporate-finance application of regressions with mismeasured regressors.

Sunday 05.06.2022 10:35 - 12:15 Parallel Session G – EcoSta2022

EO373 Room 101 (Hybrid 1) ADVANCES IN CHANGE-POINT DETECTION METHODS Chair: Ali Shojaie

E0227: **Change point localization in dependent dynamic nonparametric random dot product graph**
Presenter: **Oscar Hernan Padilla**, UCLA, United States

The change point localization problem is studied in a sequence of dependent nonparametric random dot product graphs. To be specific, assume that at every time point, a network is generated from a nonparametric random dot product graph model, where the latent positions are generated from unknown underlying distributions. The underlying distributions are piecewise constant in time and change at unknown locations, called change points. Most importantly, we allow for dependence among networks generated between two consecutive change points. This setting incorporates edge dependence within networks and temporal dependence between networks, which is the most flexible setting in the published literature. To accomplish the task of consistently localizing change points, we propose a novel change point detection algorithm, consisting of two steps. First, we estimate the latent positions of the random dot product model, our theoretical result being a refined version of the state-of-the-art results, allowing the dimension of the latent positions to grow unbounded. Subsequently, we construct a nonparametric version of the CUSUM statistic that allows for temporal dependence. Consistent localization is proved theoretically and supported by extensive numerical experiments, which illustrate the state-of-the-art performance

E0790: **Detection of relevant changes in the frequency domain**
Presenter: **Yan Liu**, Waseda University, Japan

The problem of detecting relevant changes in the frequency domain is considered. The relevant changes are formulated in the framework of nonparametric functionals of the spectral density of the time series in consideration. We propose a consistent test statistic for detecting relevant changes in the frequency domain. Specifically, we construct a new CUSUM statistic of the nonparametric estimator for the spectral density. We also elucidate the consistency of the CUSUM statistic with the relevant change. The CUSUM statistic consisting only of periodograms is not available here because it is not consistent. The proposed statistic has good features such as asymptotic convergence to the Brown bridge, and can be applied to the detection of relevant changes in hidden structures of integer-valued time series. We will also show some numerical examples and applications of this method to the real data based on the above theoretical results.

E1031: **Multiple change point detection in reduced rank high dimensional vector autoregressive models**
Presenter: **George Michailidis**, University of Florida, United States

The focus is on the problem of detecting and locating change points in high-dimensional Vector Autoregressive (VAR) models, whose transition matrices exhibit low rank plus sparse structure. We first address the problem of detecting a single change point using an exhaustive search algorithm and establish a finite sample error bound for its accuracy. Next, we extend the results to the case of multiple change points that can grow as a function of the sample size. Their detection is based on a two-step algorithm, wherein in the first step, an exhaustive search for a candidate change point is employed for overlapping windows, and subsequently, a backwards elimination procedure is used to screen out redundant candidates. The two-step strategy yields consistent estimates of the number and the locations of the change points. To reduce computation cost, we also investigate conditions under which a surrogate VAR model with a weakly sparse transition matrix can accurately estimate the change points and their locations for data generated by the original model. This work also addresses and resolves a number of novel technical challenges posed by the nature of the VAR models under consideration. The effectiveness of the proposed algorithms and methodology is illustrated on both synthetic and two real data sets.

E0844: **Multiple testing of local extrema for detection of structural breaks in linear models**
Presenter: **Dan Cheng**, Arizona State University, United States
Co-authors: Zhibing He, Yunpeng Zhao

A new approach to detect structural breaks (change points) based on differential smoothing and multiple testing is presented for long data sequences modeled as piecewise linear functions plus stationary Gaussian noise. As an application of the STEM algorithm for peak detection, the method detects change points as significant local maxima and minima after smoothing and differentiating the observed sequence. The algorithm, combined with the Benjamini-Hochberg procedure for thresholding p-values, provides asymptotic strong control of the False Discovery Rate (FDR) and power consistency, as the length of the sequence and the size of the jumps or slope changes get large. Simulations show that FDR levels are maintained in non-asymptotic conditions.

EO307 Room 102 (Hybrid 2) RECENT ADVANCES IN ECONOMETRICS AND MACHINE LEARNING Chair: Komsan Suriya

E0169: **Transfer learning under high-dimensional generalized linear models**
Presenter: **Yang Feng**, NYU, United States

The transfer learning problem is studied under high-dimensional generalized linear models (GLMs), which aim to improve the fit of target data by borrowing information from useful source data. Given which sources to transfer, we propose a transfer learning algorithm on GLM and derive its ℓ_1/ℓ_2-estimation error bounds as well as a bound for a prediction error measure. The theoretical analysis shows that under certain conditions, when the target and source are sufficiently close to each other, these bounds could be improved over those of the classical penalized estimator using only target data. When we ignore which sources to transfer, an algorithm-free transferable source detection approach is introduced to detect informative sources. The detection consistency is proved under the high-dimensional GLM transfer learning setting. Extensive simulations and a real-data experiment verify the effectiveness of our algorithms. We summarize R codes for GLM transfer learning algorithms in a new R package glmtrans, which is available on CRAN.

E0278: **Wild bootstrap for instrumental variables regressions with weak and few clusters**
Presenter: **Wenjie Wang**, Nanyang Technological University, Singapore

The wild bootstrap inference is studied for instrumental variable regressions in the framework of a small number of large clusters in which the number of clusters is viewed as fixed and the number of observations for each cluster diverges to infinity. We first show that the wild bootstrap Wald test, with or without using the cluster-robust covariance matrix, controls size asymptotically up to a small error and has power against local alternatives as long as the parameters of endogenous variables are strongly identified in at least one of the clusters. We further develop a wild bootstrap Anderson-Rubin (AR) test for the full-vector inference and show that it controls size asymptotically up to a small error even under weak or partial identification for all clusters. We illustrate the good finite-sample performance of the new inference methods using simulations and provide an empirical application to a well-known dataset about US local labour markets.

E0507: **Small tuning parameter selection for the debiased lasso**
Presenter: **Akira Shinkyu**, Kobe University, Japan

The debiased Lasso has been proposed for statistical inference in high dimensional linear regression models. It needs an estimate of the column vector of the precision matrix to correct the bias of the Lasso, and usually, we get it by the node-wise Lasso. It is common to set the order of the tuning parameter of the node-wise Lasso as $\sqrt{\log p/n}$. However, the debiased Lasso with the tuning parameter has a large bias when the column vector of the precision matrix is not sparse, so the number of nonzero coefficients should be much small such that $o(\sqrt{n}/\log p)$ for asymptotic normality. Motivated by this issue, we show that by setting the order of the tuning parameter of the node-wise Lasso as $1/\sqrt{n}$, the bias of the debiased Lasso can be removed more without making the variance diverge and sparsity conditions on the column vector of the precision matrix. This implies that the debiased Lasso is asymptotically normal even if the number of nonzero coefficients is $o(\sqrt{n/\log p})$, although it may not be efficient. We also propose a tuning parameter selection procedure for the node-wise Lasso.

E0927: **Impact-based budgeting model and the simulation to target sectoral marginal impacts of R&D investment**
Presenter: **Komsan Suriya**, Thailand Science Research and Innovation, Thailand

This model investigates the impacts of R&D investment on sectoral economic growth and GDP growth with the evaluation of targeted marginal impacts of R&D investment in each economic sector. The simulation technique is applied to solve the model. The model begins with the production function and value of production whose factors of production include capital, labor, intermediate factors, estates and terrains and ecological system and climate. It treats the adjustment factor to translate the impacts of R&D investment and R&D related components on the economic growth. In this adjustment factor, R&D investment directly impacts the economy through the advancement of science and technology per se and induces the higher productivities of skilled researchers, unskilled researchers, research facilities and research utilization. Once the ratio of R&D investment in the next period over this period is identified with all other well-simulated variables, adjustment parameters and amplifier parameters, the model presents the results of sectoral growth and overall economic growth. The marginal impact of R&D investment in each sector can be directly calculated by the differences between the value of production in the situations with and without R&D investment. The simulated impacts point to the targeted economic growth in each sector which contribute to the targeted GDP growth. The findings of the model yield the targeted marginal impacts of R&D investment in each economic sector.

EO383 **Room 103 (Hybrid 3)** **RECENT DEVELOPMENTS IN HIGH-DIMENSIONAL DATA ANALYSIS** **Chair: Zhihua Su**

E0208: **Testing the linear mean and constant variance conditions in sufficient dimension reduction**
Presenter: **Yuexiao Dong**, Temple University, United States

Sufficient dimension reduction (SDR) methods characterize the relationship between the response and the predictors through a few linear combinations of the predictors. Sliced inverse regression and sliced average variance estimation are among the most popular SDR methods as they do not involve multi-dimensional smoothing and are easy to implement. However, these inverse regression-based methods require the linear conditional mean (LCM) and(or) the constant conditional variance (CCV) assumption. We propose novel tests to check the validity of the LCM and the CCV conditions through the martingale difference divergence. Extensive simulation studies and a real data application are performed to demonstrate the effectiveness of our proposed tests.

E0319: **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, our graphical model is based on conditional independence given a set of sufficient predictors with a substantially reduced dimension. In this way, we avoid the curse of dimensionality that comes with a high-dimensional kernel. We develop the population-level properties, convergence rate, and variable selection consistency of our estimate. By simulation comparisons and an analysis of the DREAM 4 Challenge data set, we demonstrate that our method outperforms the existing methods when the Gaussian or copula Gaussian assumptions are violated, and its performance remains excellent in the high-dimensional setting.

E0320: **Dimension reduction and data visualization for Frechet regression**
Presenter: **Lingzhou Xue**, Penn State University, United States
Co-authors: Qi Zhang, Bing Li

Frechet regression models provide a promising framework for regression analysis with metric space-valued responses. We introduce a flexible sufficient dimension reduction (SDR) method for Frechet regression to achieve two purposes: to mitigate the curse of dimensionality caused by high-dimensional predictors, and to provide a tool for data visualization for Frechet regression. The approach is flexible enough to turn any existing SDR method for Euclidean (X,Y) into one for Euclidean X and metric space-valued Y. The basic idea is to first map the metric-space valued random object Y to a real-valued random variable $f(Y)$ using a class of functions, and then perform classical SDR to the transformed data. If the class of functions is sufficiently rich, then we are guaranteed to uncover the Frechet SDR space. We showed that such a class, which we call an ensemble, can be generated by a universal kernel. We established the consistency and asymptotic convergence rate of the proposed methods. The finite-sample performance of the proposed methods is illustrated through simulation studies for several commonly encountered metric spaces that include Wasserstein space, the space of symmetric positive definite matrices, and the sphere. We illustrated the data visualization aspect of our method in real applications.

E0807: **Functional directed acyclic graphs**
Presenter: **Bing Li**, The Pennsylvania State University, United States
Co-authors: Kuang-Yao Lee, Lexin Li

A new method is introduced to estimate directed acyclic graphs from multi-variate functional data, based on the notion of faithfulness that relates a directed acyclic graph with a set of conditional independence relations among the random functions. To characterize and evaluate these relations, we propose two linear operators, the conditional covariance operator and the partial correlation operator. Based on these operators, we adapt and extend the PC-algorithm to estimate the functional directed graph, so that the computation time depends on the sparsity rather than the full size of the graph. We study the asymptotic properties of the two operators, derive their uniform convergence rates, and establish the uniform consistency of the estimated graph, all of which are obtained while allowing the graph size to diverge to infinity with the sample size. We demonstrate the efficacy of our method through both simulations and an application to a time-course proteomic dataset.

EO237 Room 104 (Hybrid 4) RECENT ADVANCES IN STATISTICAL INFERENCE Chair: Gourab Mukherjee

E0694: **Motif estimation via subgraph sampling: The fourth-moment phenomenon**
Presenter: **Bhaswar Bhattacharya**, University of Pennsylvania, United States
Co-authors: Sayan Das , Sumit Mukherjee

Network sampling has emerged as an indispensable tool for understanding features of large-scale complex networks where it is practically impossible to search/query over all the nodes. Examples include social networks, biological networks, internet and communication networks, and socio-economic networks, among others. We will discuss a unified framework for statistical inference for counting motifs, such as edges, triangles, and wedges, in the widely used subgraph sampling model. In particular, we will provide precise conditions for the consistency and the asymptotic normality of the natural Horvitz-Thompson (HT) estimator, which can be used for constructing confidence intervals and hypothesis testing for the motif counts. As a consequence, an interesting fourth-moment phenomenon for the asymptotic normality of the HT estimator and connections to fundamental results in random graph theory will emerge.

E0766: **Global testing for dependent Bernoullis**
Presenter: **Sumit Mukherjee**, Columbia University, United States
Co-authors: Nabarun Deb, Rajarshi Mukherjee, Ming Yuan

Suppose $(X_1,\ldots,X_n)$ are independent Bernoulli random variables with $\mathbb{E}(X_i) = p_i$, and we want to test the global null hypothesis that $p_i = \frac{1}{2}$ for all i, versus the alternative that there is a sparse set of size s on which $p_i \geq \frac{1}{2} + A$. The detection boundary of this test in terms of (s,A) is well understood, both in the case when the signal is arbitrary, and when the signal is present in a segment. We study the above questions when the Bernoullis are dependent, and the dependence is modeled by a graphical model (Ising model). In this case, contrary to what typically happens, dependence can allow the detection of smaller signals than in the independent case. This phenomenon happens over a wide range of graphs, for both arbitrary signals and segment signals.

E0874: **On PC adjustments for high dimensional association studies**
Presenter: **Rajarshi Mukherjee**, Harvard T.H. Chan School of Public Health, United States

The focus is on the effect of Principal Component (PC) adjustments while inferring the effects of variables on outcomes. This is motivated by the EIGENSTRAT procedure in genetic association studies where one performs PC adjustment to account for population stratification. We consider simple statistical models to obtain an asymptotically precise understanding of when such PC adjustments are supposed to work. We also verify these results through extensive numerical experiments.

E0962: **Cross-validation for signal denoising**
Presenter: **Sabyasachi Chatterjee**, University of Illinois at Urbana Champaign, United States

A general cross-validation framework for signal denoising is discussed. We will then discuss how to apply this framework to two non parametric regression methods Trend Filtering and Dyadic CART. The resulting cross-validated versions would attain nearly the same rates of convergence as are known for the optimally tuned analogues. There did not exist any previous theoretical analyses of cross validated versions of Trend Filtering or Dyadic CART before this work. This framework is very general and potentially applicable to a wide range of estimation methods which use tuning parameters.

EO295 Room 105 (Hybrid 5) THEORY AND ALGORITHMS FOR HIGH-DIMENSIONAL REGRESSION FOR BIG DATA Chair: Peng Zeng

E0293: **A robust joint model of longitudinal trajectories and time-to-event data at biobank scale**
Presenter: **Hua Zhou**, UCLA, United States
Co-authors: Jin Zhou, Gang Li

Motivated by the analysis of massive electronic health record (EHR) and wearable device data in modern biobanks, a robust and scalable M-estimator, termed the joint model robust estimator (JMRE), is proposed for estimating the accelerated failure time (AFT) model for a right-censored event time jointly with a linear mixed model (LMM) for the longitudinal biomarker trajectory. As a semiparametric estimator, JMRE is robust to distribution misspecification in both AFT and LMM models. It is scalable to biobank data with $10^5 \sim 10^8$ individuals, intensive longitudinal measurements, and a large number of random effects. It can simultaneously model the time-varying effects on both mean and within-subject variance of the longitudinal biomarker. Furthermore, it is easily extensible to data with multiple longitudinal biomarkers.

E0606: **Residual projection for quantile regression**
Presenter: **Nan Lin**, Washington University in St. Louis, United States
Co-authors: Ye Fan

The alternating direction method of multipliers (ADMM) has been a popular solution to the computational challenges for quantile regression in big data. However, its relatively slow convergence can be a bottleneck when communication cost dominates local computational consumption, such as in the Internet of Things (IoT) networks. We propose an alternative technique using residual projection that converges faster. We proved the convergence property of the new technique and further extended it to composite quantile regression (CQR).

E0877: **ODE-on-scalar regression with an application on COVID-19 data**
Presenter: **Peng Zeng**, Auburn University, United States

Since the outbreak in late December 2019, COVID-19 quickly spread around the world. Different countries observed slightly different patterns of how the disease spread within their borders. We focus on the confirmed cases in the first 30 days after the first occurrence was reported in a country. The primary goal is to understand how the spread of COVID-19 is affected by socioeconomic indicators of a country. The spread of COVID-19 is modeled by the Susceptible-Infected-Recovered (SIR) model. The problem is formulated as a regression with the response being a function determined by an ODE system and the predictors being scalars. A Bayesian approach is proposed to fit this ODE-on-scalar regression. A Metropolis-Hasting algorithm is designed to sample from the posterior distributions.

E0993: **Outlier detection in robust regression via chance-constrained programming**
Presenter: **Hao Zhang**, University of Arizona, United States

Outlier detection is a critical step in data analysis to identify heterogeneous points in data. For high dimensional and extremely noisy data, many challenges are posed by outlier points, such as estimating the number of outliers, providing probabilistic confidence statements on identified outliers, fitting robust models against outliers, and achieving high breakdown points with a guarantee. To address these issues, we propose a chance-constrained outlier detection (CCOD) model that integrates robust regression and outlier diagnostics in one unified methodology. Theoretically, we prove that the new method can achieve a high breakdown point. To tackle the nonconvex computational problem, we propose a tractable and scalable convex approximation. Numerical results show that our CCOD model outperforms the state-of-art methodologies in terms of estimation accuracy, robustness, and computational time.

EO367 Room 106 (Hybrid 6) RECENT ADVANCES IN COMPLEX NETWORK ANALYSIS Chair: Yuan Zhang

E0188: **L-2 regularized maximum likelihood for beta-model estimation in large and sparse networks**
Presenter: **Yuan Zhang**, Ohio State University, United States

The beta-model is a powerful tool for modeling networks driven by degree heterogeneity. Its simple yet expressive nature particularly well-suits large and sparse networks, where most models are infeasible due to computational challenge and observation scarcity. However, existing algorithms for beta-model do not scale up; and theoretical understandings remain limited to dense networks. Several major improvements to the method and theory of -model are brought to address urgent needs of practical applications. The contributions include: 1. method: we propose a new L-2 penalized MLE scheme; we design a novel algorithm that can comfortably handle sparse networks of millions of nodes, much faster and more memory-parsimonious than any existing algorithm; 2. theory: we present new error bounds on beta-models under much weaker assumptions; we also establish new lower-bounds and new asymptotic normality results; distinct from existing literature, our results cover both small and large regularization scenarios and reveal their distinct asymptotic dependency structures; 3. application: we apply our method to large COVID-19 network data sets and discover meaningful results.

E0631: **Heterogeneous block covariance model for community detection**
Presenter: **Xiang Li**, The George Washington University, United States
Co-authors: Yunpeng Zhao, Qing Pan, Ning Hao

Community detection is a clustering method based on objects' pairwise relationships such that objects classified in the same group are more densely connected than objects from different groups. Most of the model-based community detection methods such as the stochastic block model and its variants are designed for networks in which the connections between nodes are described by discrete values, which ignores the practical scenarios where the pairwise relationships between nodes can be continuous. The heterogeneous block covariance model (HBCM) proposes a novel clustering structure applicable to signed and continuous connections between nodes such as a covariance matrix, taking into account the characteristics of each individual object additional to the group-level information, and uses the variational EM algorithm to estimate the optimal group membership and parameters. The statistical property of the HBCM is studied and its practical performance is demonstrated by extensive numerical simulations. The HBCM is applied to the yeast gene expression data.

E0869: **Classically boosted network embeddings**
Presenter: **Joel Nishimura**, Arizona State University, United States
Co-authors: Yunpeng Zhao

Network embeddings are a popular and effective preprocessing step when performing machine learning with network data. We demonstrate that standard boosting techniques, AdaBoost and Real AdaBoost can be applied to network embedding techniques to increase performance, particularly in terms of link prediction on test data in a cross-validation context. These approaches produce results competitive with other state-of-the-art embedding approaches when applied to a number of empirical networks. Additionally, we show on simulated data that Real AdaBoost can de-aggregate some networks, wherein networks created by two independent latent features can have those separate latent features inferred by different boosted rounds. Further analysis of the performance of these boosted methods shows that they retain the characteristic robustness to over-fitting as boosting methods in classical settings.

E0920: **An autoregressive beta-model for dynamic networks**
Presenter: **Binyan Jiang**, The Hong Kong Polytechnic University, Hong Kong

An autoregressive model is proposed which directly depicts the dynamic change of the edges over time for dynamic network processes with degree heterogeneity. The model facilitates efficient statistical inference such as the maximum likelihood estimators and can be consistently estimated as long as the number of network observations is great or equal to 2. Some preliminary theoretical results and numerical results will be presented to illustrate the promising performance of the proposed model

EO281 Room 107 (Hybrid 7) ADVANCES IN FUNCTIONAL DATA AND NETWORKS ANALYSIS Chair: Jeng-Min Chiou

E0581: **Modeling time-varying random objects and dynamic networks**
Presenter: **Paromita Dubey**, University of Southern California, United States

Samples of dynamic or time-varying networks and other random object data such as time-varying probability distributions are increasingly encountered in modern data analysis. Common methods for time-varying data such as functional data analysis are infeasible when observations are time courses of networks or other complex non-Euclidean random objects that are elements of general metric spaces. In such spaces, only pairwise distances between the data objects are available and a strong limitation is that one cannot carry out arithmetic operations due to the lack of an algebraic structure. We combat this complexity by a generalized notion of mean trajectory taking values in the object space. For this, we adopt pointwise Fréchet means and then construct pointwise distance trajectories between the individual time courses and the estimated Fréchet mean trajectory, thus representing the time-varying objects and networks by functional data. Functional principal component analysis of these distance trajectories can reveal interesting features of dynamic networks and object time courses and is useful for downstream analysis. The utility of the proposed methodology is illustrated with dynamic networks, time-varying distribution data and longitudinal growth data.

E0736: **Inference for high-dimensional exchangeable arrays with an application to network data**
Presenter: **Kengo Kato**, Cornell University, United States
Co-authors: Yuya Sasaki, Harold Chiang

Inference is considered for high-dimensional separately and jointly exchangeable arrays where the dimensions may be much larger than the sample sizes. For both exchangeable arrays, we first derive high-dimensional central limit theorems over the rectangles and subsequently develop novel multiplier bootstraps with theoretical guarantees. These theoretical results rely on new technical tools such as Hoeffding-type decomposition and maximal inequalities for the degenerate components in the Hoeffiding-type decomposition for the exchangeable arrays. We exhibit applications of our methods to uniform confidence bands for density estimation under joint exchangeability and penalty choice for L1-penalized regression under separate exchangeability. Extensive simulations demonstrate precise uniform coverage rates. We illustrate by constructing uniform confidence bands for international trade network densities.

E0454: **Basis expansions for functional snippets**
Presenter: **Qixian Zhong**, Xiamen University, China
Co-authors: Zhenhua Lin, Jane-Ling Wang

Estimation of mean and covariance functions is fundamental for functional data analysis. While this topic has been studied extensively in the literature, a key assumption is that there are enough data in the domain of interest to estimate both the mean and covariance functions. Mean and covariance estimation is investigated for functional snippets in which observations from a subject are available only in an interval of length strictly, and often much, shorter than the length of the whole interval of interest. For such a sampling plan, no data is available for direct estimation of the off-diagonal region of the covariance function. This challenge is tackled via a basis representation of the covariance function. The proposed estimator enjoys a convergence rate that is adaptive to the smoothness of the underlying covariance function, and has superior finite-sample performance in numerical studies.

E0793: **Online estimation for functional data**
Presenter: **Fang Yao**, Peking University, China
Co-authors: Ying Yang

Functional data analysis has attracted considerable interest and is facing new challenges, one of which is the increasingly available data in a streaming manner. We present an online nonparametric method to dynamically update the estimates of mean and covariance functions for functional data. The kernel-type estimates can be decomposed into two sufficient statistics depending on the data-driven bandwidths. We propose to approximate the future optimal bandwidths by a sequence of dynamically changing candidates and combine the corresponding statistics across blocks to form the updated estimation. The proposed online method is easy to compute based on the stored sufficient statistics and the current data block. We derive the asymptotic normality and, more importantly, the relative efficiency lower bounds of the online estimates of mean and covariance functions. This provides insight into the relationship between estimation accuracy and computational cost driven by the length of the candidate bandwidth sequence. Simulations and real data examples are provided to support such findings.

EO363 Room Virtual R1 RECENT ADVANCES IN STATISTICAL LEARNING AND STATISTICAL MODELING Chair: Jiwei Zhao

E0224: **Optimal estimation of average treatment effect on the treated under endogeneous treatment assignment**
Presenter: **Trinetri Ghosh**, University of Wisconsin-Madison, United States
Co-authors: Menggang Yu, Jiwei Zhao

When evaluating a complex intervention, instead of average treatment effect (ATE), researchers are more interested in the average treatment effect on the treated (ATT), which is the quantity most relevant to policymakers. We consider the ATT estimation motivated by a case study, where the treatment assignment might depend on the potential untreated outcome and hence is endogenous. We study the scenario that the ATT can be identified. We investigate the optimal estimation of ATT by characterizing the geometric structure of the model. We derive the semiparametric efficiency bound for ATT estimation and propose an estimator that can achieve this bound. Consistency and asymptotic normality of the proposed estimator are established. The finite-sample performance of the proposed estimator is studied through comprehensive simulations and an application to our motivated study.

E0209: **Group network Hawkes process**
Presenter: **Ganggang Xu**, University of Miami, United States
Co-authors: Guanhua Fang, Haochen Xu, Xuening Zhu, Yongtao Guan

The purpose is to study the event occurrences of individuals interacting in a network. To characterize the dynamic interactions among the individuals, we propose a group network Hawkes process (GNHP) model whose network structure is observed and fixed. In particular, we introduce a latent group structure among individuals to account for the heterogeneous user-specific characteristics. A maximum likelihood approach is proposed to simultaneously cluster individuals in the network and estimate model parameters. A fast EM algorithm is subsequently developed by utilizing the branching representation of the proposed GNHP model. Theoretical properties of the resulting estimators of group memberships and model parameters are investigated under both settings when the number of latent groups G is over-specified or correctly specified. A data-driven criterion that can consistently identify the true G under mild conditions is derived. Extensive simulation studies and an application to a data set collected from Sina Weibo are used to illustrate the effectiveness of the proposed methodology.

E0340: **Causal inference via artificial neural networks: From prediction to causation**
Presenter: **Shujie Ma**, University of California-Riverside, United States

Recent technological advances have created numerous large-scale datasets in observational studies, which provide unprecedented opportunities for evaluating the effectiveness of various treatments. Meanwhile, the complex nature of large-scale observational data post great challenges to the existing conventional methods for causality analysis. We will introduce a new unified approach that we have proposed for efficiently estimating and inferring causal effects using artificial neural networks. We develop a generalized optimization estimation through moment constraints with the nuisance functions approximated by artificial neural networks. This general optimization framework includes the average, quantile and asymmetric least squares treatment effects as special cases. The proposed methods take full advantage of the large sample size of large-scale data and provide effective protection against misspecification bias while achieving dimensionality reduction. We also show that the resulting treatment effect estimators are supported by reliable statistical properties that are important for conducting causal inference.

E0210: **Distribution-invariant differential privacy**
Presenter: **Xuan Bi**, University of Minnesota, United States
Co-authors: Xiaotong Shen

Differential privacy is becoming one gold standard for protecting the privacy of publicly shared data. It has been widely used in social science, data science, public health, information technology, and the U.S. decennial census. Nevertheless, to guarantee differential privacy, existing methods may unavoidably alter the conclusion of original data analysis, as privatization often changes the sample distribution. This phenomenon is known as the trade-off between privacy protection and statistical accuracy. We break this trade-off by developing a distribution-invariant privatization (DIP) method to reconcile both high statistical accuracy and strict differential privacy. As a result, any downstream statistical or machine learning task yields essentially the same conclusion as if one used the original data. Numerically, under the same strictness of privacy protection, DIP achieves superior statistical accuracy in two simulations and on three real-world benchmarks.

EO361 Room Virtual R10 SEMIPARAMETRIC METHODS FOR CAUSAL INFERENCE Chair: Kendrick Li

E0335: **Estimating heterogeneous treatment effects with right-censored data via causal survival forests**
Presenter: **Yifan Cui**, National University of Singapore, Singapore
Co-authors: Michael Kosorok, Erik Sverdrup, Stefan Wager, Ruoqing Zhu

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

E0486: **Semiparametric efficient G-estimation with invalid instrumental variables**
Presenter: **BaoLuo Sun**, National University of Singapore, Singapore
Co-authors: Zhonghua Liu, Eric Tchetgen Tchetgen

Mendelian randomization leverages one or multiple genetic markers as instrumental variables for causal inference in the presence of possible unmeasured confounding. In order to improve efficiency, multiple genetic markers are routinely used, leading to concerns about bias due to possible violation of exclusion restriction of no direct effect of any instrument on the outcome other than through the exposure in view. To address this concern, we introduce a new class of g-estimators that are guaranteed to remain consistent for the causal effect of interest provided that a set of at least k out of K candidate instrumental variables are valid, for some k less than or equal to K set by the analyst ex-ante, without necessarily knowing the identity of the valid and invalid instruments. We provide formal semiparametric theory supporting the results and characterize the semiparametric efficiency bound for the exposure causal effect which cannot be improved upon by any regular estimator with our favorable robustness property. Both simulation studies and applications to the UK Biobank data demonstrate the superior empirical performance of our estimators compared to competing methods.

E0866: **Double negative control inference in test-negative design studies of vaccine effectiveness**
Presenter: **Kendrick Li**, University of Michigan Department of Biostatistics, United States
Co-authors: Xu Shi, Wang Miao, Eric Tchetgen Tchetgen

The test-negative design (TND) has become a standard approach to evaluating vaccine effectiveness against the risk of acquiring infectious diseases in real-world settings, such as Influenza, Rotavirus, Dengue fever, and more recently COVID-19. In a TND study, individuals who experience symptoms and seek care are recruited and tested for the infectious disease which defines cases and controls. Despite TND's potential to reduce unobserved differences in healthcare-seeking behavior (HSB) between vaccinated and unvaccinated subjects, it remains subject to various potential biases. First, residual confounding bias may remain due to unobserved HSB, occupation as a healthcare worker, or previous infection history. Second, because selection into the TND sample is a common consequence of infection and HSB, collider stratification bias may exist when conditioning the analysis on testing, which further induces confounding by latent HSB. We present a novel approach to identifying and estimating vaccine effectiveness in the target population by carefully leveraging a pair of negative control exposure and outcome variables to account for potential hidden bias in TND studies. We illustrate our proposed method with extensive simulation and an application to study COVID-19 vaccine effectiveness using data from the University of Michigan Health System.

E0852: **Semiparametric data fusion with external summary statistics**
Presenter: **Wang Miao**, Peking University, China

Using external summary statistics to improve the efficiency of internal data analysis has attracted a lot of attention in recent years. Efficient estimation using external summary statistics has been developed for parametric models. We establish the theory for efficient estimation of a general functional under semiparametric or nonparametric models when combining internal individual-level data with external summary statistics. We propose an estimator that can achieve this efficiency bound. The estimator is also robust against the bias of external summary statistics when they are obtained with biased data whose distribution departs from the internal data. We illustrate with simulations and apply our methods to a Helicobacter pylori infection dataset.

EO407 Room Virtual R11 ECOLOGICAL STATISTICS Chair: Wen-Han Hwang

E0594: **A general algorithm for error-in-variables modelling using monte carlo expectation maximization**
Presenter: **Jakub Stoklosa**, University of New South Wales, Australia

Measurement error models are often needed to correct for uncertainty arising from measurements of covariates in regression modeling. The literature on measurement error modeling is plentiful, however, general algorithms and software for maximum likelihood estimation of models with measurement error are not as readily available, in a form that they can be used by applied researchers without relatively advanced statistical expertise. We develop a novel algorithm for errors-in-variables modeling, which could in principle take any regression model fitted by maximum likelihood, or penalized likelihood, and extend it to account for uncertainty in predictor variables. This is achieved by exploiting an interesting property of the Monte Carlo Expectation-Maximization (MCEM) algorithm, namely that it can be expressed as an iteratively reweighted maximization of complete data likelihoods (formed by imputing the missing values). Thus we can take any regression model for which we have an algorithm for (penalized) likelihood estimation when predictors are error-free, nest it within our proposed iteratively reweighted (MCEM) algorithm, and thus account for uncertainty in predictor variables. The approach is demonstrated in various ecological examples. Because the method uses maximum (penalized) likelihood, it inherits advantageous optimality and inferential properties, as illustrated by simulation. We also study the robustness to some violations in predictor distributional assumptions.

E0623: **Multispecies occupancy modelling of New Zealand bird species**
Presenter: **Nokuthaba Sibanda**, Victoria University of Wellington, New Zealand

Birds New Zealand is undertaking the third Bird Atlas for the country, which started on the 1st of June 2019 and will end on 31 May 2024. Single species occupancy models were applied to analyse data from the previous two atlases (1969-1979 and 1999-2004) to create maps of occupancy estimates for 64 bird species. We illustrate the application of multi-species occupancy models for New Zealand citizen science data. To provide more recent estimates of occupancy, we use data from the eBird program to cover the years 2015 to 2019. eBird data were first processed to make them suitable for occupancy models and to be consistent with the NZ Bird Atlas sampling design. We focused on three species and restricted the study region to the Greater Wellington area. We considered one introduced species (Platycercus eximius) and two native species (Cyanoramphus novaezelandiae and Petroica macrocephala). Single season occupancy models were first fitted to each of the three species in turn. A spatial component was included using restricted spatial regression. Multispecies occupancy models were then fitted to pairs of species and then for all three species. All models were fitted using a Bayesian approach in the R package rStan. Logistic distribution priors were used for the model coefficients. Results were presented using maps of occupancy probabilities for each species and for different pairs of species, with presence observations overlaid.

E0625: **A machine learning method for estimating the probability of presence using presence-background data**
Presenter: **Yan Wang**, RMIT University, Australia
Co-authors: Chathuri Samarasekara, Lewi Stone

Estimating the prevalence or the absolute probability of the presence of a species from presence-background data has become a controversial topic in species distribution modelling. We propose a new method by combining both statistics and machine learning algorithms that help overcome some of the known existing problems. We have also revisited the popular but highly controversial Lele and Keim (LK) method by evaluating its performance and assessing the RSPF condition it relies on. Simulations show that the LK method with unfounded model assumptions would render a fragile estimation/prediction of the desired probabilities. Rather we propose the local knowledge condition, which relaxes the pre-determined population prevalence condition that has so often been used in much of the existing literature. Simulations demonstrate the performance of the CLK method utilising the local knowledge as an assumption to successfully estimate the probability of presence. The local knowledge extends the local certainty or the prototypical presence location assumption and has significant implications for demonstrating the necessary condition for identifying absolute (rather than relative) probability of presence without absence data in species distribution modelling.

E0620: **On the occupancy models with time-to-detection data**
Presenter: **Wen-Han Hwang**, National Chung Hsing University, Taiwan

The detection time occupancy model introduced as an alternative method can estimate occupancy parameters even with a single visit, which is beneficial to improving the estimation efficiency of the common presence/absence occupancy model. However, the efficiency obtained from this method is only based on some examples and simulation studies in the literature. Compared with the presence/absence model, we derive the asymptotic relative efficiency and conduct simulation studies to strictly evaluate the effectiveness of the exponential detection time model. We also propose a mixed exponential detection time model, which can be linked to a negative binomial model if count data is available. The results obtained by analyzing the data of 63 bird species collected in the Karoo region of South Africa show that the proposed detection time model is generally more suitable for the data than the presence/absence model and the exponential detection time model.

EO345 Room Virtual R12 MODERN STATISTICAL METHODS FOR EXPERIMENTAL DESIGN Chair: Ming-Chung Chang

E0662: **Multi-fidelity surrogate modeling with confidence: Stacking experimental design with cost complexity guarantees**
Presenter: **Chih-Li Sung**, Michigan State University, United States

In an era where scientific experimentation is costly, multi-fidelity emulation (i.e., predictive modeling using data of multiple fidelities, or accuracies) is becoming a crucial tool for scientific discovery. Such emulators allow experimenters to maximize predictive power and thus scientific inference given an experimental budget. There has, however, been little work exploring the problems of design and sample size determination for multi-fidelity emulation, both of which are critical for cost-efficient predictive modeling. We thus propose a novel experimental design framework that addresses both problems under a novel multi-level emulator model. We prove a novel complexity theorem that shows, under the proposed sequential design, that the resulting emulator achieves a prediction accuracy given a computational cost. We then demonstrate the effectiveness of the proposed sequential design in a suite of simulation experiments and an application to finite-element analysis.

E0583: **Uniform projection designs and strong orthogonal arrays**
Presenter: **Cheng-Yu Sun**, National Tsing Hua University, Taiwan
Co-authors: Boxin Tang

The connections between uniform projection designs and strong orthogonal arrays of strength 2+ are explored. Both of these classes of designs are suitable designs for computer experiments and space-filling in two-dimensional margins, but they are motivated by different considerations. Uniform projection designs are introduced to capture two-dimensional uniformity using the centered L2-discrepancy whereas strong orthogonal arrays of strength 2+ are brought forth by He, Cheng, and Tang as they achieve stratifications in two-dimensions on finer grids than ordinary orthogonal arrays. We first derive a new expression for the centered L2 discrepancy, which gives a decomposition of the criterion into a sum of squares where each square measures one aspect of design uniformity. This result is not only insightful in itself but also allows us to study strong orthogonal arrays in terms of the discrepancy criterion. More specifically, we show that strong orthogonal arrays of strength 2+ are optimal or nearly optimal under the uniform projection criterion.

E0820: **Simulator selection with applications in cell biology**
Presenter: **Li-Hsiang Lin**, Louisiana State University, United States

Computer simulators are widely used for the study of complex systems. In many applications, there are multiple simulators available with different scientific interpretations of the underlying mechanism, and the goal is to identify an optimal simulator based on the observed physical experiments. To achieve the goal, we propose a selection criterion based on leave-one-out cross-validation. This criterion consists of a goodness-of-fit measure and a generalized degree of freedom penalizing the simulator's sensitivity to perturbations in the physical observations. Asymptotic properties of the selected optimal simulator are discussed. It is shown that the proposed procedure includes a conventional calibration method as a special case. The finite sample performance of the proposed procedure is demonstrated through numerical examples. In the application of cell biology, an optimal simulator is selected, which can shed light on the T cell recognition mechanism in the human immune system.

E0564: **Generating optimal order-of-addition designs with flexible run sizes**
Presenter: **Shin-Fu Tsai**, National Taiwan University, Taiwan

In some industrial, chemical and pharmaceutical studies, physical and/or chemical properties of process outputs can be very different depending on the order in which materials are added. A series of trials conducted for exploring the impact of sequentially adding the materials in various orders is called an order-of-addition experiment. We will introduce a new method to construct optimal designs for these kinds of experiments. The main idea of the proposed method is to generate an order-of-addition design by juxtaposing several isotopic Latin squares. First, a computer-assisted search procedure will be introduced to generate optimal designs with small to moderate run sizes. On the basis of computer-generated designs, a recursive method will be applied to construct optimal designs with large run sizes. By combining these approaches, a series of new optimal designs can be generated for future work.

EO153 Room Virtual R13 STATISTICAL APPROACHES FOR FUNCTIONAL OBSERVATIONS Chair: Ci-Ren Jiang

E0248: **Eigen-adjusted functional principal component analysis**
Presenter: **Ci-Ren Jiang**, Academia Sinica, Taiwan
Co-authors: Eardi Lila, John Aston, Jane-Ling Wang

Functional Principal Component Analysis (FPCA) has become a widely-used dimension reduction tool for functional data analysis. When additional covariates are available, existing FPCA models integrate them either in the mean function or in both the mean function and the covariance function. However, methods of the first kind are not suitable for data that display second-order variation, while those of the second kind are time-consuming and make it difficult to perform subsequent statistical analyses on the dimension-reduced representations. To tackle these issues, we introduce an eigen-adjusted FPCA model that integrates covariates in the covariance function only through its eigenvalues. In particular, different structures on the covariate-specific eigenvalues – corresponding to different practical problems – are discussed to illustrate the model's flexibility as well as utility. To handle functional observations under different sampling schemes, we employ local linear smoothers to estimate the mean function and the pooled covariance function, and a weighted least square approach to estimate the covariate-specific eigenvalues. The convergence rates of the proposed estimators are further investigated under the different sampling schemes. In addition to simulation studies, the proposed model is applied to functional Magnetic Resonance Imaging scans, collected within the Human Connectome Project, for functional connectivity investigation.

E0309: **Functional principal component analysis of cointegrated functional time series**
Presenter: **Won-Ki Seo**, University of Sydney, Australia

Functional principal component analysis (FPCA) has played an important role in the development of functional time series analysis. The purpose is to investigate how FPCA can be used to analyze cointegrated functional time series and proposes a modification of FPCA as a novel statistical tool. Our modified FPCA not only provides an asymptotically more efficient estimator of the cointegrating vectors but also leads to novel FPCA-based tests for examining some essential properties of cointegrated functional time series. As an empirical illustration, our methodology is applied to two empirical examples: U.S. age-specific employment rates and earning densities.

E0496: **Geometrically weighted compositional data analysis for forecasting life-table death counts**
Presenter: **Han Lin Shang**, Macquarie University, Australia

Age-specific life-table death counts observed over time are examples of densities. Non-negativity and summability are two constraints that prevent the direct implementation of standard statistical methods. Compositional data analysis presents a one-to-one mapping from constrained to unconstrained space to rectify the constraints. We introduce a weighted compositional data analysis for modeling and forecasting life-table death counts. Our proposed method assigns higher weights to more recent data and provides a modeling scheme that is easily adapted to allow for constraints. We illustrate our method using Swedish age-specific life-table death counts from 1751 to 2020 and show that the weighted compositional data analytic method improves forecast accuracy compared to their unweighted counterparts.

E0737: **Partially linear models for functional data**
Presenter: **Ming-Yueh Huang**, Academia Sinica, Taiwan

A class of partially linear models is introduced to predict a scalar response using functional covariates. Different from existing approaches, the models allow the non-linear part to be fully nonparametric. To avoid the curse of dimensionality, we further apply a dimension reduction technique to detect detailed structures of the non-linear part. To estimate the introduced model, an iterative method is proposed. In this method, a novel gradient-based estimation is proposed to estimate the dimension reduction subspace of the non-linear part. We will also discuss the special cases when some of the covariates are multivariate.

EO031 **Room Virtual R2** MULTIDIMENSIONAL/MULTIMODAL NEUROIMAGING DATA ANALYSIS **Chair: Yi Zhao**

E0184: **Statistical modeling issues in brain age prediction**
Presenter: **Fengqing Zhang**, Drexel University, United States

Brain age prediction based on neuroimaging data and machine learning models has emerged as a promising approach for characterizing typical brain development and neuropsychiatric disorders. However, few studies examine multi-modal imaging features derived from MRI, DTI as well as rs-fMRI for brain age prediction. In addition, several studies report that the predicted brain age is underestimated for older subjects and overestimated for younger subjects. We examine this systematic bias and propose different approaches to correct for the bias. We also compare different machine learning models to integrate different combinations of multi-modal imaging features. Current methods of brain age prediction provide a single value representing the whole brains average developmental or ageing status. This approach might miss the divergent patterns of change in various brain structures. We, therefore, propose a novel multidimensional brain-age index. The proposed methods are evaluated using large neuroimaging data sets.

E0225: **Probabilistic joint and individual variation explained**
Presenter: **Benjamin Risk**, Emory University, United States
Co-authors: Raphiel Murden, Gavin Tian, Deqiang Qiu

Collecting multiple types of data on the same set of subjects is common in modern scientific applications including genomics, metabolomics, and neuroimaging. Joint and Individual Variation Explained (JIVE) seeks a low-rank approximation of the joint variation between two or more sets of features captured on common subjects and isolates this variation from that unique to each set of features. We propose a probabilistic model for the JIVE framework with subject random effects and develop an expectation-maximization (EM) algorithm to estimate the parameters of interest. The model extends probabilistic PCA to multiple data sets. Extensive simulation studies show that ProJIVE achieves greater accuracy compared to other methods and is robust to model misspecification. We apply ProJIVE to measures of brain morphometry and cognition from the Alzheimer's Disease Neuroimaging Initiative. ProJIVE learns biologically meaningful sources of variation in brain morphometry and cognition. The joint morphometry and cognition subject scores are related to existing biomarkers.

E0234: **Neurodevelopment subtyping via multidimensional brain functional connectomes**
Presenter: **Yize Zhao**, Yale University, United States

Individual differences in neurodevelopment contribute to a broad range of psychiatric disorders. Regardless of the precise mechanism or behavior, the underlying assumption of all neurodevelopmental models of risk is that at the population level, there exist subgroups of individuals that share similar patterns of neural function and development and these subgroups reflect different risk profiles. However, the presence of multiple neurodevelopmental subgroups, as defined by brain functional connectivity has not been assessed previously. We propose a nonparametric Bayesian clustering scheme based on brain functional connectomes across both resting and task states, which dissects subtypes integrating multidimensional brain functional organizations. We simultaneously learn the network parcellations under each cognitive construct and identify the informative subnetwork units playing the key roles during subtyping. To facilitate posterior inference, we develop an efficient variational Bayes algorithm that allows the practical use of the proposed network clustering model. After intensive simulations, we apply the method to the motivated Adolescent Brain Cognitive Development (ABCD) Study and identified four network-driven neurodevelopment subtypes, which are verified by their distinct cognitive profiles.

E0616: **Multidimensional/multimodal neuroimaging data analysis**
Presenter: **Brian Caffo**, Johns Hopkins University, United States

Neuroimaging and neuroscience measurement is intrinsically multidimensional and complex. In this session, we will discuss novel methodology in this exciting and dynamic area. I will be the discussant for the session.

EO077 **Room Virtual R3** CAUSAL INFERENCE **Chair: Yen-Tsung Huang**

E0516: **Causal inference, competing events, and mechanism**
Presenter: **Jessica Young**, Harvard Medical School and Harvard Pilgrim Health Care Institute, United States

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

E0984: **Disentangling confounding and nonsense associations due to dependence**
Presenter: **Elizabeth Ogburn**, Johns Hopkins University, United States

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

E0997: **Identification and estimation of natural mediation effect in the presence of treatment induced confounding**
Presenter: **Kwun Chuen Gary Chan**, University of Washington, United States
Co-authors: Fan Xia

Natural mediation effects are desirable estimands for studying causal mechanisms in a population, but complications arise in estimating natural indirect effects in the presence of treatment-induced confounding. For instance, the usual sequential ignorability assumption no longer guarantees identification. We will consider additional assumptions needed to guarantee identifiability and its relationship to interventional effects. Moment-type and locally efficient estimators will be presented.

E1005: **From linear structural equation modeling to generalized multiple mediation formula**
Presenter: **Sheng-Hsuan Lin**, Institute of Statistics, Taiwan

Causal mediation analysis is advantageous for mechanism investigation. In settings with multiple causally ordered mediators, path-specific effects (PSEs) have been introduced to specify the effects of certain combinations of mediators. However, most PSEs are unidentifiable. The interventional analogue of PSE (iPSE) is adapted to address the non-identifiability problem. Moreover, previous studies only focused on cases with two or three mediators due to the complexity of the mediation formula in a large number of mediators. We provide a generalized definition of traditional PSEs and iPSEs with a recursive formula, along with the required assumptions for nonparametric identification. The three major contributions are: First, we develop a general approach for causal mediation analysis with an arbitrary number of multiple ordered mediators and with time-varying confounders. Second, we demonstrate identified formula of iPSE is a general form of previous mediation analysis. It is reduced to a linear structural equation model under a linear or log-linear model, to causal mediation formula when only one mediator. Third, a flexible algorithm built based on the g-computation algorithm is proposed along with user-friendly software online. All methods and software contribute to comprehensively decomposing a causal effect confirmed by data science and help to disentangle causal mechanisms when multiple ordered mediators exist, which make the natural pathways complicated.

EO243 Room Virtual R4 NEW IDEAS IN EMPIRICAL BAYES Chair: Sihai Zhao

E0580: **Adventures in sparsity and shrinkage with the normal means model**
Presenter: **Matthew Stephens**, University of Chicago, United States

The normal means model has been the canonical model for illustrating the ideas and benefits of shrinkage estimation and has been the subject of considerable theoretical study. By comparison, practical applications of the normal means model are relatively rare, and it has generally been overshadowed by methods like L1-regularization as a way of inducing sparsity. We argue that this should change: we describe some recently-developed Empirical Bayes ways to solve the normal means model and describe how they can be applied to induce shrinkage, sparsity and smoothness in a range of practical applications, including False Discovery Rates, non-parametric regression, sparse regression, and sparse principal components analysis or factor analysis.

E0730: **Covariate-powered empirical Bayes estimation**
Presenter: **Nikolaos Ignatiadis**, Stanford University, United States
Co-authors: Stefan Wager

Methods are studied for simultaneous analysis of many noisy experiments in the presence of rich covariate information. The goal of the analyst is to optimally estimate the true effect underlying each experiment. Both the noisy experimental results and the auxiliary covariates are useful for this purpose, but neither data source on its own captures all the information available to the analyst. We propose a flexible plug-in empirical Bayes estimator that synthesizes both sources of information and may leverage any black-box predictive model. We show that our approach is within a constant factor of minimax for a simple data-generating model. Furthermore, we establish an extension to the classic result of James-Stein, whereby our proposed estimator dominates the sample mean of the experimental results under quadratic risk; even if the auxiliary covariates contain no information about the true effects. Finally, we exhibit promising empirical performance of the method on both real and simulated data.

E0818: **Asymptotically optimal simultaneous gaussian mean estimation with nonparametric regression**
Presenter: **Alton Barbehenn**, University of Illinois Urbana-Champaign, United States
Co-authors: Sihai Zhao

Simultaneous estimation of multiple parameters has received a great deal of recent interest, with applications in multiple testing, causal inference, and large-scale data analysis. Most approaches to simultaneous estimation use empirical Bayes methodology. We propose an alternative, completely frequentist approach based on nonparametric regression. We show that simultaneous estimation can be viewed as a constrained and penalized least-squares regression problem, so that empirical risk minimization can be used to estimate the optimal estimator within a certain class. We show that under mild conditions, our data-driven decision rules have an asymptotically optimal risk that can match the best-known convergence rates for this compound estimation problem. Our approach provides another perspective to understand sufficient conditions for asymptotic optimality of simultaneous estimation. Our proposed estimators demonstrate comparable performance to state-of-the-art empirical Bayes methods in a variety of simulation settings and our methodology can be extended to apply to many practically interesting settings.

E0978: **A nonparametric integrative Tweedie approach to empirical Bayes estimation with side information**
Presenter: **Wenguang Sun**, Zhejiang University, China

Compound estimation of normal means with auxiliary data collected from related source domains is considered. The empirical Bayes framework provides an elegant interface to pool information across different samples and construct efficient shrinkage estimators. We propose a nonparametric integrative Tweedie (NIT) approach to incorporating structural knowledge encoded in the auxiliary data to assist the simultaneous estimation of primary vector of parameters. NIT uses convex optimization tools to directly estimate the gradient of the log-density through an embedding in the reproducing kernel Hilbert space (RKHS), which is induced by the Steins discrepancy metric. Most popular structural constraints can be easily incorporated into our estimation framework. We characterize the asymptotic Lp risk of NIT by first rigorously analyzing its connections to the RKHS risk, and second establishing the rate at which NIT converges to the oracle estimator. The improvements in the estimation risk and the deteriorations in the learning rate are precisely tabulated as the dimension of side information grows. The numerical performance of NIT and its superiority over existing methods are illustrated through the analysis of both simulated and real data.

EO219 Room Virtual R5 HIGH-DIMENSIONAL DATA ANALYSIS IN ECONOMETRICS AND STATISTICS Chair: Sungkyu Jung

E0703: **Estimation of eigenvectors for linear combinations of high-dimensional covariance matrices and its application**
Presenter: **Kazuyoshi Yata**, University of Tsukuba, Japan
Co-authors: Aki Ishii, Makoto Aoshima

High-dimensional data often have a non-sparse and low-rank structure which contains strongly spiked eigenvalues. We call it the strongly spiked eigenvalue (SSE) model. We note that, under the SSE model, asymptotic normality is not valid because it is heavily influenced by strongly spiked eigenvalues. Recently, consistent estimators of eigenvectors for each high-dimensional covariance matrix have been given by developing a new PCA method called the noise-reduction (NR) methodology. A data transformation technique that transforms the SSE model into the non-SSE model has also been provided by using the NR method. Under the non-SSE model, we can ensure high accuracy for inferences by using the asymptotic normality. We consider the estimation of eigenvectors for linear combinations of the high-dimensional covariance matrices. We give a consistent estimator of the eigenvectors by developing the NR method. By using the estimator, we give a new data transformation technique that transforms the SSE model into the non-SSE model for the linear combinations. We propose a statistic for the linear combinations of mean vectors and prove that the statistic establishes the asymptotic normality. Finally, we investigate the performance of the statistic in actual data analyses.

E0713: **Principal weighted least square support vector machine: An online dimension-reduction tool for binary classification**
Presenter: **Seung Jun Shin**, Korea University, Korea, South

As relevant technologies advance, steamed data are frequently encountered in various applications, and the need for scalable algorithms becomes urgent. We propose the principal weighted least square support vector machine(PWLSSVM) as a novel tool for SDR in a binary classification where most SDR methods suffer since they assume continuous Y. We further show that the PWLSSVM can be employed for the online SDR for the streamed data. Namely, the PWLSSVM estimator can be directly updated from the new data without having old data. We explore the asymptotic properties of the PWLSSVM estimator and demonstrate its promising performance in terms of both estimation accuracy and computational efficiency for both simulated and real data.

E0946: **James-Stein for eigenvectors**
Presenter: **Lisa Goldberg**, University of California, Berkeley, United States
Co-authors: Alexander Shkolnik, Alec Kercheval

Estimated covariance matrices are widely used to construct portfolios with variance-minimizing optimization, yet the embedded sampling error produces portfolios with systematically underestimated variance. This effect is especially severe when the number of securities greatly exceeds the number of observations. In this high dimension low sample size (HL) regime, we show that a dispersion bias in the leading eigenvector of the estimated covariance matrix is a material source of distortion in the minimum variance portfolio. We correct the bias with the data-driven GPS (Global Positioning System) shrinkage estimator, which improves with the size of the market, and which is structurally identical to the James-Stein estimator for a collection of averages. We illustrate the power of the GPS estimator with a numerical example, and conclude with open problems that have emerged.

E0945: **Direction penalized principal component analysis**
Presenter: **Alexander Shkolnik**, University of California, Santa Barbara, United States
Co-authors: Youhong Lee

A regularization method is proposed called direction penalized principal component analysis (dPCA). This approach penalizes the first principal component, i.e., the direction of maximum variance of the data, for deviations away from some target direction. While the latter vector has an obvious interpretation in terms of a Bayesian prior, our main contributions lay elsewhere. In particular, we derive an optimal penalty parameter that, for any target, always reduces the asymptotic L2-loss function relative to that of the raw principal component. The optimal penalty parameter is determined solely from the data and an iterative algorithm efficiently computes the dPCA estimator. We prove our results by adopting a high-dimension, low sample size framework that is increasingly relevant for modern applications. To shed some insight into the dPCA estimator, we develop interesting connections to Ledoit-Wolf constant correlation shrinkage as well a recently proposed James-Stein estimator for the first principal component. We demonstrate the performance of dPCA by benchmarking against both of these estimators.

EO205 Room Virtual R6 FAIR, ROBUST, EFFICIENT AND EXPLAINABLE MACHINE LEARNING MODELS Chair: Yao Li

E0231: **Does enforcing fairness mitigate algorithmic biases due to distributional shift?**
Presenter: **Yuekai Sun**, University of Michigan, United States

Many instances of algorithmic bias are caused by distributional shifts. A particularly prominent class of examples is algorithmic biases caused by the under-representation of samples from minority groups in the training data. We study whether enforcing algorithmic fairness during training mitigates such biases in the target domain. On one hand, we show that there are scenarios in which enforcing fairness does not improve model performance (in the target domain). In fact, it may even harm performance. On the other hand, we derive sufficient conditions under which enforcing group and individual fairness successfully mitigate algorithmic biases due to distributional shifts.

E0325: **Detecting adversarial examples with Bayesian neural network**
Presenter: **Yao Li**, University of North Carolina at Chapel Hill, United States
Co-authors: Tongyi Tang, Thomas Lee, Cho-Jui Hsieh

A new framework is proposed to detect adversarial examples motivated by the observations that random components can improve the smoothness of predictors and make it easier to simulate the output distribution of a deep neural network. With these observations, we propose a novel Bayesian adversarial example detector, short for BATer, to improve the performance of adversarial example detection. In specific, we study the distributional difference of hidden layer output between natural and adversarial examples, and propose to use the randomness of Bayesian neural network (BNN) to simulate hidden layer output distribution and leverage the distribution dispersion to detect adversarial examples. The advantage of BNN is that the output is stochastic while neural networks without random components do not have such characteristics. Empirical results on several benchmark datasets against popular attacks show that the proposed BATer outperforms the state-of-the-art detectors in adversarial example detection.

E0453: **Time-frequency analysis of scalp EEG with Hilbert-Huang transform and deep learning**
Presenter: **Jingyi Zheng**, Auburn University, United States

Electroencephalography (EEG) is a brain imaging approach widely used in neuroscience and clinical settings. The conventional EEG analyses usually require pre-defined frequency bands when characterizing neural oscillations and extracting features for classifying EEG signals. However, neural responses are naturally heterogeneous. Failure to account for such variations might result in information loss and classifiers with a low accuracy but high variation across individuals. To address these issues, we present a systematic time-frequency analysis approach for analyzing scalp EEG signals. In particular, we propose a data-driven method to compute the subject-specific frequency bands for brain oscillations via the Hilbert-Huang Transform, lifting the restriction of using fixed frequency bands for all subjects. Then, we propose two novel metrics to quantify the power and frequency aspects of brainwaves represented by sub-signals decomposed from the EEG signals. The effectiveness of the proposed metrics is tested on two scalp EEG datasets and compared with four commonly used feature sets extracted from wavelet and Hilbert-Huang Transform. The validation results show that the proposed metrics are more discriminatory than other features leading to accuracies in the range of 94.93% to 99.84%. Besides classification, the proposed metrics show great potential in the quantification of neural oscillations and serve as biomarkers in neuroscience research.

E0492: **How to trust a black-box: Formal verification of deep neural networks**
Presenter: **Huan Zhang**, CMU, United States

Neural networks have become a crucial element in modern artificial intelligence. However, they are often black-boxes and can behave unexpectedly and produce surprisingly wrong results. When applying neural networks to mission-critical systems such as autonomous driving and aircraft control, it is often desirable to formally verify their trustworthiness such as safety and robustness. We will first introduce the problem of neural network verification and the challenges involved to guarantee neural network output given bounded input perturbations. Then, we will discuss the bound propagation-based neural network verification algorithms such as CROWN and beta-CROWN, which efficiently propagate linear inequalities through the network in a backward manner. State-of-the-art verification techniques will be highlighted which are used in our alpha-beta-CROWN verifier, a scalable, powerful and GPU-accelerated neural network verifier that won the 2nd International Verification of Neural Networks Competition (VNN-COMP21) with the highest total score.

EO253 Room Virtual R7 RECENT ADVANCES IN SURVIVAL ANALYSIS Chair: Sy Han Chiou

E0503: **Boosting method for length-biased and interval-censored survival data subject to high-dimensional error-prone covariates**
Presenter: **Li-Pang Chen**, National Chengchi University, Taiwan
Co-authors: Bangxu Qiu

Analysis of length-biased and interval-censored data is an important topic in survival analysis, and many methods have been developed to address this complex data structure. However, these methods focus on low-dimensional data and assume the covariates to be precisely measured, while high-dimensional data subject to measurement error are frequently collected in applications. We explore a valid inference method for handling high-dimensional length-biased and interval-censored survival data with measurement error in covariates under the accelerated failure time model. We primarily employ the SIMEX method to correct for measurement error effects and propose the boosting procedure to do variable selection and estimation. 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.

E0638: **Goodness-of-fit test for Cox model under isotonic constraint**
Presenter: **Huan Chen**, University of Texas at Dallas, United States
Co-authors: Chuan-Fa Tang

Cox proportional hazard model has been widely used as it shows the effect of covariates on the hazard. However, in many applications researchers are only willing to assume the hazard is isotonic to covariate such that the Cox proportional hazard model may introduce biases. In this case, a general isotonic proportional hazards model is more accurate. In order to determine which model is more appropriate based on the researchers' data, a likelihood ratio goodness-of-fit test is proposed via the bootstrap method, where the null hypothesis is Cox proportional hazard model while the alternative is the isotonic proportional hazards model. Starting from the time-independent univariate cases, the test can be generalized to time-dependent univariate and partial linear multivariate cases. The pseudo-iterative convex minorant algorithm is used for the estimation of the monotone hazard to guarantee efficiency.

E0761: **Weighted least squares estimation for semiparametric accelerated failure time model with regularization**
Presenter: **Ying Chen**, The university of Texas at Dallas, United States
Co-authors: Chuan-Fa Tang, Sy Han Chiou, Min Chen

Clustered failure time data arise when failure times are sampled in clusters. Depending on the sampling schemes, the selected sample of clusters might not be representative of the population. For example, case-cohort sampling is commonly adopted for studying large cohorts with rare events, and subjects who experienced events of interest are more likely to be sampled. The semiparametric accelerated failure time (AFT) model is an appealing method in survival analysis as it directly relates the failure times to a linear combination of covariates. However, most approaches in the AFT model framework are rank-based and rely on solving nonsmooth estimating equations. To facilitate inference procedures for AFT models, we extended a generalized estimating equation (GEE) embedded approach to account for the within-cluster correlations and the sampling bias. In a high-dimensional data setting, we propose to penalize the corresponding GEE for variable selection. We demonstrate the effectiveness of the proposed methods with high-dimensional data simulated under generalized stratified case-cohort designs. The large-scale simulation shows the proposed methods are more efficient than the estimator which ignores the sampling weights or within-cluster dependence.

E1043: **Transformation model based regression with dependently truncated and independently censored data**
Presenter: **Jing Qian**, University of Massachusetts, Amherst, United States
Co-authors: Sy Han Chiou, Rebecca Betensky

Truncated survival data arise when the event time is observed only if it falls within a subject-specific region. The conventional risk-set adjusted Kaplan-Meier estimator or Cox model can be used for estimation of the event time distribution or regression coefficient. However, the validity of these approaches relies on the assumption of quasi-independence between truncation and event times. One model that can be used for the estimation of the survival function under dependent truncation is a structural transformation model that relates a latent, quasi-independent truncation time to the observed dependent truncation time and the event time. The transformation model approach is appealing for its simple interpretation, computational simplicity and flexibility. We extend the transformation model approach to the regression setting. We propose three methods based on this model, in addition to a piecewise transformation model that adds greater flexibility. We investigate the performance of the proposed models through simulation studies and apply them to a study on cognitive decline in Alzheimer's disease.

EO337 Room Virtual R8 INNOVATIVE METHODS FOR GRAPHICAL MODELS AND NETWORKS Chair: Jeffrey Morris

E0885: **Connectivity regression**
Presenter: **Neel Desai**, University of Pennsylvania, United States

One key scientific problem in neuroscience involves assessing how functional connectivity networks in the brain vary across individuals and subject-specific covariates. We introduce a general framework for regressing subject-specific connectivity networks on covariates while accounting for inter-edge dependence within the network. The approach utilizes a matrix-logarithm function to transform the network object into an alternative space in which Gaussian assumptions are justified and positive semidefinite constraints are automatically satisfied. Multivariate regression models are fit in this space, with the covariance accounting for inter-edge network dependence, and multivariate penalization is used to induce sparsity in regression coefficients and covariance elements. We use permutation tests to perform multiplicity-adjusted inference to identify which covariates affect connectivity, and stability selection scores to indicate which network circuits vary by covariate. Simulation studies validate the inferential properties of the proposed method and demonstrate how estimating and accounting for inter-edge dependence when present leads to more efficient estimation, more powerful inference, and more accurate selection of which network circuits vary by covariates. We apply our method to data from the Human Connectome Project Young Adult study, revealing insights into how connectivity varies across language processing covariates and structural brain features.

E0985: **Functional Bayesian networks**
Presenter: **Fangting Zhou**, Texas AM University, United States

Multivariate functional data arise in a wide range of applications. One fundamental task is to understand the causal relationships among these functional objects of interest. We develop a novel Bayesian network model for multivariate functional data where the conditional independence and causal structure are represented by a directed acyclic graph. Specifically, we allow the functional objects to deviate from the common Gaussian process assumption, which is key for unique causal structure identification even when the functional data are purely observational and measured with noise. A fully Bayesian framework is designed to infer the functional Bayesian network model with natural uncertainty quantification through posterior summaries. Simulation studies and a real data application with brain electroencephalogram records demonstrate the utility of the proposed model.

E0989: **Probabilistic learning of treatment trees in cancer**
Presenter: **Tsung-Hung Yao**, University of Michigan at Ann Arbor, United States
Co-authors: Zhenke Wu, Karthik Bharath, Jinju Li, Veera Baladandayuthapani

Accurate identification of synergistic treatment combinations and their underlying biological mechanisms is critical across many domains. In oncology, preclinical systems such as patient-derived xenografts (PDX) have emerged as a unique study design evaluating multiple treatments administered to samples from the same human tumor implanted into mice. We propose a novel Bayesian probabilistic tree-based framework for PDX data to investigate the hierarchy between treatments by inferring treatment cluster trees, referred to as treatment trees (Rx-tree). The framework motivates a new metric of mechanistic similarity between two or more treatments accounting for the inherent uncertainty in tree estimation. Building upon Dirichlet Diffusion Trees, we derive a closed-form marginal likelihood encoding the tree structure, which facilitates computationally efficient posterior inference via a new two-stage algorithm. Simulation studies demonstrate the superior performance of the proposed method in recovering the tree structure and treatment similarities. Our analyses of a recently collated PDX dataset produce treatment similarity estimates that show a high degree of concordance with known biological mechanisms across treatments in five different cancers. More importantly, we uncover new and potentially effective combination therapies that confer synergistic regulation of specific downstream biological pathways for future clinical investigations.

E0986: **Bayesian functional graphical model for dynamic functional connectivity network inference**
Presenter: **Lin Zhang**, University of Minnesota, United States

A Bayesian functional graphical modeling framework is developed for correlated multivariate functional data, which allows the graphs to vary over the functional domain. The model involves estimation of graphical models that evolve functionally in a nonparametric fashion while accounting for within-functional correlations and borrowing strength across functional positions so contiguous locations are encouraged but not forced to have similar graph structure and edge strength. We utilize a strategy that combines nonparametric basis function modeling with modified Bayesian graphical regularization techniques, which induces a new class of hypoexponential normal scale mixture distributions that not only leads to adaptively shrunken estimators of the conditional cross-covariance but also facilitates a thorough theoretical investigation of the shrinkage properties. The approach scales up to large functional datasets collected on a fine grid. We show through simulations and real data analysis that the Bayesian functional graphical model can efficiently reconstruct the functionally–evolving graphical models by accounting for within-function correlations.

EO019 Room Virtual R9 RECENT DEVELOPMENTS IN COMPLEX IMAGING DATA ANALYSIS Chair: Dayu Sun

E0201: **Challenges and opportunities in statistical harmonization for multi-center neuroimaging studies**
Presenter: **Russell Shinohara**, University of Pennsylvania, United States

As multi-center studies in imaging science become increasingly commonplace, there is a need for understanding and mitigating biases associated with the acquisition of multiple scanners. We will review the state-of-the-art in image harmonization including harmonization in new settings such as longitudinal study designs, distributed analyses, and complex covariance structures in imaging features. We will also discuss the harmonization complex data objects such as those representing connectomic measurements. We will conclude with a discussion of future directions and areas for improvement in study design and analysis.

E0681: **A sparse blind source separation method for probing human whole-brain connectomes**
Presenter: **Ying Guo**, Emory University, United States
Co-authors: Yikai Wang

In neuroscience research, imaging-based network connectivity measures have become the key for understanding brain organizations, potentially serving as individual neural fingerprints. There are major challenges in analyzing connectivity matrices including the high dimensionality of brain networks, unknown latent sources underlying the observed connectivity, and the large number of brain connections leading to spurious findings. We propose a novel blind source separation method with low-rank structure and uniform sparsity (LOCUS) as a fully data-driven decomposition method for network measures. Compared with existing methods that vectorizes connectivity matrices ignoring brain network topology, LOCUS achieves more efficient and accurate source separation for connectivity matrices using the low-rank structure and a novel angle-based uniform sparsity regularization. We propose a highly efficient iterative Node-Rotation algorithm to solve the non-convex optimization problem for learning LOCUS. We illustrate LOCUS through extensive simulation studies and application to a resting state fMRI data.

E0826: **Bayesian spatially varying weight neural networks with the soft-thresholded Gaussian process prior**
Presenter: **Jian Kang**, University of Michigan, United States

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 a good prediction accuracy and the model fitting results can be difficult to interpret. We construct a novel single-layer Bayesian neural network (BNN) with spatially varying weights for the scalar-on-image regression. The goal is to select interpretable image regions and to achieve high prediction accuracy with limited training samples. We assign the soft-thresholded Gaussian process (STGP) prior to the spatially varying weights and develop an efficient posterior computation algorithm based on stochastic gradient Langevin dynamics (SGLD). The BNN-STGP provides large prior support for sparse, piecewise-smooth, and continuous spatially varying weight functions, enabling efficient posterior inference on image region selection and automatically determining the network structures. We establish the posterior consistency of model parameters and selection consistency of image regions when the number of voxels/pixels grows much faster than the sample size. We compared our methods with state-of-the-art deep learning methods via analyses of multiple real data sets including the task fMRI data in the Adolescent Brain Cognitive Development (ABCD) study.

Sunday 05.06.2022 13:15 - 14:55 Parallel Session H – EcoSta2022

EV463 Room Virtual R7 CONTRIBUTIONS IN CAUSAL INFERENCE Chair: Subir Ghosh

E0814: **Nonlinear and nonseparable structural functions in fuzzy regression discontinuity designs**
Presenter: **Haitian Xie**, UC San Diego, United States

The aim is to examine the identification and estimation of the structural function in fuzzy Regression Discontinuity (RD) designs with a continuous treatment variable. Under a dual monotonicity condition, we show that the nonlinear and nonseparable structural function can be nonparametrically identified at the RD cutoff. The dual monotonicity condition requires that the structural function and the treatment choice be strictly increasing in the unobserved causal factor. This condition is satisfied by standard parametric models used in practice. The identification result contrasts with the local average treatment effect literature, where only a certain weighted average of the structural function is identified. We propose a three-step semiparametric estimation procedure and derive the asymptotic distribution of the estimator. The semiparametric estimator achieves the same convergence rate as in the case of a binary treatment variable. As an application of the method, we estimate the causal effect of sleep time on health status by the discontinuity in natural light timing at time-zone boundaries.

E0825: **Joint diagnostic test of regression discontinuity designs: multiple testing problem**
Presenter: **Koki Fusejima**, The University of Tokyo, Japan
Co-authors: Takuya Ishihara, Masayuki Sawada

Diagnostic tests with a large number of covariates have been a norm to validate regression discontinuity (RD) designs. Such a procedure lacks its validity because of the multiple testing problem. Testable restrictions are to verify a single identification restriction, and therefore a single joint null hypothesis should be tested. In a meta-analysis of economics top five publications, the joint null was over-rejected and the null distribution of test statistics is distorted possibly by publication bias. We provide joint testing procedures based on the newly shown joint asymptotic normality of RD estimators. Simulation evidence demonstrates their favorable performances over Bonferroni correction for dimensions fewer than 10 covariates. However, neither Bonferroni correction nor our procedure guaranteed its size control with a larger number of covariates.

E0947: **Fuzzy Wald ratio difference-in-differences and changes-in-changes estimator for spatiotemporal and spatial data**
Presenter: **Andrej Srakar**, University of Ljubljana, Slovenia

Difference-in-differences (DiD) literature is a fast growing field in econometrics. Topics such as presence of serial correlation, aggregating data, clustered standard errors, arbitrary covariance structures, parallel growth assumption, synthetic control, multiple and continuous treatments and staggered treatment adoption have been subject to recent research. In previous contributions DiD has been extended to spatial data using Hadamard product conditioned calculus. We extend this in a treatment effect with network interference context by controlling for violated stable unit treatment values assumption (SUTVA), inherent for such analysis but seldom controlled so far and in a spatiotemporal autoregressive setting. We develop a time-corrected Wald ratio DiD estimator based on fuzzy DiD approach with extensions to changes-in-changes estimation. We provide asymptotic analysis using functional delta and Stein approaches and present results of Monte Carlo simulations. In an application, we study causal effects of the yearly Venice carnival, being able to isolate the effect respective to other competing large events in Venice in the studied period. The application uses three stage approach of using ARIMA models in the first stage, Frechet mean and median based derivation of spatial matrices in the second stage, and our new estimator in the final third stage. In conclusion, we consider extensions using spillover double robust DiD and Bayesian approaches.

E0743: **Causal impact of policy measures and behavior on the COVID pandemic in Germany**
Presenter: **Jenny Bethaeuser**, Justus Liebig University Giessen, Germany

Critics protest loudly against restrictions imposed by politicians during the COVID pandemic: Mandatory masks, lockdowns, school and business closures. The aim is to examine (1) the extent to which these policies have indirectly contributed to limiting the number of COVID cases and deaths by forcing people to practice social distancing, and (2) the extent to which people have adjusted their social distancing behavior on their own based on information about the national cases and fatality numbers and therefore directly limited the number of COVID cases and deaths. The panel analysis at the federal state level in Germany between 03/2020 and 12/2021 finds that substantial declines in COVID cases and death growth rates are attributable to private behavioral response, but policies played an important role as well. A change in policies explains a large fraction of changes in social distancing behavior, why both policies and national information are important determinants of federal COVID cases and deaths. Due to the lack of cross-sectional variation, there is uncertainty about the effect of the mask mandate.

EO055 Room 101 (Hybrid 1) RECENT ADVANCES IN HIGH-DIMENSIONAL COVARIANCE ESTIMATION Chair: Kei Hirose

E0380: **Asymptotic theory of sparse factor models in high-dimension**
Presenter: **Benjamin Poignard**, Osaka University, Japan
Co-authors: Yoshikazu Terada

The problem of estimating a factor model-based variance-covariance matrix is considered when the factor loading matrix is assumed sparse. We develop a penalized estimating function framework to handle the identifiability issue of the factor loading matrix while fostering sparsity in potentially all its entries. We prove the oracle property of the penalized estimator for the factor model, that is the penalization procedure can recover the true sparse support and the estimator is asymptotically normally distributed. Consistency and support recovery are established when the number of parameters is diverging. The non-penalized loss functions are deduced from the class of Bregman divergence losses, providing new estimators for factor modelling. These theoretical results are supported by empirical studies.

E0644: **Accelerating dependency modeling with graphics processing units**
Presenter: **Po-Hsien Huang**, National Chengchi University, Taiwan

Dependency modeling among response variables is a crucial task in multivariate analysis. A general strategy for this task is to introduce latent variables (or random effects) to capture the common part of the variables. By integrating out the latent variables, the so-called marginal maximum likelihood (MML) can be conducted for parameter estimation. However, when the number of latent factors is large, the MML generally becomes infeasible. We demonstrate how to use graphics processing units (GPU) computing and vectorization to greatly speed up the training process of MML. In particular, the MML for item factor analysis (IFA) is considered. IFA could be understood as a generalization of factor analysis for handling polytomous data. A python package called xifa was developed. Our numerical experiments show that xifa could be 33 times faster than its CPU counterpart. Furthermore, when the number of latent factors is equal to or larger than 5, xifa is much faster than the competing implementations, including the Bock-Aitkin expectation-maximization and the MHRM implemented by mirt (on CPU), and the importance-weighted autoencoder (on GPU). We believe GPU computing would play a central role in large-scale statistical modeling in the near future.

E0744: **Multilinear common component analysis for tensor data based on Kronecker product approach**
Presenter: **Shuichi Kawano**, The University of Electro-Communications, Japan
Co-authors: Kohei Yoshikawa

The common component analysis is a multivariate method that extracts a common structure from several datasets. We present multilinear common component analysis (MCCA) based on Kronecker products of mode-wise covariance matrices in order to extract a common structure from multiple tensor datasets. We develop an estimation algorithm and establish the convergence properties of the algorithm. Numerical studies are given to show the effectiveness of MCCA.

E0781: **Computationally efficient forecasting algorithm in the SUTSE model and its properties**
Presenter: **Wataru Yoshida**, Kyushu University, Japan
Co-authors: Kei Hirose

The problem of forecasting multivariate time series is considered by using a Seemingly Unrelated Time Series Equations (SUTSE) model. In the SUTSE model, multiple univariate time series equations are combined to express a single state-space model, resulting in the coefficient matrices for system and observation models to be diagonal. The SUTSE model usually assumes the correlations among error variables. In this case, however, the model estimation requires heavy computational loads due to a large matrix computation, especially for high-dimensional data. To alleviate the computational issue, we propose a two-stage procedure for forecasting. First, we conduct the Kalman filter as if the correlations among error variables do not exist; that is, univariate time series analyses are conducted separately. Next, the forecast value is computed by estimating a covariance matrix of the forecast error. The proposed algorithm is much faster than the ordinary SUTSE model because we do not require a large matrix computation. Some theoretical properties of our proposed estimator are presented. Monte Carlo simulation is conducted to investigate the effectiveness of our proposed method. The results show that our proposed method is comparable with the ordinary SUTSE in prediction accuracy.

EO411 **Room 102 (Hybrid 2)** **STATISTICAL LEARNING FOR FUNCTIONAL DATA** **Chair: Yousri Slaoui**

E0760: **Unsupervised classification method based on nonparametric functional mode estimation**
Presenter: **Yousri Slaoui**, University of Poitiers, France

The focus is on an unsupervised classification problem in the framework of nonparametric functional data. We first, proposed a classification method based on a recursive estimation of the mode of the distribution of a functional random variable, this estimator is based on a pseudo-density estimator. Moreover, we study the asymptotic properties of these two estimators. We then showed the performance of the proposed unsupervised classification estimator by considering a real electroencephalography dataset. Finally, we compare our estimator to a parametric approach based on a Stochastic block Model for node-weighted networks based on two emission laws.

E0796: **Using Bayesian neural network as an actor in actor-critic methods**
Presenter: **Leo Grill**, Universite de Poitiers, France
Co-authors: Yousri Slaoui, David Nortershauser, Stephane Le Masson

Reinforcement learning and deep learning lack statistical theory background. Bayesian approaches lead to more robust learning. The Bayesian theory has already started enlightening Deep learning. It can be used to deal with issues such as overfitting or black-box modeling. Bayesian methods are used to train the neural network encoding the actor and find a policy in deep actor-critic algorithms. The results of simulations are presented to show the interest in using these methods for deep reinforcement learning. It is particularly important for the exploration phase and for the stability of learning. The Bayesian neural network also benefits from the advantages of ensemble methods, it can be used to realize coherent exploration. The prior and posterior approaches can be either used to regularize the model or two bring some knowledge during the training.

E0802: **Statistical learning in nonparametric q-kernel q-density estimation**
Presenter: **Oumaima Ben Mrad**, CNRS and University of Poitiers, France
Co-authors: Yousri Slaoui, Afif Masmoudi

In the context of quantum calculus, the interest is in statistical learning of nonparametric kernel density estimation. Firstly, we propose two q-density estimations. The first one is based on a q-Uniform kernel and the second is based on a q-Gaussian kernel. Secondly, we focus on characteristics related to Jackson's q-integral and q-derivative and we investigate the asymptotic properties of the two proposed q-kernel q-density estimators. Moreover, we conduct a numerical study to show the efficiency and the feasibility of the two proposed estimators by considering various values of the q parameters and various sample sizes.

E0895: **Manifold MCMC algorithm for Gamma-GPD mixture model**
Presenter: **Salah El Adlouni**, Universite de Moncton, Canada

Modelling extreme events is important in many domains, including environmental variables, civil engineering, reliability, financial risk, and computer security. The peaks over threshold approach (POT) describes the main characteristics of the observed extreme series, yet the threshold selection is challenging and might affect the results. Mixture models offer more flexibility to represent samples with heterogenous data. The Gamma-Generalized Pareto mixture model (GAM-GP) is presented for extreme risk estimation. The model is developed in its general form, where the observed events depend on multi-dimensional covariates and non-linear link functions. A new Monte Carlo Markov Chain algorithm, based on Riemannian Manifold, is developed to estimate the parameters in a Bayesian framework. Results show the capacity of the proposed algorithm to converge to the posterior distribution of the parameters even for the high dimension of the covariates space. The approach is illustrated on simulated data and a daily streamflow dataset for the Saint John River at Fort-Kent (upstream) New-Brunswick (Canada).

EO041 **Room 103 (Hybrid 3)** **THEORIES AND METHODOLOGIES FOR HIGH-DIMENSIONAL DATA** **Chair: Kazuyoshi Yata**

E0643: **Test for outlier detection by high-dimensional PCA**
Presenter: **Yugo Nakayama**, Kyoto University, Japan
Co-authors: Kazuyoshi Yata, Makoto Aoshima

Outlier detection for high-dimensional, low sample size (HDLSS) data is studied. Theories and methodologies for high-dimensional data have become increasingly important in many fields. In particular, there has been a strong demand for HDLSS analysis. Principal component analysis (PCA) is investigated under the HDLSS settings. However, there are still some areas where analysis methods have not been fully established, one of which is outlier detection. For high-dimensional data, classical methods based on the Mahalanobis distance are usually not applicable, so an alternative is needed. One of the methods for the univariate data is the Smirnov-Grubbs test. We propose a new outlier detection by applying high-dimensional PC scores to the Smirnov-Grubbs test. By using the asymptotic properties of the PC scores, we evaluate its size and power. Our results show that the proposed method gives preferable performances in the HDLSS setting. Finally, we check the performance of the outlier detection method in both numerical and real data analysis.

E0773: **Condition of GIC to select the model minimizing KL-loss function in high-dimensional multivariate linear regression**
Presenter: **Ryoya Oda**, Hiroshima University, Japan
Co-authors: Hirokazu Yanagihara

The focus is on the variable selection method based on minimizing the generalized information criterion (GIC) for selecting explanatory variables in a normality-assumed multivariate linear regression. The GIC is defined as the sum of -2 times maximum log-likelihood and a penalty term included in a positive parameter. From the viewpoint of the prediction ability, it is often hoped that the model which minimizes a loss function among all candidate models is chosen. Hence, it is important to examine whether or not the GIC has the asymptotic property that the probability of selecting the model minimizing the KL-loss function converges to 1. We call this property consistency. Recently, there has been significant attention in the literature to statistical methods for high-dimensional data. We obtain conditions for consistency of the GIC based on minimizing the KL-loss function under the following high-dimensional asymptotic framework: the sample size tends to infinity and the dimension of response variables divided by the sample size converges to a positive constant within [0,1). Then, using the obtained conditions, we propose a consistent variable selection criterion under the high-dimensional asymptotic framework. Through simulation experiments, it is shown that the probability of selecting the model minimizing the KL-loss function by our proposed criterion is high even when the dimension is large.

E0915: **Non-linear variable selection via kernel regression with high-dimensionality**
Presenter: **Yuta Umezu**, Nagasaki University, Japan

In a high-dimensional scenario, where the number of covariates is much larger than the sample size, marginal screening is a fundamental technique for model selection. We focus on extracting non-linear relationships between response and covariates based on kernelized marginal regression models with sparsity inducing penalty. From the representer theorem and KKT conditions, our screening method can simply be implemented and enjoys the sure screening property under some mild conditions, that is, all active covariates will be retained with probability converging to one even in a high-dimensional setting. In addition, we consider choosing a thresholding value for screening. Since our screening score can be considered as generalized V-statistics, its asymptotic distribution can be derived, so we can asymptotically control the expected false positive rate at the same time as assuring sure screening property. We will also present several simulation studies and a real data example for checking the performance of our method.

E0414: **Testing allometric extension in high-dimensional and spiked eigenvalue situations**
Presenter: **Koji Tsukuda**, Kyushu University, Japan
Co-authors: Shun Matsuura

In multivariate allometry, the first principal component vector of a covariance matrix is of interest sometimes. In particular, when there are two groups, if the first principal component vectors of the two groups have the same direction and the direction is identical to the difference of the mean vectors, then one group is called an allometric extension of the other group. Several studies have considered statistical hypothesis testing of the allometric extension. However, previous studies dealt with the cases where the dimension of observed variables is smaller than sample sizes, and have not supposed high-dimensional situations. Statistical procedures that are not justified under high-dimensional asymptotic regimes do not work in high-dimensional situations often. Therefore, we propose a high-dimensional test procedure for allometric extension; in particular, we show the asymptotic normality of the test statistic and the consistency of the test under a high-dimensional asymptotic regime with an assumption that eigenvalues of covariance matrices are spiked.

EO191 **Room 104 (Hybrid 4)** STATISTICAL INFERENCE ON VARIOUS MANIFOLD **Chair: Toshihiro Abe**

E0517: **An extended sine-skewed circular distribution and its extension to a model on cylinder**
Presenter: **Yoichi Miyata**, Takasaki City University of Economics, Japan

The sine skewed circular distribution is a tractable circular probability model that can be asymmetric in shape and that has the advantage that the sine and cosine moments can be written in explicit forms. We use the framework of Ley and Verdebout to propose a new family of probability distributions, including the sine skewed circular distribution. This family includes distributions that can give stronger asymmetry around the mode than the sine skewed circular distribution. Furthermore, we show that a subfamily of the extended sine-skewed wrapped Cauchy distributions is identifiable with respect to parameters, and all distributions in the subfamily have explicit sine and cosine moments. We will also discuss an extension of the proposed circular distribution to probability models on cylinder.

E0521: **Complex valued time series modeling in relation to directional statistics**
Presenter: **Takayuki Shiohama**, Nanzan University, Japan
Co-authors: Takayuki Shiohama

Stationary time series fluctuation often shows periodic behavior and these patterns are usually summarized via a spectral density. Since the spectral density is a periodic function, it can be modeled by using a circular distribution function. Several time series models are studied in relation to a circular or a cylindrical distribution. First, as an introduction, we illustrate how to model bivariate time series data using complex-valued time series in the context of circular distribution functions. Next, some other time series modeling by incorporating cylindrical distributions is illustrated. The maximum likelihood estimation procedures are introduced to estimate unknown model parameters. Some real data analyses are also performed to illustrate the proposed models' applicability.

E0540: **Construction of a circular distribution from a discrete distribution and its extension**
Presenter: **Tomoaki Imoto**, University of Shizuoka, Japan

In diverse scientific fields, a data sample is often represented as a point in the circumference of a unit circle. Typical examples are wind direction at some point and event time measured on a 24-h clock. Such data are called circular data and should be modeled by a distribution defined on the circle, called circular distribution. The trigonometric moments play an important role in characterizing the circular distribution and statistical analysis through the circular distribution. A method is provided for constructing a circular distribution by assuming that the trigonometric moments follow a discrete distribution on the line. The probability density function of the resulting distribution is expressed by using the characteristic function of the discrete distribution. A multivariate extension is also obtained by assuming that the joint trigonometric moments follow a multivariate discrete distribution. Some examples constructed from the proposed method are shown, and fitting examples to real data are given,

E0601: **Smoothing parameter selection of circular kernel density estimation**
Presenter: **Yasuhito Tsuruta**, The University of Nagano, Japan

Nonparametric density estimations, such as kernel density estimation, enable flexible estimations. Therefore, nonparametric density estimations of circular data have been studied extensively for decades. Circular kernel density estimators are affected by the choice of the smoothing parameter. Unfortunately, the optimal parameter, which minimizes the mean integrated square error, depends on a derivative of an unknown density. Therefore, many studies have proposed smoothing parameter selectors. A few studies discuss the asymptotic properties of these selectors. The aim is to investigate some properties of the selectors: least-squares cross-validation and direct plug-in rule. The result shows that the convergence rate of the direct plug-in rule is faster than that of least squares cross-validation. The numerical experiment shows the performance of least squares cross-validation and direct plug-in rule under small samples.

EO331 **Room 105 (Hybrid 5)** **RECENT DEVELOPMENT ON FUNCTIONAL ANALYSIS OF COMPLEX DATA** **Chair: Weichi Wu**

E0185: **A unified approach to hypothesis testing for functional linear models**
Presenter: **Zhenhua Lin**, National University of Singapore, Singapore
Co-authors: Yinan Lin

A unified approach is developed to hypothesis testing for various types of widely used functional linear models, such as scalar-on-function, function-on-function and function-on-scalar models. In addition, the proposed test applies to models of mixed types, such as models with both functional and scalar predictors. In contrast with most existing methods that rest on the large-sample distributions of test statistics, the proposed method leverages the technique of bootstrapping max statistics and exploits the variance decay property that is an inherent feature of functional data, to improve the empirical power of tests, especially when the sample size is limited and the signal is relatively weak. Theoretical guarantees on the validity and consistency of the proposed test are provided uniformly for a class of test statistics.

E0666: **Semiparametric function-on-function quantile regression model with dynamic single-index interactions**
Presenter: **Yuanyuan Zhang**, Soochow University, China
Co-authors: Hanbing Zhu, Yehua Li

A new semiparametric function-on-function quantile regression model with time-dynamic single-index interactions is proposed. Our model is very flexible in taking into account the nonlinear time-dynamic interaction effects of the multivariate longitudinal/functional covariates on the longitudinal response, and most existing quantile regression models for longitudinal data are special cases of our proposed model. We propose to approximate the bivariate nonparametric coefficient functions by tensor product B-splines and employ a check loss minimization approach to estimate the bivariate coefficient functions and the index parameter vector. Under some mild conditions, we establish the asymptotic normality of the estimated single-index coefficients using the projection orthogonalization technique and obtain the convergence rates of the estimated bivariate coefficient functions. Furthermore, we propose a score test to examine whether there exist interaction effects between the covariates. The finite sample performance of the proposed method is illustrated by Monte Carlo simulations and an empirical data analysis.

E0816: **Simultaneous inference for the scalar response functional linear regression models**
Presenter: **Yan Cui**, Harbin Institute of Technology, China

The focus is on the problem of joint simultaneous confidence band (JSCB) construction for regression coefficient functions of time series scalar-on function linear regression when the regression model is estimated by either truncated basis expansions or roughness penalizations with flexible choices of orthonormal basis functions. A simple and unified multiplier bootstrap methodology is proposed for the JSCB construction which is shown to achieve the correct coverage probability asymptotically. Furthermore, the JSCB is robust to inconsistently estimated standard deviations of the model. The proposed methodology is applied to an environmental and a financial time series data set to visually investigate and formally test the overall regression relationship as well as perform model validation. A uniform Gaussian approximation and comparison result over all Euclidean convex sets for normalized sums of a class of moderately high-dimensional stationary time series is established, which may be of separate interest. Finally, the proposed methodology can be readily applied to simultaneous inference for scalar-on-function linear regression of independent cross-sectional data.

E1000: **Estimation of functional treatment effect using generalized empirical likelihood stabilized weights**
Presenter: **Ruoxu Tan**, The University of Hong Kong, Hong Kong
Co-authors: Wei Huang, Guosheng Yin, Zheng Zhang

Most studies concerning the effect of a functional variable on an outcome are restricted to exploring the association rather than the casual relationship. In the areas where causal treatment effect analysis has many applications, functional data has become popular in the recent decade. Due to the lack of definition of probability density function for functional data, estimating the propensity score for functional treatment is challenging. The limited literature on the functional treatment effect tackled this problem by replacing the functional treatment with its functional principal component scores in the definition of the propensity score. However, such an approximation does not guarantee asymptotic consistency. We propose not using the principal component scores but identifying the average treatment effect by a weighted conditional expectation of the observed outcome given the functional treatment. The weights, called the stabilized weights, can be well defined in terms of a functional treatment. We then estimate the stabilized weights using a generalized empirical likelihood method and show the consistency of our estimator. After that, a functional linear estimator of the average treatment effect is proposed. We study the theoretical and numerical properties of the estimator. A real data application demonstrates the practical value of our method.

EO339 **Room Virtual R1** **NATURE-INSPIRED METAHEURISTIC METHODS AND APPLICATIONS** **Chair: Frederick Kin Hing Phoa**

E0771: **Metaheuristic optimization on tensor-type solution via swarm intelligence**
Presenter: **Hsinping Liu**, National Taiwan University, Taiwan
Co-authors: Frederick Kin Hing Phoa, Jessica Yun Heh Chen-Burger, Shau Ping Lin

Nature-inspired metaheuristic optimization has been widely used in many problems in industry and scientific investigations, but their applications in designing selling schemes are rare because the solution space in this kind of problem is usually high-dimensional, and their constraints are sometimes cross-dimensional. Recently, the Swarm Intelligence Based (SIB) method is proposed for problems in discrete domains, and it is widely applied in many mathematical and statistical problems that common metaheuristic methods seldom approach. We introduce an extension of the SIB method that handles solutions with many dimensions, or tensor solutions in mathematics. We further speed up our method by implementing our algorithm with the use of CPU parallelization. We then apply this extended framework to real applications in designing selling schemes, showing that our proposed method helps to increase the profit of a selling scheme compared to those suggested by traditional methods.

E0828: **Finding time series motif by using swarm intelligence method**
Presenter: **Hendri Sutrisno**, Academia Sinica, Taiwan
Co-authors: Frederick Kin Hing Phoa

Time series motif discovery has been one of the most discussed problem domains in data mining. Most of the methods proposed for discovering time series motifs are computationally exhaustive, mainly on more extensive time-series data. We propose a swarm intelligence method to approximate the motif location. In the methodology, the solutions in the search space were clustered into several sub-optimum groups based on an automatic clustering mechanism to enable the local-global search strategies. The local search mechanism bounds the search space based on the clustering result into regions to improve exploitation ability, and the global search mechanism promotes information changing between the local best solutions to improve the exploration ability. The experiment results on both synthetic and real datasets reconfirm that our method can speed up dramatically compared to the current techniques for discovering the time series motifs.

E0931: **An efficient method to scatter network nodes on a spherical surface via swarm intelligence**
Presenter: **Chao-Hui Huang**, National Tsing Hua University, Taiwan
Co-authors: Frederick Kin Hing Phoa

The space-filling problem has been an important topic in many scientific and practical aspects. There have been many theoretical and methodological results when the space is flat and regular, but few discussions are found for evenly distributing points on a spherical surface. Fibonacci lattice is an elegant solution to this problem when the points are all independent, but the condition is hardly fulfilled especially when we consider the nodes in a network. Although this problem easily becomes very complex and time-consuming with the existence of clusters, it is highly useful and practical in the visualization of network data. We provide an efficient two-steps method to arrange the network nodes uniformly on a spherical surface. We partition the spherical surface associated with a criterion about the edge/point ratio, then we scatter the nodes on the respective subspace according to the relationship between nodes and modularity. In order to reduce the computational efforts, we first uniformly distribute points on a two-dimensional plane uneven with a functional gradient, then we stereographically project all points from a gradient plane back to a sphere. Some networks are used for demonstration.

E0992: **Traveling salesman problem via swarm intelligence**
Presenter: **Pei-Chen Yen**, The University of Melbourne, Australia
Co-authors: Frederick Kin Hing Phoa

An efficient method via swarm intelligence is introduced to handle the traveling salesman problem, which is widely applied in many real-world applications. This method can be seen as a discrete version of the PSO with some variants. Compared to the classic Ant Colony Optimization method, the proposed SIB method performs well in terms of efficiency and accuracy in the TSP problem. For TSP with cities size between 15 to 25, SIB has a significantly lower average executing time to obtain an adequate solution with a close distance.

EO177 Room Virtual R10 RECENT DEVELOPMENTS ON DATA INTEGRATION FOR STATISTICS Chair: Anne Ruiz-Gazen

E0792: **Guidelines on areal interpolation methods**
Presenter: **Thibault Laurent**, Universite Toulouse 1 Capitole, France
Co-authors: Anne Vanhems, Van Huyen Do

The objective is to delve deeper into the understanding and practical implementation of classical areal interpolation methods using R software. Based on a survey paper, we focus on four classical methods used in the area-to-area interpolation problem: point-in-polygon, areal weighting interpolation, dasymetric method with auxiliary variable and dasymetric method with control zones. Using the departmental election database for Toulouse in 2015, we find that the point-in-polygon method can be applied if the sources are much smaller than the targets; the areal interpolation method provides good results if the variable of interest is related to the area, but otherwise, a good alternative is to use the dasymetric method with another auxiliary variable; and finally, the dasymetric method with control zones allows us to benefit from both areal interpolation and dasymetric method and, from that perspective, seems to be the best method.

E0934: **Spatial multivariate trees for integrating geospatial data from multiple sources**
Presenter: **Michele Peruzzi**, Duke University, United States
Co-authors: David Dunson

High-resolution geospatial data are challenging because standard geostatistical models based on Gaussian processes are known to not scale to large data sizes. While progress has been made towards methods that can be computed more efficiently, considerably less attention has been devoted to methods for large scale data that characterize complex relationships between several outcomes recorded at high resolutions by different sensors or data sources. In these settings, popular coregionalization models along with assumptions of conditional independence across spatial neighbors may be inappropriate when the spatial resolution from one data source is much lower than others. Our spatial multivariate trees (SpamTrees) are based on conditional independence assumptions on latent random effects based on a treed directed acyclic graph. SpamTrees can be interpreted as a multiscale method for multivariate data in which outcomes that are more sparsely observed are placed at tree heights corresponding to coarser scales. Information-theoretic arguments and considerations on computational efficiency guide the construction of the tree and the related efficient sampling algorithms in these imbalanced settings. We illustrate SpamTrees using a large climate data set which combines high-resolution satellite data with sparsely observed land-based station data.

E0939: **Improving finite population inference by data integration**
Presenter: **Anne Ruiz-Gazen**, Toulouse School of Economics, France
Co-authors: Estelle Medous, Camelia Goga, Jean-Francois Beaumont, Alain Dessertaine, Pauline Puech

Combining survey sample data and big databases is an important current challenge in finite population inference. While survey sample data are obtained through a probability sampling design, big data consist usually of non-probability samples. Many well-known unbiased or approximately unbiased methods exist for estimating finite population parameters from a probability sample. Inference from a non-probability sample is, however, often subject to selection bias. Recently, a data integration approach has been proposed that allows handling the selection bias of non-probability samples by incorporating a probability sample. We propose to revisit their approach and study in detail the gain in terms of efficiency of the estimators based on a probability sample when taking into account non-probability samples.

E0937: **Statistical data integration using a prediction approach**
Presenter: **Estelle Medous**, University of Toulouse 1, France
Co-authors: Anne Ruiz-Gazen, Camelia Goga, Jean-Francois Beaumont, Alain Dessertaine, Pauline Puech

In a finite population setting, it is possible to improve the efficiency of estimators based on a probability sample by using non-probability big data sources. However, the target variable may not be observable in the big data sources, while the auxiliary information present in these sources may not be measured in the probability sample. In such a situation, new estimators can be proposed with a prediction approach. These estimators are either design-based, model-based, or cosmetic. Their properties in terms of bias and efficiency are studied using some theoretical and simulation results. The interest of the new proposals is illustrated in the context of the French postal service, where the objective is to estimate the monthly postal traffic through a survey of the mailmen rounds while taking advantage of the database containing information on the automatically processed postal mail.

EO317 Room Virtual R2 ADVANCES IN STATISTICAL METHODS FOR RELIABILITY ANALYSIS Chair: Man Ho Ling

E0189: **Robust statistical inference for one-shot devices**
Presenter: **Elena Castilla**, Universidad Rey Juan Carlos, Spain

One-shot device testing is an increasingly important problem in the area of reliability. This is an extreme case of interval censoring, where one only knows if the device works when it is tested. The existing literature on one-shot devices is extensive and focuses particularly on the development of techniques for estimating the maximum likelihood estimator (MLE). The MLE however, presents an important lack of robustness in the presence of outliers. We present alternative divergence-based estimators, which are seen to be more robust with an unavoidable loss of efficiency.

E0290: **Lamination scheme of curing degree at multiple levels of temperature with location-scale regression**
Presenter: **Chien-Tai Lin**, Tamkang University, Taiwan
Co-authors: Chih-Chun Tsai, Narayanaswamy Balakrishnan

Solar power has become a key green source of energy. An important factor that affects the reliability and lifetime of solar modules is the quality of encapsulation through the lamination process, which melts the ethylene-vinyl acetate (EVA) to make the solar cells combine with the front glass side and the rear side units. The degree of crosslinking or curing degree for EVA sheets, when the EVA sheet gets heated, can affect the efficiency of the performance and power conversion of solar modules. For this reason, motivated by lamination data, we construct here a statistical model for describing the relationship between the curing degree and the lamination factors (temperature and time). Then, based on some specification limits on the curing degree, the optimal lamination time for solar modules can be determined at different temperatures. Moreover, the optimal sample size allocation in a test for measuring EVA sheets can also be determined. A simulation study is finally carried out to show the closeness of simulation results to the asymptotic results.

E0809: **Order restricted inference for adaptive progressively censored competing risks data**
Presenter: **Debanjan Mitra**, Indian Institute of Management Udaipur, India

Under adaptive progressive Type-II censoring schemes, order restricted inference based on competing risks data is discussed. The latent failure lifetimes for the competing causes are assumed to follow Weibull distributions, with an order restriction on the scale parameters of the distributions. The practical implication of this order restriction is that one of the risk factors is dominant, as often observed in competing risks scenarios. In this setting, likelihood estimation for the model parameters, along with bootstrap-based techniques for constructing asymptotic confidence intervals are presented. Bayesian inferential methods for obtaining point estimates and credible intervals for the model parameters are also discussed. Through a detailed Monte Carlo simulation study, the performance of order restricted inferential methods are assessed. In addition, the results are also compared with the case when no order restriction is imposed on the estimation approach. The simulation study shows that order restricted inference is more efficient between the two when this additional information is taken into consideration. A numerical example is provided for illustrative purposes.

E0921: **Interval estimation and hypothesis testing for the generalized Pareto distribution under non-regularity conditions**
Presenter: **Hideki Nagatsuka**, Chuo University, Japan
Co-authors: Narayanaswamy Balakrishnan

The generalized Pareto distribution (GPD), introduced by Pickands, is widely used to model exceedances over thresholds. It is well known that inference for the GPD is a difficult problem since the moments exist only for a limited range of parameters and the GPD violates the classical regularity conditions in the maximum likelihood method. For parameter estimation, most existing methods do not perform satisfactorily for all ranges of parameters. Furthermore, the interval estimation and hypothesis tests have not been studied well in the literature. We introduce a novel framework for inference for the GPD, which works successfully for all values of the shape parameter. Specifically, a new method of parameter estimation for the GPD is constructed, and some asymptotic properties of the proposed estimators and related statistics are derived. The existence and uniqueness of the proposed estimates are also established. Based on the asymptotic properties of the proposed estimators and related statistics, new confidence intervals and hypothesis tests are developed. The performances of the proposed estimators of parameters, confidence intervals and hypothesis tests are then shown by Monte Carlo simulation and real data examples.

EO021 Room Virtual R3 RECENT ADVANCES IN SPATIAL SCAN STATISTICS Chair: Matthieu Marbac

E0195: **Spatial scan statistics in statistical models**
Presenter: **Tonglin Zhang**, Purdue University, United States

Spatial scan statistics are powerful in detecting spatial clusters for disease data, but the method is rarely combined with statistical models. To incorporate the method into a statistical model, a straightforward idea is to specify a group of artificial explanatory variables for spatial clusters with each cluster candidate explained by an explanatory variable. Based on this formulation, the spatial scan test can be carried out by a variable selection procedure with the estimates of coefficients for the strength for spatial clusters and the size of explanatory variables for the number of clusters. To implement this idea, the research develops a method to connect variable selection with cluster detection under the framework of generalized linear models. It then proposes a generalized information criterion approach to estimate both the number of clusters and their shapes.

E0276: **Multivariate scan statistics for spatial data**
Presenter: **Lionel Cucala**, Universite de Montpellier, France

A parametric and a nonparametric scan method for multivariate data indexed in space are introduced. The parametric one relies on a generalized likelihood ratio associated with the multivariate Gaussian distribution whereas the nonparametric one is completely distribution-free as it is based on so-called multivariate ranks. In contrast to existing scan methods, they both take into account the covariance structure between all observed variables. These methods are compared through a simulation study and then applied to a dataset recording the levels of metallic pollutants for two areas in the North of France.

E0294: **Identification of geographic clusters for temporal heterogeneity with application to dengue surveillance**
Presenter: **Pei-Sheng Lin**, National Health Research Institutes, Taiwan

Identifying the transmission of hot spots with temporal trends is important for reducing infectious disease propagation. Cluster analysis is a particularly useful tool to explore underlying stochastic processes between observations by grouping items into categories by their similarity. In a study of epidemic propagation, clustering geographic regions that have similar time series could help researchers track diffusion routes from a common source of an infectious disease. We propose a two-stage scan statistic to classify regions into various geographic clusters by their temporal heterogeneity. The proposed scan statistic is more flexible than traditional methods in that contiguous and non-proximate regions with similar temporal patterns can be identified simultaneously. A simulation study and data analysis for a dengue fever infection are also presented for illustration.

E0541: **A new spatial scan statistic for multiple spatial clusters**
Presenter: **Mohamed Salem Ahmed**, University of Lille, France
Co-authors: Michael Genin, Matthieu Marbac

The focus is on the development of a spatial scan statistic able to detect multiple spatial clusters and test their significance, as well as ensure a reasonable computation time for large spatial data. The proposed method is based on generalized linear models in which the spatial clusters are integrated assuming that they have arbitrary parametric shapes (allowing elliptical and rectangular cluster shapes). This allows detecting spatial clusters by estimating their parametric shapes rather than proceeding with an exhaustive search over a set of candidate clusters. We propose a new Monte-Carlo procedure to evaluate the statistical significance of the detected spatial clusters and to estimate the actual number of spatial clusters. Simulations and a case study show that the proposed method is able to consistently and efficiently detect multiple spatial clusters.

EO059 Room Virtual R4 BLOCKCHAIN, DIGITAL CURRENCIES, AND DECENTRALISED FINANCE Chair: Stephen Chan

E0497: **eXplainable AI for credit risk management**
Presenter: **Branka Hadji Misheva**, ZHAW Zurich University of Applied Sciences, Switzerland
Co-authors: Joerg Osterrieder, Ali Hirsa

Artificial Intelligence (AI) has created the single biggest technology revolution the world has ever seen. For the finance sector, it provides great opportunities to enhance customer experience, democratize financial services, ensure consumer protection and significantly improve risk management. While it is easier than ever to run state-of-the-art machine learning models, designing and implementing systems that support real-world finance applications have been challenging. In large part, this is due to the lack of transparency and explainability which in turn represent important factors in establishing reliable technology. The research on this topic with a specific focus on applications in credit risk management has been limited. We implement different advanced post-hoc model agnostic explainability techniques to machine learning (ML)-based credit scoring models applied to loan performance data. We present multiple comparison scenarios and we discuss in detail the practical challenges associated with the implementation of these state-of-art eXplainable AI (XAI) methods.

E0553: **Topological data analysis of dynamic Ethereum token networks**
Presenter: **Yuzhou Chen**, Princeton University, United States

Forecasting price in the dynamic Ethereum token networks data is indispensable for understanding the blockchain dynamics and measuring the risk connectedness among the cross-cryptocurrency trades. In the last few years, Geometric Deep Learning (GDL), e.g., Graph Convolutional Networks (GCNs), have emerged as a powerful alternative to more conventional time-series predictive models. Despite their proven success, GCNs tend to be limited in their ability to simultaneously infer latent temporal relations among entities. We make the first step on a path of bridging the two emerging directions, namely, time-aware GDL with time-conditioned topological representations of complex dynamic Ethereum token networks. To summarize such time-conditioned topological properties, we develop novel topological representations. We then propose topology-based GDL models which allow us to simultaneously learn co-evolving intra- and inter-dependencies in the dynamic Ethereum token networks data.

E0574: **Asymmetric tail dependence modeling, with application to cryptocurrency market data**
Presenter: **Yan Gong**, KAUST, Saudi Arabia
Co-authors: Raphael Huser

Since the inception of Bitcoin in 2008, cryptocurrencies have played an increasing role in the world of e-commerce, but the recent turbulence in the cryptocurrency market in 2018 has raised some concerns about their stability and associated risks. For investors, it is crucial to uncover the dependence relationships between cryptocurrencies for more resilient portfolio diversification. Moreover, the stochastic behavior in both tails is important, as long positions are sensitive to a decrease in prices (lower tail), while short positions are sensitive to an increase in prices (upper tail). In order to assess both risk types, we develop a flexible copula model which is able to distinctively capture asymptotic dependence or independence in its lower and upper tails simultaneously. We apply our model to the historical closing prices of five leading cryptocurrencies, which share large cryptocurrency market capitalizations. The results show that our proposed copula model outperforms alternative copula models and that the lower tail dependence level between most pairs of leading cryptocurrencies-and in particular Bitcoin and Ethereum-has become stronger over time, smoothly transitioning from an asymptotic independence regime to an asymptotic dependence regime in recent years, whilst the upper tail has been relatively more stable overall at a weaker dependence level.

E1010: **An analysis of the return-volume relationship in decentralized finance (DeFi)**
Presenter: **Stephen Chan**, American University of Sharjah, United Arab Emirates
Co-authors: Jeffrey Chu, Yuanyuan Zhang

The decentralized finance sector has recently experienced a surge in popularity and has emerged from the shadows of the cryptocurrency space. Although the purposes of the currencies used in this new sector differ from traditional cryptocurrencies, they still possess monetary value and can be traded using fiat currencies on specialized decentralized exchanges. This paper investigates the dynamic volume-return relationship of the five largest decentralized finance tokens, to better understand this relationship given the similarities with cryptocurrencies and the possible benefits for traders and practitioners. We implement the quantile-on-quantile regression and an extreme value theory approach to examine the relationship between the daily returns of the prices and trading volumes of decentralized finance tokens at varying quantiles and at the extreme tails. Our results suggest that when trading volume is experiencing large increases, the returns of the prices of tokens appear to be significantly positive for some cases but negative for others. The extreme volume-return dependence is found to be asymmetric in the extreme negative and positive tails of the distributions, where the dependence below extreme negative thresholds is essentially non-existent but above extreme positive thresholds, it is significant. This extreme dependence between return and volume may be beneficial for developing trading strategies that incorporate trading volume data.

EO149 Room Virtual R5 SPATIO-TEMPORAL MODELS FOR ENVIRONMENTAL AND HEALTH APPLICATIONS Chair: Chae Young Lim

E0268: **An interaction Neyman-Scott point process model for COVID-19**
Presenter: **Jaewoo Park**, Yonsei University, Korea, South
Co-authors: Won Chang, Boseung Choi

With rapid transmission, the COVID-19 has led to over three million deaths worldwide, posing significant societal challenges. Understanding the spatial patterns of patient visits and detecting local cluster centers are crucial to controlling disease outbreaks. We analyze COVID-19 contact tracing data collected from Seoul, which provide a unique opportunity to understand the mechanism of patient visit occurrence. Analyzing contact tracing data is challenging because patient visits show strong clustering patterns, while cluster centers may have complex interaction behavior. To account for such behaviors, we develop a novel interaction Neyman-Scott process that regards the observed patient visit events as offsprings generated from a parent cluster center. Inference for such models is challenging since the likelihood involves intractable normalizing functions. To address this issue, we embed an auxiliary variable algorithm into our Markov chain Monte Carlo. We fit our model to several simulated and real data examples under different outbreak scenarios and show that our method can describe the spatial patterns of patient visits well. We also provide useful visualizations that can inform public health interventions for infectious diseases, such as social distancing.

E0275: **Robust distance-based clustering for sparse multivariate functional data**
Presenter: **Zhuo Qu**, KAUST, Saudi Arabia

A novel elastic time distance for sparse multivariate functional data is introduced and a robust distance-based clustering algorithm is proposed. The elastic time distance serves as a foundation for clustering functional data with various time measurements per subject. With the elastic time distance, classical distance-based clustering methods such as K-medoids and agglomerative hierarchical clustering are extended to the sparse multivariate functional case. Unlike most model-based, and the aforementioned classical, clustering methods our approach, besides being applicable to unbalanced multivariate functional data directly, is outlier-resistant and can detect outliers that do not belong to any clusters. Numerical experiments on simulated data highlight the excellent performance of the proposed algorithm compared to existing methods. Using Pacific Northwest cyclone track data as a motivating example, we demonstrate the effectiveness of the proposed approach.

E0515: **Estimating concurrent climate extremes: A conditional approach**
Presenter: **Whitney Huang**, Clemson University, United States
Co-authors: Adam Monahan, Francis Zwiers

Simultaneous concurrence of extreme values across multiple climate variables can result in large societal and environmental impacts. Therefore, there is growing interest in understanding these concurrent extremes. In many applications, not only the frequency but also the magnitude of concurrent extremes are of interest. One way to approach this problem is to study the distribution of one climate variable given that another is extreme. We develop a statistical framework for estimating bivariate concurrent extremes via a conditional approach, where univariate extreme value modeling is combined with dependence modeling of the conditional tail distribution using techniques from quantile regression and extreme value analysis to quantify concurrent extremes. We focus on the distribution of daily wind speed conditioned on daily precipitation taking its seasonal maximum. The Canadian Regional Climate Model large ensemble is used to assess the performance of the proposed framework both via a simulation study with a specified dependence structure and via an analysis of the climate model-simulated dependence structure.

E0576: **A spectral adjustment for spatial confounding**
Presenter: **Yawen Guan**, University of Nebraska - Lincoln, United States
Co-authors: Garritt Page, Brian Reich, Massimo Ventrucci, Shu Yang

Adjusting for an unmeasured confounder is generally an intractable problem, but in the spatial setting, it may be possible under certain conditions. We derive necessary conditions on the coherence between the treatment variable and the unmeasured confounder that ensure the causal effect of the treatment is estimable. We specify our model and assumptions in the spectral domain to allow for different degrees of confounding at different spatial resolutions. The key assumption that ensures identifiability is that confounding present at global scales dissipates at local scales. We show that this assumption in the spectral domain is equivalent to adjusting for global-scale confounding in the spatial domain by adding a spatially smoothed version of the treatment variable to the mean of the response variable. Within this general framework, we propose a sequence of confounder adjustment methods that range from parametric adjustments based on the Matern coherence function to more robust semi-parametric methods that use smoothing splines. These ideas are applied to areal and geostatistical data for both simulated and real datasets.

EO440 **Room Virtual R6** SOME RECENT DEVELOPMENTS IN HIGH DIMENSIONAL STATISTICS **Chair: Lei Huang**

E0271: **High-dimensional central limit theorems by Stein's method**
Presenter: **Xiao Fang**, The Chinese University of Hong Kong, Hong Kong

Explicit error bounds are obtained for the d-dimensional normal approximation on hyperrectangles for a random vector that has a Stein kernel, or admits an exchangeable pair coupling, or is a non-linear statistic of independent random variables or a sum of n locally dependent random vectors. We assume the approximating normal distribution has a non-singular covariance matrix. The error bounds vanish even when the dimension d is much larger than the sample size n. We prove our main results using a previous approach in Stein's method, together with modifications of a given estimate and a smoothing inequality. For sums of n independent and identically distributed isotropic random vectors having a log-concave density, we obtain an error bound that is optimal up to a $\log n$ factor. We also discuss an application to multiple Wiener-Itôintegrals.

E0286: **Relative error-based model averaging**
Presenter: **Xiaochao Xia**, Chongqing University, China

A relative error-based model averaging (REMA) approach is proposed to predict the positive response data under a set of multiplicative error models. To estimate the parameters in each candidate multiplicative model, we utilize a relative error-type loss as empirical objective function. Specifically, two commonly used losses: the least product relative error (LPRE) and the least absolute relative error (LARE) are considered, under which two model averaging estimators, REMA-LPRE and REMA-LARE, are proposed accordingly. The involved optimal weight vector, $\mathbf{w}$, is chosen by minimizing a jackknife version of the relative error loss over $\mathcal{H}_n = \{\mathbf{w} \in [0,1]^M, \sum_{m=1}^M w_m = 1\}$, where M denotes the number of candidate models. Theoretically, it is shown that under some technical conditions, our proposed model averaging estimators enjoy asymptotic optimality under the two losses, respectively, in the sense that its loss defined by a final prediction error (FPE) is asymptotically identical to that of the infeasible but best model averaging estimator. Furthermore, we present an extension to relaxing the summation constraint in $\mathcal{H}_n$, in which the asymptotic optimality for the LPRE-based model averaging estimator is still established. Extensive simulations and empirical applications are conducted to demonstrate the usefulness of our approach.

E0430: **Limiting spectral distribution of large dimensional Spearman's rank correlation matrices**
Presenter: **Cheng Wang**, Shanghai Jiao Tong University, China

The empirical spectral distribution of Spearman's rank correlation matrices is studied under the assumption that the observations are independent and identically distributed random vectors and the features are correlated. We show that the limiting spectral distribution is the generalized Marcenko-Pastur law with the covariance matrix of the observation after standardized transformation. With these results, we compare several classical covariance/correlation matrices including the sample covariance matrix, Pearson's correlation matrix, Kendall's correlation matrix and Spearman's correlation matrix.

E0490: **On singular values of data matrices with general independent columns**
Presenter: **Chen Wang**, University of Hong Kong, Hong Kong
Co-authors: Tianxing Mei, Jeff Yao

The focus is on singular values of a large $p \times n$ data matrix $X_n = (x_{n1},...,x_{nn})$ where the columns $\{x_{nj}\}$ are independent p-dimensional vectors, possibly with different distributions. Assuming that the covariance matrices $n_j = Cov(x_{nj})$ of the column vectors can be asymptotically simultaneously diagonalizable, with appropriately converging spectra, we establish a limiting distribution for the singular values of X_n when both dimension p and n grow to infinity in comparable magnitude. The matrix model goes beyond and includes many existing works on different types of sample covariance matrices, such as the weighted sample covariance matrix, the Gram matrix model and the sample covariance matrix of linear times series models. Furthermore, we develop three applications of our general approach. First, we obtain the existence and uniqueness of a new limiting spectral distribution of realized covariance matrices for a multi-dimensional diffusion process with anisotropic time-varying co-volatility processes. Secondly, we derive the limiting spectral distribution for singular values of the data matrix for a recent matrix-valued auto-regressive model. Finally, for a generalized finite mixture model, the limiting spectral distribution for singular values of the data matrix is obtained.

EO255 Room Virtual R8 DESIGN AND ANALYSIS OF SCREENING EXPERIMENTS Chair: Rakhi Singh

E0523: Screening using locating arrays

Presenter: **Violet Syrotiuk**, Arizona State University, United States

Screening experiments are often the first step in identifying potentially important factors that significantly impact the response variables of a system. We propose an experimental design and analysis method for screening experiments based on a *locating array*, a new combinatorial design. Locating arrays have an expected number of runs that grows logarithmically in the number of factors. Thus, our methods can support a very large factor-space. They also support categorical factors due to the level-wise nature of both the design and the analysis. We provide a precise definition of locating arrays and describe our algorithm used for analysis, based on orthogonal matching pursuit. Results on real data sets provide a check of the validity of our methods.

E0467: Effective algorithms for constructing two-level QB-optimal designs for screening experiments

Presenter: **Alan Vazquez**, University of California, Los Angeles, United States

Co-authors: Weng Kee Wong, Peter Goos

Optimal two-level screening designs are widely applied in the manufacturing industry to identify factors that explain most of the product variability. These designs feature each factor at two settings and are traditionally constructed using standard algorithms, which rely on a pre-specified linear model. Since the assumed model may depart from the truth, two-level QB-optimal designs have been developed to provide efficient estimates for parameters in a large set of potential models as well. The optimal designs also have an overarching goal that models that are more likely to be the best for explaining the data are estimated more efficiently than the rest. Despite these attractive features, there are no good algorithms to construct these designs. Therefore, we propose two algorithms. The first algorithm, which is rooted in mixed-integer programming, guarantees convergence to the two-level QB-optimal designs. The second algorithm, which is based on metaheuristics, employs a novel formula to assess these designs and it is computationally efficient. Using numerical experiments, we demonstrate that our mixed integer programming algorithm is attractive to find small optimal designs, and our heuristic algorithm is an effective approach to constructing both small and large designs.

E0372: A factor screening approach for supersaturated experiments with an exponential family response via Dantzig selector 2.0

Presenter: **Jing-Wen Huang**, National Tsing-Hua University, Taiwan

Co-authors: Frederick Kin Hing Phoa, Yu-Wei Chen

Dantzig selector is a powerful method for the analysis of experiments conducted in a supersaturated design. It strikes a balance between variable selection and orthogonal projection estimation. However, its underlying assumption on the normally distributed response is not always valid in real applications. The aim is to generalize the formulation of the Dantzig selector to analyze experiments with a response that follows exponential family distribution by a maximum likelihood estimation approach. It results in an approximate linear program for any fixed tuning parameters that features low computational complexity and short computational time. Moreover, we propose a binary search algorithm for tuning parameter selection. We demonstrate the performance of our proposed method via simulation studies of a supersaturated design with a logistic binary response. Our method shows good performance by comparing it with several conventional methods.

E0656: Optimal two-level designs under model uncertainty

Presenter: **Steven Gilmour**, KCL, United Kingdom

Co-authors: Pi-Wen Tsai

Two-level designs are widely used for screening experiments where the goal is to identify a few active factors which have major effects. Most work on two-level designs is based on the effect hierarchy assumption that lower-order effects are of more importance than higher-order effects, so the focus is on two-level designs with level balance and pairwise orthogonality. We apply the model-robust Q_B criterion for the selection of optimal two-level designs by incorporating experimenters' prior knowledge on the importance of each effect into the optimality criterion. We find a smooth relationship between the choice of designs and the experimenters' prior beliefs. Additionally, we provide a coordinate exchange algorithm for the construction of Q_B-optimal designs without the restrictions of level-balance and pairwise orthogonality.

EO071 Room Virtual R9 RECENT DEVELOPMENTS IN STATISTICAL DEEP LEARNING Chair: Il Do Ha

E0570: Deep learning-based residual control chart for count data

Presenter: **Jong-Min Kim**, University of Minnesota at Morris, United States

Co-authors: Il Do Ha

Statistical process control for count data has difficulty overcoming multicollinearity. We propose a new deep learning residual control chart based on the asymmetrical count response variable when there are highly correlated explanatory variables. We implement and compare different methods such as neural network, deep learning, principal component analysis based Poisson regression, principal component analysis based negative binomial regression, nonlinear principal component analysis based Poisson regression, and nonlinear principal component analysis based negative binomial regression in terms of the root mean squared error. Using two asymmetrical simulated datasets generated by the combined multivariate normal, binary and copula functions, the neural network and deep learning have a smaller mean, median, and interquartile range when compared to the principal component analysis based Poisson regression, principal component analysis based negative binomial regression, nonlinear principal component analysis based Poisson regression, and nonlinear principal component analysis based negative binomial regression. We also compare the deep learning and neural network based residual control charts in terms of the average run length with the copula based asymmetrical simulated and real bids number of takeover bids data.

E0589: An ensemble model of CNN-BiLSTMs for forecasting NASDAQ volatility index

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

Co-authors: DongWan Shin

A new forecast method is proposed based on artificial neural networks (ANNs), ensemble CNN-BiLSTM, which is an ensemble of three CNN-BiLSTMs constructed with the combination of Convolution Neural Network (CNN) and Bidirectional Long Short-Term Memory (BiLSTM). The new forecast method effectively handles the strong long memory serial dependence feature of the daily VXN by the ensemble CNN-BiLSTM together with proper normalization and batch size. The long memory features arising from time-dependent mean and variance are largely reduced by normalizing the data with local mean and local standard deviation (SD). The batch size is determined by the optimal block length of the moving block bootstrap which reflects the long memory. The ensemble CNN-BiLSTM concentrates on 1-day, 1-week, and 2-week features of the normalized VXN data. An out-of-sample forecast comparison reveals that (i) the proposed ensemble CNN-BiLSTM has better forecast performance than the autoregressive model, DNN, LSTM, BiLSTM, and individual CNN-BiLSTMs; (ii) the local mean-SD normalization has superior forecast performance to the standard global mean-SD normalization; (iii) and the optimal block length improves the forecast performance over a batch size considered in the literature.

E0721: **Variable and architecture selection in neural networks**
Presenter: **Andrew McInerney**, University of Limerick, Ireland
Co-authors: Kevin Burke

Feedforward neural networks can be viewed as non-linear regression models, where covariates enter the model through a combination of weighted summations and non-linear functions. Although these models have similarities to the models typically used in statistical modelling, the majority of neural network research has been conducted outside of the field of statistics. This has resulted in a lack of statistically-based methodology, and, in particular, there has been little emphasis on model parsimony. Determining the input layer structure is analogous to variable selection while determining the structure for the hidden layer(s) relates to model complexity. However, the calculation of an associated likelihood function opens the door to information-criteria-based variable and architecture selection. A novel top-down model selection method is proposed using the Bayesian information criterion for feedforward neural networks, wherein the optimal weights for one model are carried over to the next. Simulation studies are used to evaluate the performance of the proposed method, and an application on real data is investigated.

E0821: **Understanding deep learning via statistical modelling approaches**
Presenter: **Il Do Ha**, Pukyong National University, Korea, South
Co-authors: Jihun Kim

Recently, deep learning (DL) have provided breakthrough results for prediction problems including classification for a wide variety of applications. In particular, the core architectures that currently dominate the DL are deep feed-forward neural networks (DFNN), CNN, RNN, LSTM, AE and GAN, etc. The DL models are represented as structured neural networks consisting of three layers (input, hidden and output layers) for constructing (or modelling) the functional relationship between input and output variables, and the main goal is to find a nonlinear predictor of the output Y given the input X. The output models of DL can be expressed as structured mean models, leading that the estimation of such mean provides the prediction of Y. It is thus interested to study the DL in terms of statistical perspective. The DL models can be viewed as a highly nonlinear and semi-parametric generalization of statistical models such as the generalized linear model (GLM). The fitting (i.e. learning) of DL models based on train data is usually implemented using likelihood-based methods including the construction of loss function or regularization. We present how to understand the DL models via the GLM framework, and then extend it to survival models allowing for censoring and to random-effect models, with practical examples.

Sunday 05.06.2022 15:25 - 16:40 Parallel Session I – EcoSta2022

EV460 Room Virtual R2 CONTRIBUTIONS IN APPLIED ECONOMETRICS Chair: James Flegal

E0357: **Measuring the impact of cybersecurity breaches on bitcoin returns**
Presenter: **George Milunovich**, Macquarie University, Australia

Digital exchanges, which convert funds between national currencies and cryptocurrencies, are often the victims of cybersecurity attacks. We investigate the impact of such cybersecurity breaches on bitcoin returns. Using several alternative specifications we test the hypothesis that bitcoin returns experience a decrease on the dates associated with cybersecurity breaches of cryptocurrency exchanges. We find a negative impact where bitcoin declines between 1.288% and 1.470% on cyberattack days, depending on which model is applied and what control variables are included. The finding sheds light on this important but not widely recognized source of cryptocurrency market risk.

E0463: **Can unlisted firms benefit from market information? A data-driven approach**
Presenter: **Michele Modina**, University of Molise, Italy
Co-authors: Alessandro Bitetto, Stefano Filomeni

A sample of 10,136 Italian micro-, small-, and mid-sized enterprises (MSMEs) that borrow from 113 cooperative banks is employed to examine whether market pricing of public firms adds additional information to accounting measures in predicting default of private firms. Specifically, we first match the asset prices of listed firms following a data-driven clustering by means of Neural Networks Autoencoder so to evaluate the firm-wise probability of default (PD) of MSMEs. Then, we adopt three statistical techniques, namely linear models, multivariate adaptive regression spline, and random forest to assess the performance of the models and to explain the relevance of each predictor. We find a significant improvement in model performance when including the estimated PD in the predictive specifications.

E0679: **Optimal fiscal policies in booms and in recessions: An econometric case study for Slovenia**
Presenter: **Reinhard Neck**, Alpen-Adria Universitaet Klagenfurt, Austria
Co-authors: Dmitri Blueschke, Klaus Weyerstrass

Optimal fiscal policies for the next few years are determined for Slovenia under alternative assumptions about global development. In particular, we distinguish between a scenario of a recession (assuming the recent Covid-19 crisis continues over a considerable period) and a scenario of the boom (assuming that the crisis is over by summer 2022). We use the macroeconometric model SLOPOL11 and assume an intertemporal objective function for Slovenian policymakers containing output, unemployment, inflation, the budget deficit, public debt, and the current account as the main arguments. Using the OPTCON algorithm, approximately optimal policies are calculated under both scenarios. It turns out that the design of fiscal policies is rather similar in both cases, showing the relatively low effectiveness of the fiscal instruments with respect to their influence on the small open economy within the Euro Area.

EV451 Room Virtual R6 CONTRIBUTIONS IN COMPUTATIONAL STATISTICS AND APPLICATIONS Chair: Abdelaati Daouia

E0922: **A Gaussian mixture model with a modified Hard EM algorithm in clustering problems**
Presenter: **Samyajoy Pal**, LMU Munich, Germany
Co-authors: Christian Heumann

Hard EM or Viterbi Training is often used for complex unsupervised learning models as it is less computationally intensive and easy to implement. However, it is considered to be inferior to standard EM as it is known to have some theoretical disadvantages, like biased estimates and lack of consistency. Also, in what circumstances it is to be preferred over the other is not well understood. We have revisited the issue of Hard EM for cluster analysis. We have proposed some modifications to the Hard EM algorithm to build Gaussian Mixture Models. The performance of the model has been assessed over different situations (increasing number of clusters, increasing dimension, increasing overlap, imbalance, etc.) on five benchmark data sets. Then the results are compared with standard EM to investigate if it really works as badly as assumed. The study also includes an analysis of two real data sets from biological science to explore the convenience of the proposed models.

E0173: **Expectation-maximization algorithm with combinatorial assumption**
Presenter: **Loc Nguyen**, Loc Nguyen's Academic Network, Vietnam

The expectation-maximization (EM) algorithm is a popular and powerful mathematical method for parameter estimation in case there exist both observed data and hidden data. The EM process depends on an implicit relationship between observed data and hidden data which is specified by a mapping function in traditional EM and a joint probability density function (PDF) in practical EM. However, the mapping function is vague and impractical whereas the joint PDF is not easy to be defined because of heterogeneity between observed data and hidden data. The aim is to improve the competency of EM by making it more feasible and easier to be specified, which removes the vagueness. Therefore, it is assumed that observed data is the combination of hidden data which is realized as an analytic function where data points are numerical. In other words, observed points are supposedly calculated from hidden points via a regression model. Mathematical computations and proofs indicate the feasibility and clearness of the proposed method which can be considered an extension of EM.

E0929: **The importance of morphology data in predicting the risks of aneurysm rupture**
Presenter: **Ehsan Kharatikoopaei**, Durham University, United Kingdom
Co-authors: Nasima Akhter, Amanda Ellison, Nicki Richards, Boguslaw Obara, Adetayo Kasim, Steve Steve Bonner, Edel McCauley, Nitin Mukerji

Spontaneous Subarachnoid Haemmorrhage (SAH), mainly caused by rupture of Intracranial aneurysms (IA), is a common cerebrovascular disorder affecting worldwide mortality and morbidity. Approximately, 3.5 million people in the UK are likely to have unruptured IA. While the management of unruptured IA is controversial and not all aneurysms rupture, there is a lot more to understand about the possibility of SAH and risk factors associated with it. Aneurysm morphology may influence aneurysm rupture, and advanced methods investigating the relative importance of morphological data in predicting aneurism rupture have been underutilized. Random forest (RF) classification is applied to a database of over 400 patients attending James Cook Hospital in North East England, UK, to predict rupture of aneurysms. RF showed that when demographic data were used in addition to the morphological data, the prediction accuracy goes up to 71% with age at haemorrhage, aneurysm dome diameter (d5), log ratio of d5 and aneurysm dome height, years of smoking, log ratio of aneurysm neck diameter and d5 being the top five most important variables. This shows the non-ignorable importance of the factors which are not commonly noted and can be useful to support clinical decisions in this context.

EV454 Room Virtual R7 CONTRIBUTIONS IN FINANCIAL ECONOMETRICS Chair: Pavel Krupskiy

E0748: **Spurious tail risk factors and asset prices**
Presenter: **Maxime Nicolas**, Universite Paris 1 Pantheon-Sorbonne, France

It is argued that recent findings on the predictive ability of tail risk exposure are likely spurious. We argue that these results are related to biases in the estimation procedure of the tail dependence coefficient (TDC) computed based on the joint behavior of equity returns, market returns, or other factors. Backed by a simulation framework, we show how this coefficient may capture a high level of correlation rather than asymptotic dependence. Then, we replicate recent studies finding a relationship between crash risk exposure and future excess returns. We proceed to show that these results do not hold when we control for the correlation coefficient and other past return behavior.

E0881: **Cryptocurrencies' quantile and tail expectation forecasting**
Presenter: **Kokulo Kpai Lawuobahsumo**, University of Calabria, Italy
Co-authors: Bernardina Algieri, Arturo Leccadito

The aim is to jointly predict conditional quantiles and tail expectations for the returns of the most popular cryptocurrencies (Bitcoin, Ethereum, Ripple, and Litecoin) using financial and macroeconomic indicators as explanatory variables. The financial variables we use are Nasdaq Composite, WTI Futures, Gold Fixing Price 3:00 P.M. (London time), and CBOE Volatility Index (VIX). The economic variables we considered are U.S. Dollar Index, 5-Year Forward Inflation Expectation Rate, 10-Year Breakeven Inflation Rate, 10-Year Treasury Constant Maturity, and 10-Year Market Yield on U.S. Treasury Securities. We use daily data and the Monotone Composite quantile regression neural network model (MCQRNN) to make one-step-ahead and five-step-ahead predictions for Value-at-Risk (VaR) and Expected Shortfall (ES) on a rolling basis and compare the performance of our model against the standard GARCH (1,1) model. The superior set of models is then chosen by backtesting α-VaR and α-ES using a Model Confidence Set (MCS) procedure with a loss function. Our results show that the MCQRNN performed better than the benchmark model for jointly predicting VaR and ES. The result is consistent for 1% and 5% levels of α both in the right and left tails for all cryptocurrencies.

E0474: **A novel approach to bank marketing campaign**
Presenter: **Yuzhi Cai**, Swansea University, United Kingdom

Banks usually need to identify a group of customers in order to target them with a specific financial product that will allow them to retain existing and attract new customers. A binary response model could be used to predict the response probability for each customer, which could then be used for classification. However, as the predicted response probability depends on the mean of an unobserved variable, it is not clear whether these response probability forecasts are able to provide optimal classification results for, e.g. bank marketing campaigns. We show that these response probability forecasts usually do not lead to optimal classification results because they use only one specific piece of information about the response variable and this piece of information may not be representative of the response variable. We also propose a novel binary quantile function model and a new classification method that allow us to use a set of information about the response variable from which optimal classification results can be obtained. We illustrate this by analysing some real bank marketing data.

EV459 Room Virtual R8 CONTRIBUTIONS IN FORECASTING Chair: Matthieu Marbac

E0323: **On the uncertainty of a combined forecast: The critical role of correlation**
Presenter: **Andrey Vasnev**, University of Sydney, Australia
Co-authors: Jan Magnus

The purpose is to show that the effect of the zero-correlation assumption in combining forecasts can be huge, and that ignoring (positive) correlation can lead to confidence bands around the forecast combination that are much too narrow. In the typical case where three or more forecasts are combined, the estimated variance increases without a bound when correlation increases. Intuitively, this is because similar forecasts provide little information if we know that they are highly correlated. Although we concentrate on forecast combinations and confidence bands, our theory applies to any statistic where the observations are linearly combined. We apply our theoretical results to explain why forecasts by Central Banks (in our case, the Bank of Japan) are so frequently misleadingly precise. In most cases, a correlation above 0.7 is required to produce reasonable confidence bands.

E0843: **Forecasting half-hourly electricity prices using a mixed-frequency VAR framework: The case of New Zealand market**
Presenter: **Nuttanan Wichitaksorn**, Auckland University of Technology, New Zealand
Co-authors: Gaurav Kapoor, Wenjun Zhang

A mixed-frequency vector autoregressive (VAR) framework is employed to forecast half-hourly electricity prices while having several weather variables and electricity demand that come with another frequency. In addition to a standard VAR model used in the analysis, we propose a VAR model extending from a single-equation RU-MIDAS model. LASSO is also incorporated to help with the variable selection. These models are estimated using a range of techniques including least squares, Gibbs sampling, and Variational Bayes. We compare our forecasting results with those from random subspace regressions, e.g., subset and projection regressions. We found our results are favorable, especially those with LASSO.

E0864: **Using the yield curve to forecast recessions: The role of fragmentation in the Euro area**
Presenter: **Jean-Baptiste Hasse**, Universite Catholique de Louvain, Belgium
Co-authors: Quentin Lajaunie

A new early warning system (EWS) model is developed which is based on the yield curve to forecast recessions. Using a panel logit model and a unique dataset covering 11 Euro area member countries over the period 2001-2021, we empirically show that EWS based on both the level of the short-term government bond rate and the yield spread give better predictive performance than models with the yield spread alone. This result is robust to different econometric specifications, controlling for monetary policy stance and recession risk factors. Furthermore, via an innovative cluster analysis, we give the empirical evidence of the role of the sovereign bond market fragmentation between Core and Periphery European countries. Our results provide a useful toolbox for monitoring economic cycles.

EO081 Room 102 (Hybrid 2) GAUSSIAN APPROXIMATION FOR HIGH-DIMENSIONAL DATA Chair: Yuta Koike

E0344: **Gaussian approximation and spatially dependent wild bootstrap for high-dimensional spatial data**
Presenter: **Daisuke Kurisu**, Yokohama National University, Japan
Co-authors: Kengo Kato, Xiaofeng Shao

A high-dimensional CLT is established for the sample mean of p-dimensional spatial data observed over irregularly spaced sampling sites in $\mathbb{R}^d$, allowing the dimension p to be much larger than the sample size n. We adopt a stochastic sampling scheme that can generate irregularly spaced sampling sites in a flexible manner and include both pure increasing domain and mixed increasing domain frameworks. To facilitate statistical inference, we develop the spatially dependent wild bootstrap (SDWB) and justify its asymptotic validity in high dimensions by deriving error bounds that hold almost surely conditionally on the stochastic sampling sites. The dependence conditions on the underlying random field cover a wide class of random fields such as Gaussian random fields and continuous autoregressive moving average random fields. Through numerical simulations and a real data analysis, we demonstrate the usefulness of our bootstrap-based inference in several applications, including joint confidence interval construction for high-dimensional spatial data and change-point detection for spatio-temporal data.

E0910: **Gaussian approximations for high-dimensional non-degenerate U-statistics via exchangeable pairs**
Presenter: **Zhi Liu**, University of Macau, China

Some recent results are reported on the non-asymptotic bound for Gaussian approximations for centered high-dimensional non-degenerate U-statistics over the class of hyper-rectangles via. We improved the upper bound of the convergence rate from the order one-sixth of the sample size to a quarter of the sample size, up to a polynomial factor of the logarithm of the dimension.

E0944: **On Gaussian approximations for M-estimators**
Presenter: **Masaaki Imaizumi**, The University of Tokyo, Japan

The purpose is to develop a non-asymptotic Gaussian approximation theory for distributions of M-estimators, which are defined as maximizers of empirical criterion functions. In existing mathematical statistics literature, numerous studies have focused on approximating the distributions of the M-estimators for statistical inference. In contrast to the existing approaches, which mainly focus on limiting behaviors, we employ a non-asymptotic approach, establish abstract Gaussian approximation results for maximizers of empirical criteria, and propose a Gaussian multiplier bootstrap approximation method. Our developments can be considered as extensions of previous work on the approximation theory for distributions of suprema of empirical processes toward their maximizers. We shed new light on the statistical theory of M-estimators. Our theory covers not only regular estimators, such as the least absolute deviations, but also some non-regular cases where it is difficult to derive or approximate numerically the limiting distributions such as non-Donsker classes and cube root estimators.

EO127 Room 103 (Hybrid 3) STOCHASTIC CONTROL AND DATA SCIENCE IN ECONOMICS AND FINANCE Chair: Seyoung Park

E0442: **Optimal retirement with disability risk**
Presenter: **Jiwon Chae**, Pohang University of Science and Technology, Korea, South
Co-authors: Bong-Gyu Jang, Seyoung Park

Long-term disability significantly changes an individual's financial planning and quality of life. We illustrate the effects by modeling lifecycle consumption and investment while considering disability risk and retirement. The high intensity of disability shock lowers the optimal retirement wealth level; therefore, it leads to early retirement. Our model shows that the income risk due to such a high disability shock can lower the consumption-to-wealth ratio. Moreover, we find that the optimal risky portfolio-to-wealth ratio is reduced when the possibility of disability events rises, and its drop gets even larger for the poor. Our analysis of the certainty equivalent wealth (CEW) implies that people with low initial wealth are willing to pay for a large portion of their wealth to eliminate the disability risk.

E0558: **A generalization of Ramsey's on discount rate with regime-switching by martingale approach**
Presenter: **Qi Li**, Pusan National University, Korea, South

The aim is to generalize the following Ramsey's rule on discount rate with regime-switching by a martingale approach: the discount rate is the sum of the rate of pure time preference and the product of the consumption elasticity of marginal utility and the consumption growth rate. We first characterize an implicit form of the equivalent martingale measure under regime switch. The ability to derive the unique state price density concerning the risk-neutral Poisson intensity under regime switching in closed form is another distinguishing feature. We show that Ramsey's rule can be extended to regime-dependent interest-rate formulas for discounting future regime changes. Furthermore, we also show that the effect of pure time preference is overwhelmingly dominated by the effect of the regime-switching parameter. This is closely related to the consumption smoothing consequences across regimes in the long term.

E0590: **Optimal consumption and savings decisions with disastrous income risk: Revisiting Rietz's rare disaster risk hypothesis**
Presenter: **Chusu He**, University of Bath, United Kingdom
Co-authors: Alistair Milne, Seyoung Park

An analytically tractable framework is developed for optimal consumption and savings decisions with disastrous income risk. In the context of Rietz's rare disaster risk hypothesis, we explain high-risk premia through a new channel of idiosyncratic income risk premium. The effects of idiosyncratic income risk premium are revealed in the agent's optimal decisions by the precautionary savings, thereby making the agent consume less and save more. We also investigate the important role of insurance with a focus on the recovery of income in a disaster. We highlight how the extent of the disastrous income risk to which the agent is exposed and her income recovery in the income shock jointly affect the agent's optimal decisions. Overall, the availability of insurance can be particularly important for both the poor and the wealthy in the sense that they could even consume more, save less, and invest more in the income shock as long as their future income is (partly) recovered.

EO139 Room 104 (Hybrid 4) METHODS FOR SURVIVAL DATA ANALYSIS II Chair: Takeshi Emura

E0897: **Survival analysis with several classes of functional data as covariates**
Presenter: **Yuko Araki**, Tohoku University, Japan

The survival analysis which contains several classes of functional data is investigated. First, we identify the length and class of individual trajectories by model selection and the proposed functional clustering method. Further, as a second stage, class information is used in the Cox proportional hazards regression models to assess the risk of mortality during the follow-up period. We assess the performance of the proposed model in a simulation study and its application to the long-term cohort study in Japan.

E0163: **Mediation analysis for mixture Cox proportional hazards cure models**
Presenter: **Xinyuan Song**, Chinese University of Hong Kong, Hong Kong

Mediation analysis aims to decompose a total effect into specific pathways and investigate the underlying causal mechanism. Although existing methods have been developed to conduct mediation analysis in the context of survival models, none of these methods accommodates the existence of a substantial proportion of subjects who never experienced the event of interest, even if the follow-up is sufficiently long. Mediation analysis is considered for mixture Cox proportional hazards cure models that cope with the cure fraction problem. Path-specific effects on restricted mean survival time and survival probability are assessed by introducing a partially latent group indicator and applying the mediation formula approach in a three-stage mediation framework. A Bayesian approach with P-splines for approximating the baseline hazard function is developed to conduct analysis. An application of the Alzheimer's Disease (AD) Neuroimaging Initiative dataset investigates the causal effects of APOE-epsilon 4 allele on AD progression.

E0543: **Optimal stratification of survival data via Bayesian nonparametric mixtures**
Presenter: **Bernardo Nipoti**, University of Milan Bicocca, Italy

The stratified proportional hazards model represents a simple solution to account for heterogeneity within the data while keeping the multiplicative effect on the hazard function. Strata are typically defined a priori by resorting to the values taken by a categorical covariate. A general framework is proposed, which allows for the stratification of a generic accelerated lifetime model, including as a special case the Weibull proportional hazard model. The stratification is determined a posteriori by taking into account that strata might be characterized by different baseline survivals as well as different effects of the predictors. This is achieved by considering a Bayesian nonparametric mixture model and the posterior distribution it induces on the space of data partitions. The optimal stratification is then identified by means of the variation of information criterion and, in turn, stratum-specific inference is carried out. The performance of the proposed method and its robustness to the presence of right-censored observations are investigated by means of an extensive simulation study. Our findings are further illustrated by analysing a data set extracted from the University of Massachusetts AIDS Research Unit IMPACT Study.

EO065 **Room 105 (Hybrid 5)** **INFERENCE FOR DYNAMIC SYSTEMS** **Chair: Francoise Anne Kemp**

E0801: **Using Bayes factors to compare dynamical models of hydrological systems**
Presenter: **Mingo Ndiwago Damian**, University of Luxembourg, Luxembourg
Co-authors: Remko Nijzink, Christophe Ley, Stanislaus Schymanski, Jack Hale

Hydrologists are often faced with selecting the 'best' model from a set of competing rainfall-runoff models that differ widely in their complexity and ability to reproduce both past and future data. The Bayes factor is one tool for selecting between models. It is relatively robust and easy–to–use as it implicitly and automatically balances model complexity and goodness of fit to data under a few simplifying assumptions. However, it requires the computation of the marginal likelihood which is a very expensive and difficult integration problem. This expense can be attributed to three factors; the necessity of many likelihood calculations with moderate run–times due to the repeated solution of the rainfall-runoff model; the multi–modal and highly correlated nature of the posterior; and finally the inherent difficulty of the marginal likelihood integration problem. We show that by combining recent advances in differentiable programming languages, modern gradient-based Markov Chain Monte Carlo algorithms and thermodynamic integration, the Bayes factor is now a practical and robust tool for comparing rainfall–runoff models. We illustrate our approach with the problem of choosing from a set of HBV-type models with increasing dynamical complexity calibrated against both synthetically generated and real runoff data. We show that the Bayes factor not only selects a parsimonious model but can also be computed using a reasonable amount of computational resources.

E0974: **Dynamical modelling of COVID-19 pandemic**
Presenter: **Francoise Anne Kemp**, University of Luxembourg, Luxembourg
Co-authors: Daniele Proverbio, Atte Aalto, Christophe Ley, Jorge Goncalves, Alexander Skupin, Stefano Magni

Worldwide 488,532,505 confirmed cases and 6,144,226 dead people have been identified to be infected with SARS-CoV-2 (severe acute respiratory syndrome coronavirus 2) by 1st April 2022. COVID-19 is a new strain of coronavirus SARS-COV-2 (severe acute respiratory syndrome coronavirus 2). The first outbreak was reported in December 2019 in Wuhan, China. COVID-19 is spreading across the globe and was declared as a pandemic by WHO in March 2020. Epidemiological modelling and forecasting of the evolution of the epidemic represent a useful tool to assist in designing better non-pharmaceutical measures and prepare for the stress on healthcare systems. We develop a new compartment-model based on Susceptible-Exposed-Infectious-Removed (SEIR) model and described by ordinary differential equations to simulate the time evolution of the number of positive cases, the number of people entering hospital and intensive care unit (ICU) and of deaths. The model is fitted to Luxembourgish time-series data and provides fruitful insights into the dynamics behind this pandemic and the evolution of immunity in the society. Overall, we show that mathematical modelling represents a powerful tool to test mechanistic hypotheses and to identify underlying principles of complex biological systems.

E0595: **Dynamic conditional correlation models with time-varying parameters incorporating realized covariance matrices**
Presenter: **Hideto Shigemoto**, Kwansei Gakuin University, Japan
Co-authors: Takayuki Morimoto

Novel realized dynamic conditional correlation (Realized DCC) models with measurement errors are proposed to forecast the covariance of asset returns. The aim is to estimate the conditional covariance of asset returns and to ensure the bound of the forecasted correlation matrix by incorporating the measurement error in estimating the conditional variance and using the BEKK and HAR models to estimate the conditional correlation matrix. The models that we propose can incorporate measurement errors into not only variance estimations but also correlation estimations. These models can keep the persistence of volatility and correlation at high levels when the asymptotic variances of realized volatility and realized correlation are small. By incorporating measurement errors, models can decrease the persistence when the asymptotic variances are large. We show, in our empirical results, a significant improvement in predictive performance over the benchmark BEKK-type model and the usual Realized DCC model.

EO017 **Room Virtual R1** **ECONOMETRICS OF SPILLOVER EFFECTS AND SOCIAL INTERACTIONS** **Chair: Ryo Okui**

E0232: **Causal inference with noncompliance and unknown interference**
Presenter: **Takahide Yanagi**, Kyoto University, Japan
Co-authors: Tadao Hoshino

The aim is to investigate a treatment effect model in which individuals interact in a social network and they may not comply with the assigned treatments. We introduce a new concept of exposure mapping, which summarizes spillover effects into a fixed dimensional statistic of instrumental variables, and we call this mapping the instrumental exposure mapping (IEM). We investigate identification conditions for the intention-to-treat effect and the average causal effect for compliers, while explicitly considering the possibility of misspecification of IEM. Based on our identification results, we develop nonparametric estimation procedures for the treatment parameters. Their asymptotic properties, including consistency and asymptotic normality, are investigated using an approximate neighborhood interference framework. For an empirical illustration of our proposed method, we revisit experimental data on the anti-conflict intervention school program.

E0261: **Recovering latent linkage structures and spillover effects with structural breaks in panel data models**
Presenter: **Wendun Wang**, Erasmus University Rotterdam, Netherlands

The aim is to capture time-varying spillover effects in a panel data setting. We consider panel models where the outcome of a unit not only depends on its own characteristics and also the characteristics of other units (spillover effects). The effect of own characteristics can be unit-specific or homogeneous (common effects). We allow the linkage structure, i.e., which units interact with which, to be latent, and the structure and the spillover effects may both change at an unknown breakpoint. To estimate the breakpoint, spillover and common effects, we solve a penalized least squares optimization and employ double machine learning procedures to improve the convergence and inference. We establish the super consistency of the breakpoint estimator and provide asymptotic properties of estimated spillover and common effects. We illustrate the theory via simulated and empirical data.

E0285: **Production network positions and risk premia: A semiparametric approach**
Presenter: **Chao Yang**, Shanghai University of Finance and Economics, China

Ever since the financial recession in 2008, more and more attention has been paid to the role that production networks play in translating idiosyncratic risks. According to the theoretical models, owing to the input-output linkages, the risks from a supplier and/or a customer of production intermediaries may be amplified and propagated. Different positions in the production network would then induce heterogeneous exposure to this network risks. By adopting a flexible semiparametric characteristics-based factor model, we investigate how the factor loadings would be influenced by firms' positions in the production network, such as in-degree, out-degree, and centrality, in the Chinese asset market.

EO155 **Room Virtual R3** **RECENT ADVANCES IN FINANCIAL ECONOMETRICS** **Chair: Seok Young Hong**

E0709: **Testing for jumps in a discretely observed price process with endogenous sampling times**
Presenter: **Shifan Yu**, Lancaster University, United Kingdom
Co-authors: Yifan Li, Ingmar Nolte, Sandra Nolte

A new nonparametric test is proposed to determine whether finite-activity jumps are present in a discretely observed price process. For a univariate semimartingale, we introduce the concept of censored increments for observations recursively sampled at exit times with a symmetric double barrier, and design a standardized test statistic to compare the sample moments of censored and uncensored increments. Simulation results show that our test has better finite-sample performance than other commonly used calendar time-based jump tests with a similar level of sampling sparseness, and is fairly robust to measurement errors including market microstructure noise and price staleness. Our empirical study provides strong evidence for the presence of jumps for 10 NYSE stocks in 2020, but the jumps are much less frequent than that suggested by some existing tests.

E0712: **Adjusted-range self-normalized autocorrelation tests**
Presenter: **Xiaolu Zhao**, Dongbei University of Finance and Economics, China
Co-authors: Jiajing Sun, Yongmiao Hong, Oliver Linton, Xiaolu Zhao

Two autocorrelation tests are proposed using the adjusted-range based self-normalization, extending previous work. Despite focusing on testing the autocorrelation, our method can be generalized to testing the significance of the class of approximately linear statistics, such as the marginal mean, marginal variance and quantiles. Through comprehensive simulation studies on both standard time series and count series data, we confirm that our adjusted-range based tests are particularly suitable for detecting the presence of serial correlation, significantly outperforming existing self-normalized tests. Empirical results on the periodic presence of unit root in realized volatility series of the world's major stock indices as well as on testing the serial autocorrelation in COVID-19 count series further confirm the validity of our approach.

E0886: **Beyond the candlesticks: Exploiting price paths via the MAED statistic**
Presenter: **Yifan Li**, The University of Manchester, United Kingdom
Co-authors: Ingmar Nolte, Sandra Nolte

The Maximal rAnge-rEturn Divergence (MAED) statistic is proposed, which is defined as the maximum distance between the price range and the absolute return on a fixed time interval. The statistic can be easily constructed when high-frequency transaction data is available. The MAED statistic summarizes the inward movement of price paths, which contains substantially different information to the candlestick data (i.e., high, low, open and close price in an interval) that mainly capture the outward movement of prices. We propose a spot volatility estimator based on the MAED-augmented candlestick data and establish its asymptotic properties in the fixed-k asymptotic setting with discrete price observations. Our analytical and simulation results show that our MAED-augmented estimator can reduce the asymptotic variance of the optimal candlestick-based spot volatility estimator by as much as 40%. The MAED statistic is robust to jump by construction, which allows us to construct a spot jump test to detect jumps in small time intervals.

EO187 **Room Virtual R4** **STATISTICAL METHODOLOGY FOR PERSONALISING ONLINE CONTENT** **Chair: Cristina Mollica**

E0501: **Learning to rank under multinomial logit choice**
Presenter: **James Grant**, Lancaster University, United Kingdom
Co-authors: David Leslie

Learning the optimal ordering of content is an important challenge in website design. The learning to rank (LTR) framework models this problem as a sequential problem of selecting lists of content and observing where users decide to click. Most previous work on LTR assumes that the user considers each item in the list in isolation, and makes binary choices to click or not on each. We introduce a multinomial logit (MNL) choice model to the LTR framework, which captures the behaviour of users who consider the ordered list of items as a whole and make a single choice among all the items and a no-click option. Under the MNL model, the user favours items that are either inherently more attractive or placed in a preferable position within the list. We propose upper confidence bound algorithms to minimise regret in two settings - where the position-dependent parameters are known, and unknown.

E0510: **Recommender systems, bandits and Bayesian neural networks**
Presenter: **Simen Eide**, University of Oslo / Schibsted, Norway

Internet platforms consist of millions or billions of different items that users can consume. To help users navigate this landscape, recommender systems have become an important component in many platforms. The aim of a recommender system is to suggest the most relevant content on the platform to the user based on previous interactions the user has done with the platform. A model used in recommender systems faces multiple sources of uncertainty: There are limited interactions per user, the signals a user makes may be noisy and not always reflect her preferences, and new items may be introduced to the platform giving few signals on these items as well. The focus will be on model uncertainty and decision making in the recommender systems. We will discuss various ways to quantify, reduce and exploit these uncertainties through the use of Bayesian neural networks, hierarchical priors and different recommender strategies.

E0977: **Pseudo-Mallows for preference learning and personalized recommendation**
Presenter: **Qinghua Liu**, University of Oslo, Norway

The Mallows model has been proven to be useful for learning personal preferences from highly incomplete data and be applied to recommender systems. However, inference based on MCMC is slow, preventing its use in real time applications. We propose the Pseudo-Mallows distribution over the set of all permutations of n items, to approximate the posterior distribution with a Mallows likelihood. The Pseudo-Mallows distribution is a product of univariate discrete Mallows-like distributions, constrained to remain in the space of permutations. In a variational setting, we optimise the variational order parameter by minimising a marginalized KL-divergence. We propose an approximate algorithm for this discrete optimization, and conjecture a certain form of the optimal variational order that depends on the data. Empirical evidence and some theory support our conjecture. Sampling from the Pseudo-Mallows distribution allows fast preference learning, compared to alternative MCMC based options, when the data exists in form of partial rankings of the items or of clicking on some items. Through simulations and a real life data case study, we demonstrate that the Pseudo-Mallows model learns personal preferences very well and makes recommendations much more efficiently, while maintaining similar accuracy compared to the exact Bayesian Mallows model.

EO161 Room Virtual R5 COPULA MODELS AND APPLICATIONS Chair: Aurora Gatto

E0784: **The impact of COVID-19 pandemic shock on the correlation between energy markets and EU ETS**
Presenter: **Hao Wang**, Jilin University, China
Co-authors: Runfan Chen

The COVID19 pandemic has created significant shocks in the energy market and carbon emissions. We study the change of correlations between energy markets and EU ETS to investigate the possible implications of the pandemic. The Vine Copula is employed to research the static correlations, and the rolling approach is adopted to study the dynamic nonlinear tail dependence. We found that the high-dimensional correlation structures remain the same, while the correlation has been reinforced after the epidemic. The gas market has the strongest direct correlation with EU ETS, followed by the relationship between EU ETS and the oil market. On the other hand, upper tail dependence is stronger than the lower one. And the probability of upper tail risks in EU ETS brought by the gas market is highest, which happened mostly after the pandemic. After the COVID19 pandemic, the correlations between EU ETS and energy markets have been strengthened, especially the correlation and upper tail dependence between EU ETS and the gas market. And the energy structure is the radical cause of this divergence. Our findings suggest that energy price risk, particularly the extreme price risk of gas, has become an impediment to Europe's transition to a low-carbon economy.

E0870: **Kendall conditional value-at-risk with application to the Italian energy market**
Presenter: **Aurora Gatto**, University of Salento, Italy
Co-authors: Fabrizio Durante, Elisa Perrone

The Conditional Value-at-Risk (CoVaR) is a dependence-adjusted version of the Value at Risk (VaR) to quantify the risk of a random variable Y with respect to another random variable X. We take into account a multivariate modification of CoVaR based on the Kendall distribution function, where the stress event is related to multiple random variables. In detail, we discuss two possible hazard scenarios that generalize the standard CoVar and use the copula theory to derive the corresponding risk quantities. As an application of the proposed methodology, we consider the main Italian Energy companies listed on the stock exchange to demonstrate how the multivariate modification of CoVaR can be useful to analyze the resilience of a system when some parts of it are under distress.

E1034: **Convergence of copulas revisited: Different notions of convergence and their interrelations**
Presenter: **Nicolas Dietrich**, Universitat Salzburg, Austria
Co-authors: Juan Fernandez Sanchez, Wolfgang Trutschnig

Building upon the one-to-one relation between the family $\mathcal{C}$ of bivariate copulas and Markov operators we consider the metric OP_p corresponding to the L_p, $p \in [1\infty]$ operator norm and study its interrelation with other metrics on $\mathcal{C}$. In particular, we prove the surprising result that OP_1 convergence implies weak conditional convergence of the transposed copulas and establish the fact that the topology induced by OP_∞ is strictly finer than the topology induced by weak conditional convergence.

EC461 Room 101 (Hybrid 1) CONTRIBUTIONS IN BAYESIAN METHODS (IN-PERSON) Chair: Shogo Nakakita

E0833: **Bayesian ridge estimators based on copula-based joint prior distributions for logistic regression parameters**
Presenter: **Yuto Aizawa**, Kitasato university, Japan
Co-authors: Hirofumi Michimae

Ridge regression was originally proposed as an alternative to OLS regression in order to address multicollinearity in linear regression and later extended to logistic regression and also to Cox regression. In the Bayesian framework, the ridge estimator is interpreted as a Bayesian posterior mode when the regression coefficients have multivariate normal priors. We previously proposed vine copula-based joint priors on the regression coefficients in linear regression including an interaction, which promote the use of ridge regression because the interaction term produces multicollinearity. We showed that the vine copula-based priors improved the estimation accuracy over the multivariate normal prior, and would be a promising approach in other types of regression, such as logistic regression. We propose the vine copula-based prior for Bayesian ridge estimators under the logistic model. Especially, we focus on the case of two covariates and their interaction term. A simulation study was carried out to compare the performance of four prior (the Clayton, Gumbel, Gaussian priors and the trivariate normal prior) on the regression coefficients. These simulation studies proved that the Archimedean (the Clayton, Gumbel) copula priors showed more accurate estimates in the presence of multicollinearity compared with the other priors.

E0908: **Bayesian fused lasso and Bayesian HORSES via horseshoe prior**
Presenter: **Yuko Kakikawa**, The University of Electro-Communications, Japan
Co-authors: Kaito Shimamura, Shuichi Kawano

Bayesian fused lasso is one of the Bayesian methods to estimate regression coefficients of a linear regression model. It shrinks both regression coefficients and their successive differences simultaneously by assuming Laplace distributions on both of them. It enables the estimation of regression coefficients with the successive ones fused. However, Bayesian fused lasso tends to over-shrink regression coefficients and their successive differences which are not supposed to be near zero. To overcome this problem, we assume a horseshoe prior on the difference of successive regression coefficients and construct the Bayesian fused lasso modeling. Because horseshoe prior has an infinite spike at zero and a Cauchy-like tail, the proposed method enables us to prevent over-shrinkage of those differences. We also propose a Bayesian hexagonal operator for regression with shrinkage and equality selection (HORSES) with horseshoe prior, which imposes priors on every pair of differences of regression coefficients. To obtain the estimates of the parameters, we develop a Gibbs sampler by using a hierarchical expression of a Laplace prior and a horseshoe prior. Monte Carlo simulations and an application to real data are conducted to investigate the effectiveness of the proposed method.

E0513: **Haar-Weave-Metropolis kernel**

Presenter: **Kengo Kamatani**, ISM, Japan

Co-authors: Xiaolin Song

Recently, many Markov chain Monte Carlo methods have been developed with deterministic reversible transform proposals inspired by the Hamiltonian Monte Carlo method. The deterministic transform is relatively easy to reconcile with the local information (gradient etc.) of the target distribution. However, as the ergodic theory suggests, these deterministic proposal methods seem to be incompatible with robustness and lead to poor convergence, especially in the case of target distributions with heavy tails. On the other hand, the Markov kernel using the Haar measure is relatively robust since it learns global information about the target distribution by introducing global parameters. However, it requires a density preserving condition, and many deterministic proposals break this condition. We carefully select deterministic transforms that preserve the value of the density function and create a Markov kernel, the Weave-Metropolis kernel, using the deterministic transforms. By combining with the Haar measure, we also introduce the Haar-Weave-Metropolis kernel.

EC432 Room 106 (Hybrid 6) CONTRIBUTIONS IN HIGH DIMENSIONAL AND COMPLEX DATA (IN-PERSON) Chair: Kazuyoshi Yata

E0677: **Testing the equality of topic distribution between documents of a corpus**

Presenter: **Louisa Kontoghiorghes**, Kings College London, United Kingdom

Co-authors: Ana Colubi

Topic modelling is a well-known text mining technique to identify the themes covered in a set of documents. We introduce a methodology to test whether two documents of a given corpus are homogeneous with respect to the topics they cover. The suggested approach uses Latent Dirichlet Allocation (LDA) to estimate the topic distributions and the Kullback-Leibler divergence to measure the distance between the distributions. Since the sampling distribution of the proposed statistics is unknown, a (frequentist) bootstrap test is suggested. The methodology is illustrated using scientific abstracts from the CMStatistics conference.

E0983: **Monitoring time dependent image processes**

Presenter: **Yarema Okhrin**, University of Augsburg, Germany

Monitoring techniques are developed for high dimensional image processes with temporal dependence. This problem is of great practical importance in visual quality control of production processes. It is assumed that the pixel intensities follow a spatial AR process. In the out-of-control state we expect a location shift. To overcome computational infeasibility of the implementation we focus on dimension reduction techniques based on averaging. Moreover, we derive necessary and sufficient conditions for the original process and spatially averaged process to follow the same spatial AR process. For monitoring purposes we apply several multivariate EWMA control charts and evaluate their performance in extensive simulations.

E0734: **Multi-task learning for compositional data based on sparse network lasso regularization**

Presenter: **Akira Okazaki**, The University of Electro-Communications, Japan

Co-authors: Shuichi Kawano

A network lasso enables us to construct a model for each sample, which is known as multi-task learning. It is used in various fields of research, in particular life science in which, the obtained data contain heterogeneity that varies among samples. In such a case, general models that are common to all samples fail to extract the effective information, which is related to heterogeneity. On the other hand, compositional data, which consist of the proportions of a composition, are used in the fields of geology and life science for microbiome analysis. Existing methods for multi-task learning cannot be applied to compositional data due to their intrinsic properties. In this research, we propose a multi-task learning method for compositional data using a sparse network lasso regularization. We focus on a symmetric form of the log-contrast model, which is a regression model with compositional covariates. The symmetric form is extended to the locally symmetric form in which each sample has a different regression coefficient vector. These regression coefficient vectors are clustered by the network lasso regularization and selected by the L1-norm regularization. The effectiveness of the proposed method is shown through simulation studies and application to gut microbiome data.

EC431 Room 107 (Hybrid 7) CONTRIBUTIONS IN TIME SERIES (IN-PERSON) Chair: Junichi Hirukawa

E0228: **Nonlinear scalar BEKK**

Presenter: **Bilel Sanhaji**, University Paris 8, France

A nonlinear conditional covariance model is proposed with five scalars to estimate. We develop the asymptotic theory of the quasi maximum likelihood estimator. We propose Lagrange Multiplier and Likelihood Ratio tests for nonlinearity in conditional covariances in multivariate GARCH models. We also show asymptotic properties through Monte Carlo simulations and provide empirical illustrations.

E0903: **Detecting change-points in noisy data sequences with continuous piecewise structures**

Presenter: **Yiming Ma**, University of Otago, New Zealand

Co-authors: Andreas Anastasiou, Ting Wang, Fabien Montiel

A new method, called singular spectrum analysis isolate-detect (SSAID), is proposed to detect change-points in noisy data sequences with an underlying continuous piecewise structure. In contrast to existing parametric change-point detection methods for signals with a predefined piecewise structure, SSAID does not require prior knowledge of the exact nature of the structural changes; for example, it can identify change-points in noisy piecewise-exponential or piecewise-quadratic signals equally well. SSAID is motivated by the need for automated detection of slow slip events (SSEs), which are a type of slow earthquakes. The SSE data have a typical piecewise-non-linear trend, but the exact structure is unknown. SSAID recasts the problem of identifying SSEs as that of detecting change-points in a piecewise-linear signal. This is achieved by obscuring the deviation from the piecewise-linear in the underlying piecewise non-linear signal. The results on both simulated and real SSE data, and simulated data with various piecewise structures suggest that our method can successfully detect change-points in signals with a wide range of piecewise structures. We further demonstrate the performance of SSAID using real data such as the number of COVID-19 daily confirmed cases in the United States and the monthly S&P 500 close price index.

Sunday 05.06.2022 16:50 - 18:30 Parallel Session J – EcoSta2022

EV457 Room Virtual R5 CONTRIBUTIONS IN TIME SERIES II Chair: Philip Yu

E0850: **Similarity-based recession predictions in different monetary policy conditions**
Presenter: **Visa Kuntze**, University of Turku, Finland
Co-authors: Henri Nyberg, Samuel Rauhala

A nonparametric similarity-based approach is developed to predict the state of the business cycle in different interest rate environments. As an alternative to the existing parametric logit and probit models, our approach provides several methodological advances and new perspectives on the usefulness of the interest rate level and the term spread as leading indicators. The empirical findings, obtained with international data from the U.S., euro area and Japan, show that the predictability of business cycles is dependent on the stance of monetary policy such as the level of the short-term interest rates.

E0718: **Time series entropy: Clustering for decision-making**
Presenter: **Miguel Angel Ruiz Reina**, Universidad de Malaga, Spain

The multidimensional classification generates information gaps for researchers, practitioners, or businesses; Information Theory clustering is a solution to understanding seasonal decision-making time series. We propose an automatic clustering system based on the multi-optional based on Shannon entropy, combining techniques that relax the usual assumptions of statistics to understand the data set. Clustering metric methods are crucial for many real-world applications, and distance metrics provide learning models with better performance than is generally achieved. This unsupervised classification method automatically adjusts spatio-temporal observations and organises the data set for decision-making. The empirical field of application is tourist accommodation decision-making among hotels, tourist apartments, campsites and rural apartments for foreign tourists visiting Spain from January 2001 to January 2022. The intracluster verification criterion confirms the similarity of the members of the group. In this way, policymakers could adjust their offers or impact policies based on the seasonal typology studied. It is possible to convert time series from High-Dimensional Time Series to Reduction-Dimensional Time Series by recognising behaviour patterns of foreign tourist demand in Spain. This statistical learning model allows for building analysis models on large volumes of data and providing unsuspected knowledge in the initial exploratory analysis of the data.

E0669: **A clustering approach for analysing the impact of COVID-19 on stock market volatility**
Presenter: **Jorge Caiado**, University of Lisbon, Portugal
Co-authors: Franscisco Santos

The COVID-19 impact on U.S. stock market volatility with a focus on 11 S&P 500 sectors is investigated. For this purpose, we introduced a model feature-based method for clustering financial time series that accounts the useful information about the dependence structure of their conditional volatilities. This clustering approach consists in fitting parsimonious threshold GARCH models to the sector stock returns and then computing the distance matrix between the autocorrelations of their estimated conditional variances. By using hierarchical and non-hierarchical clustering methods, we conclude that there is a clear change in the composition of each cluster from the period before the first U.S. COVID-19 case to the period during the pandemic, leading to the conclusion that the similarities or distances between sectors have undergone a significant change and the industries most affected by the pandemic were the Hotels, Automobile and Airline.

E0842: **On the asymptotic behavior of bubble date estimators**
Presenter: **Eiji Kurozumi**, Hitotsubashi University, Japan
Co-authors: Anton Skrobotov

The three-regime bubble model is extended to allow the fourth regime followed by the unit root process after recovery. We provide the asymptotic and finite sample justification of the consistency of the collapse date estimator in the two-regime AR(1) model. The consistency allows us to split the sample before and after the date of collapse and to consider the estimation of the date of exuberation and date of recovery separately. We have also found that the limiting behavior of the recovery date varies depending on the extent of explosiveness and recovery.

EI011 Room 101 (Hybrid 1) RISK MEASUREMENT FOR SUSTAINABLE FINANCE (VIRTUAL) Chair: Monica Billio

E0297: **Responsible investing under ambiguity induced by climate risk**
Presenter: **Monica Billio**, University of Venice, Italy
Co-authors: Massimo Guidolin, Francesco Rocciolo

The aim is to propose a theory of responsible investing under conditions of ambiguity induced by climate risk by studying the portfolio allocation problem solved a smoothly ambiguity averse representative agent. Within this setting, we find that the ambiguity risk premium is a strictly decreasing function of the environmental scores of the assets. Ambiguity-averse investors behave as environmentally motivated agents who allocate their wealth according to a mean-variance-ambiguity efficient frontier and their attitude towards risk and ambiguity. The agents rationally choose green portfolios in order to diminish their exposition towards ambiguity and maximize their ambiguity Sharpe ratio. Our theoretical predictions are consistent with the empirical literature on the rewards-to-risks trade-off of responsible investment.

E0578: **Climate-related transition risk in the European CDS market**
Presenter: **Michele Costola**, Ca' Foscari University of Venice, Italy
Co-authors: Katia Vozian

The European low-carbon transition started in the last decades and is accelerating to reach net-zero by 2050. We study how the climate-related transition risk of a European large corporate firm relates to its CDS-implied credit risk for different time horizons. We find that firms with higher GHG emissions have higher CDS-implied credit risk, even at the 30-year horizon, particularly after the 2015 Paris Agreement. The results suggest that the European CDS market is already pricing to some extent the exposure to transition risk of a firm at different time horizons but ignores firms transition risk management efforts and their exposure to the EU ETS.

E0714: **Stock market risk under transition scenarios for the euro area**
Presenter: **Claudio Morana**, Universita di Milano Bicocca, Italy

The evolution of stock market volatility since the inception of the euro area is assessed. The period is of interest and informative as it encompasses various episodes of financial disruption, from the stock market crisis in the early 2000s to the 2007-2008 financial crisis and ensuing Great Recession, the sovereign debt crisis, the Covid-19 pandemics through the most recent geopolitical crisis. The volatility analysis is functional not only for the understanding of the evolution of stock market risk over the first twenty years of existence of the euro area and the progress of financial integration but also to uncover evidence on evolving systematic risk factors, in relation to the increasing relevance of climate change and transition risk. The analysis is carried out within a conditional asset pricing framework, directly exploiting information provided by the first-step volatility analysis. The paper is in progress and no results are currently available.

EO203 Room 102 (Hybrid 2) THE STEIN METHOD AND APPLICATIONS Chair: Xiao Fang

E0274: **A unifying view on kernel stein discrepancy tests for goodness-of-fit**
Presenter: **Wenkai Xu**, University of Oxford, United Kingdom

Non-parametric goodness-of-fit testing procedures based on kernel Stein discrepancies (KSD) are promising approaches to validate general unnormalised distributions in various scenarios. We introduce various KSD-based goodness-of-fit testing including Euclidean data, survival data, directional data and compositional data. Standardisation methods have been developed in Stein's method literature to study approximation properties for normal distributions. We apply techniques inspired by the standardisation idea that enable us to present a unifying view to theoretically compare and interpret different Stein operators in performing the KSD-based goodness-of-fit testing. The unifying framework is also useful as a guide to develop novel KSD-based tests. Different choice of standardisation functions results in different Stein operators, whose corresponding KSD choices have a considerable effect on the test performances. We discuss the operator choice and kernel choice for KSD-based testing procedures. We show experimental results demonstrate that these KSD tests control type-I error well and achieve higher test power than existing approaches, including the test based on maximum-mean-discrepancy (MMD).

E0519: **Asymptotic mixed normality of the realized covariance matrix in high-dimensions**
Presenter: **Yuta Koike**, University of Tokyo, Japan

Asymptotic mixed normality of the realized covariance matrix for a multi-dimensional continuous semimartingale is established, where the dimension may be much larger than the sample size. More precisely, in such a setting, a mixed normal approximation of the error distribution is shown in terms of the Kolmogorov distance. The proof is based on a variant of the Chernozhukov-Chetverikov-Kato theory on high-dimensional central limit theorems for sums of independent random vectors, where the theory is adapted to random asymptotic covariance matrices with the help of the Malliavin-Stein method. An application to testing for residual sparsity in a continuous-time factor model is presented.

E0548: **Cramer-type moderate deviations under local dependence**
Presenter: **Songhao Liu**, Southern University of Science and Technology, China
Co-authors: Zhuo-Song Zhang

The aim is to establish Cramer-type moderate deviation theorems for sums of locally dependent random variables and combinatorial central limit theorems. Under some mild exponential moment conditions, optimal error bounds and convergence ranges are obtained. The main results are more general or shaper than the existing results in the literature. The main results follow from a more general Cramer-type moderate deviation theorem for dependent random variables without any boundedness assumptions, which is of independent interest. The proofs couple Stein's method with a recursive argument.

E0716: **Approximations to the ergodic measure of stable SDE via EM scheme**
Presenter: **Lihu Xu**, University of Macau, China

An Euler-Maruyama scheme is developed for a stable stochastic differential equation to approximate its ergodic measure, and give an error bound in Wasserstein-1 distance. The error bound is optimal.

EO353 Room 103 (Hybrid 3) DATA SCIENCE AND OTHER DEVELOPMENTS IN FINANCIAL MODELLING Chair: Rogemar Mamon

E0565: **LSTM in varying regimes: How to combine hidden Markov and machine learning models for financial risk management**
Presenter: **Christina Erlwein-Sayer**, University of Applied Sciences HTW Berlin, Germany
Co-authors: Stefanie Grimm

Time series analysis and machine learning techniques are combined to forecast financial time series. In our hidden Markov model, hidden market regimes are detected and filtered out from observed financial time series. These states are subsequently incorporated into a long-short-term memory neural network (LSTM). Through this, we develop a model to forecast corporate credit spreads over changing market regimes within an LSTM; switching regimes are included as a feature to the neural network. This HMM-LSTM model is calibrated to corporate credit spreads from three European countries. The performance of the LSTM is analysed and compared to the accuracy of an LSTM without regime-switching information. Furthermore, we propose an HMM-LSTM mixture of experts' model, where regime-switching information acts as a gating function to activate a neural network. Applications of this approach to time series forecasting in electricity markets are shown. Our findings show that in most cases the LSTM performance is improved when regime information is added.

E0633: **A residual network for valuing large portfolios of variable annuities**
Presenter: **Heng Xiong**, Wuhan University, China

The valuation of large variable annuity portfolios is a central concern for insurers considering that the commonly used Monte-Carlo (MC) simulation is computationally intensive. A spatial interpolation method was developed recently to significantly reduce the calculation time for valuation. However, such a method relies heavily on a predefined distance function. Thus, it is replaced by a neural network (NN) strategy that could select the optimal distance function automatically. We present the residual portfolio valuation network (ResPoNet), which outperforms the traditional NN by adding a loss of weight item. ResPoNet maintains the universal meaning of the distance function. The high performance of ResPoNet is also due to the insertion of a residual connection into the network training process, which in turn enables the network to learn the attributes of insurance policies. Our numerical experiments illustrate that the proposed approach effectively smooths the training process and significantly improves the accuracy of the valuation when benchmarked to the original NN.

E0767: **Jumping hedges on the strength of the Mellin transform**
Presenter: **Dr M Rodrigo**, University of Wollongong, Australia

With more looming uncertainties in our present financial climate and environment, models with jump-diffusion more than ever are necessary. They are suited to reproduce the large and sudden fluctuations in the level of the underlying variable, and mimic various statistical properties in observed time series. The jump-diffusion modelling setup, however, brings complexity to the valuation and hedging of derivative securities. We delve into the subject of hedging along with the illustration of hedgings' intimate interplay with pricing. We harness the power of the Mellin transform and its convolution property to establish hedging sensitivities that capture the many dimensions of risk in option positions. Our methodology allows for a wide class of generalised payoffs (i.e. all piecewise linear functions). Each hedging parameter is shown to contain the impact of jumps, and an explicit metric to quantify the jump risk is defined. The systematic and efficient calculations of higher-order sensitivities are demonstrated. Some examples to illustrate the ease of implementation are given.

E0778: **A multivariate-index-driven anomaly detection system with supervised learning**
Presenter: **Rogemar Mamon**, University of Western Ontario, Canada

A hybrid supervised learning system is developed to detect anomalies in multivariate time-series index data. The focus of the application is the determination of signs of possible crisis episodes that may wreak havoc on the financial market or economic stability. The proposed statistical-computing approach synthesises stochastic process modelling, hidden Markov filtering, Random Forest and XGBoost. Such an approach is capable of efficiently and accurately tracing simultaneously the financial stress indices (FSIs) of multiple countries and more importantly identifying anomalous FSIs behaviour that signals an impending financial instability. We show that our method is capable of dynamically making 6-step-ahead binary anomalous-normal classification predictions in a probabilistic sense for the benefit of industry practitioners and regulators. Our method, which also gives rise to an early-warning system, is benchmarked with other alternative financial-instability monitoring methods and its advantage is highlighted via various model validation measures.

EO349 Room 104 (Hybrid 4) RECENT ADVANCES IN BAYESIAN COMPUTATION AND APPLICATIONS Chair: Minh-Ngoc Tran

E0481: Robust generalised Bayesian inference for intractable likelihoods

Presenter: **Takuo Matsubara**, The Alan Turing Institute / Newcastle University, United Kingdom
Co-authors: Jeremias Knoblauch, Francois-Xavier Briol, Chris Oates

Generalised Bayesian inference updates prior beliefs using a loss function, rather than a likelihood, and can therefore be used to confer robustness against possible misspecification of the likelihood. We consider generalised Bayesian inference with a Stein discrepancy as a loss function, motivated by applications in which the likelihood contains an intractable normalisation constant. In this context, the Stein discrepancy circumvents evaluation of the normalisation constant and produces generalised posteriors that are either closed-form or accessible using standard Markov chain Monte Carlo. On a theoretical level, we show consistency, asymptotic normality, and bias-robustness of the generalised posterior, highlighting how these properties are impacted by the choice of Stein discrepancy. Then, we provide numerical experiments on a range of intractable distributions, including applications to kernel-based exponential family models and non-Gaussian graphical models.

E0488: Loss-based variational Bayes prediction

Presenter: **Ruben Loaiza-Maya**, Monash University, Australia
Co-authors: David Frazier, Gael Martin, Bonsoo Koo

A new method is proposed for Bayesian prediction that caters for models with a large number of parameters and is robust to model misspecification. Given a class of high-dimensional (but parametric) predictive models, this new approach constructs a posterior predictive using a variational approximation to a loss-based, or Gibbs, posterior that is directly focused on predictive accuracy. The theoretical behavior of the new prediction approach is analyzed and a form of optimality is demonstrated. Applications to both simulated and empirical data using high-dimensional Bayesian neural networks and autoregressive mixture models demonstrate that the approach provides more accurate results than various alternatives, including misspecified likelihood-based predictions.

E0522: Creating manifold structures to accelerate MCMC sampling

Presenter: **Alexandre Thiery**, National University of Singapore, Singapore

Consider the observation $y = F(x) +$ (noise) of a quantity of interest x. In Bayesian inverse problems, the quantity x typically represents the high-dimensional discretization of a continuous and unobserved field while the evaluations of the forward operator F involve solving a system of partial differential equations. In the low-noise regime, the posterior distribution concentrates in the neighborhood of a nonlinear manifold. As a result, the efficiency of standard MCMC algorithms deteriorates due to the need to take increasingly smaller steps. We present a constrained HMC algorithm that is robust in the low noise regime. Taking the observations generated by the model to be constraints on the prior, we define a manifold on which the constrained HMC algorithm generates samples. By exploiting the geometry of the manifold, our algorithm is able to take larger step sizes than more standard MCMC methods, resulting in a more efficient sampler. If time permits, we will explain how this idea can be extended to classification problems, by exploiting an auxiliary manifold.

E0604: The Bayesian learning rule

Presenter: **Mohammad Emtiyaz Khan**, RIKEN Center for AI project, Japan

It will be shown that a wide variety of machine-learning algorithms are instances of a single learning rule called the Bayesian learning rule. The key idea in deriving such algorithms is to approximate the posterior using candidate distributions estimated by using natural gradients. Different candidate distributions result in different algorithms and further approximations to natural gradients give rise to variants of those algorithms.

EO449 Room 105 (Hybrid 5) ADVANCES IN PRODUCTIVITY ANALYSIS AND MEASUREMENT Chair: Artem Prokhorov

E0435: Indirect inference of stochastic frontier models

Presenter: **Hung-pin Lai**, National Chung Cheng University, Taiwan

The standard method to estimate a stochastic frontier model is the maximum likelihood approach with the distribution assumptions of a symmetric two-sided stochastic error v and a one-sided inefficiency random component u. When v or u has a nonstandard distribution, such as v follows a generalized t distribution or u has a Chi-squared distribution, the likelihood function can be complicated. The aim is to use indirect inference to estimate the stochastic frontier models, where only least squares estimation is used. There is no need to derive the density or likelihood function, thus it is easier to handle a model with complicated distributions in practice. We examine the finite sample performance of the proposed estimator and also compare it with the standard maximum likelihood estimator as well as the maximum simulated likelihood estimator using Monte Carlo simulations. We found that our estimator performs quite well in finite samples.

E0437: Improving predictions of technical inefficiency

Presenter: **Robert James**, The University of Sydney, Australia
Co-authors: Artem Prokhorov, Peter Schmidt, Christine Amsler

The traditional predictor of technical inefficiency is a conditional expectation. We study whether, and by how much, the predictor can be improved by using auxiliary information in the conditioning set. To do so, we use simulations to study two types of stochastic frontier models. The first type is a panel data model where composed errors from past and future time periods contain information about contemporaneous technical inefficiency. The second type is when the stochastic frontier model is augmented by input ratio equations in which allocative inefficiency is correlated with technical inefficiency. We consider a standard kernel-smoothing estimator and a newer estimator based on a local linear random forest which helps mitigate the curse of dimensionality when the conditioning set is large. We also provide an illustrative empirical example.

E0471: How to measure the average rate of change

Presenter: **Mikhail Sokolov**, St Petersburg University; European University at St. Petersburg; RAS, Russia
Co-authors: Aleksandr Alekseev

A theory of the average rate of change (ARC) measurement is developed. Using an axiomatic approach, the conventional ARC measures (such as the difference quotient and the continuously compounded growth rate) are generalized in several directions: to outcome variables with arbitrary connected domains, to not necessarily time-shift invariant dependence on time, to more general (than an interval) time sets, to a path-dependent setting, and to a benchmark-based evaluation. We also revisit and generalize the relationship between the ARC measurement and intertemporal choice models.

E1033: **Assessing bank performance under volatile exchange rate**
Presenter: **Mikhail Mamonov**, CERGE-EI and CEBA, Czech Republic
Co-authors: Artem Prokhorov, Christopher Parmeter

The impact of currency fluctuations on bank performance when banks have non-trivial exposures to foreign currencies is studied. We use unique data on Russian banks between 2004 and 2020 to document that cost efficiency estimates are both downward biased and not rank preserving when these fluctuations are ignored. We find that the Demsetz efficient structure hypothesis holds only when currency fluctuations are accounted for. To ensure generalizability, we also perform a counterfactual exercise which does not rely on our unique bank-level data but rather employs banks' exposures to foreign currencies and nominal exchange rate volatility.

EO107 **Room 106 (Hybrid 6)** **HIGH-DIMENSIONAL INFERENCE AND DETECTION UNDER DEPENDENCE** **Chair: Ansgar Steland**

E0304: **Innovation algorithm of fractionally integrated processes and applications to the estimation of parameters**
Presenter: **Junichi Hirukawa**, Niigata University, Japan
Co-authors: Kou Fujimori

The long memory phenomena frequently occur in the empirical studies of various fields. The fractionally integrated process is one of the suitable candidates which appropriately represents the long memory property. There are two recursive algorithms for determining the one-step predictors of time series, that is, the Durbin-Levinson algorithm and the innovation algorithm. The Durbin-Levinson algorithm for fractionally integrated processes is well-known and widely used. It naturally derives the Cholesky factorization of the inverse matrix of the covariance matrix of the process. We derive the innovation algorithm for the fractionally integrated process. The result is also applied to the derivation of the Cholesky factorization of the covariance matrix of the process in the explicit form. Moreover, the asymptotic theory of the Gaussian maximum likelihood estimator (GMLE) is derived in terms of the innovation algorithm.

E0237: **Sequential Gaussian approximation for nonstationary time series in high dimensions**
Presenter: **Fabian Mies**, RWTH Aachen University, Germany
Co-authors: Ansgar Steland

To enable sequential inference in high-dimensional vector time series, Gaussian couplings of partial sum processes are constructed, for the regime $d = o(n^{\frac{1}{3}})$. The coupling is derived for sums of independent random vectors and subsequently extended to nonstationary time series. The new inequalities depend explicitly on the dimension and on a measure of nonstationarity and are thus also applicable to arrays of random vectors. A feasible bootstrap approximation scheme is proposed. To demonstrate the usefulness of the approximation results, applications to sequential testing and change-point detection are described.

E0649: **A Gaussian approximation result for weakly dependent random fields using dependency graphs**
Presenter: **Dennis Loboda**, RWTH Aachen University, Institute of Statistics, Germany

Non-stationary random fields under the physical dependence measure are investigated. In particular, the objective is to study the maximum of local averages given an increasing bandwidth under expanding-domain asymptotics. By defining suitable vectors based on the studied random field it becomes possible to use the concept of dependency graphs known from time series analysis. This leads to an approximation result for the maximum of local averages through a Gaussian random field which preserves the covariance structure.

E0788: **Distinguishing between breaks in the mean and breaks in persistence under long memory**
Presenter: **Mwasi Mboya**, Leibniz University Hannover, Germany
Co-authors: Simon Wingert, Philipp Sibbertsen

A procedure to discriminate between stationarity, a break in the mean and a break in persistence in a time series that may exhibit long memory is introduced. The asymptotic properties of test statistics based on the CUSUM statistic are studied. In a Monte Carlo study, we further analyze the finite sample properties of the procedure. An application to inflation rates shows the potential of our procedure for future research.

EO197 **Room Virtual R1** **SPATIAL MODELS IN ECONOMIC RESEARCH** **Chair: Pipat Wongsa-art**

E0329: **Estimating a continuous treatment model with spillovers: A control function approach**
Presenter: **Tadao Hoshino**, Waseda University, Japan

The focus is on the estimation of a continuous treatment effect model in the presence of treatment spillovers through social networks. We assume that one's outcome is affected not only by their own treatment but also by the average of their neighbors' treatments, both of which are treated as endogenous variables. Using a control function approach with appropriate instrumental variables, in conjunction with some functional form restrictions, we show that the conditional mean potential outcome can be nonparametrically identified. We also consider a more empirically tractable semiparametric model and develop a three-step estimation procedure for this model. The consistency and asymptotic normality of the proposed estimator are established under certain regularity conditions. As an empirical illustration, we investigate the causal effect of the regional unemployment rate on the crime rate using Japanese city data.

E0417: **Spatial economic models of social interactions**
Presenter: **Pascal Mossay**, Kyungpook National University, Korea, South

An economic approach to social interactions over the space is introduced. We explain how external effects arising from social interactions can lead individuals to cluster in a spatial economy. For this, we exploit the notions of spatial trade-off and economic equilibrium. This allows us to describe how the model parameters affect spatial patterns and to discuss issues related to the the social optimum and multiple equilibria. A particular attention is devoted to the spatial structure underlying the economic models.

E0724: **Varying coefficient model with correlated error components: Disparities between mental health service in England**
Presenter: **Pipat Wongsa-art**, Cardiff University, United Kingdom

The purpose is to discuss estimation procedure and various inferential methods for varying coefficient panel data models that include spatially correlated error components. Our estimation procedure is an extension of the quasi-maximum likelihood method for spatial panel data regression to the conditional local kernel-weighted likelihood. We allow both relevant and irrelevant regressors in our model and propose a variable selection procedure that we show to perform well for models that involve spatial error dependence. We also extend our procedure so that it allows empirical modelling and testing of the so-called semi-varying coefficient specification. To ensure the statistical validity of our methods, we derive a set of asymptotic properties based on a collection of primitive assumptions that appear regularly in the nonparametric literature. Finally, we use the proposed model and methods to analyse the municipal disparities in mental health service spending by local authorities in England in order to illustrate practicability and empirical relevance.

E0808: **The impact of COVID-19 on SMEs default: The role of the network**
Presenter: **Francesco Moscone**, Brunel University London and Ca Foscari of Venice, United Kingdom
Co-authors: Monica Billio, Joan Madia, Elisa Tosetti

While various policy measures enhancing firms' access to liquidity mitigated immediate insolvency risks, a large number of companies failed. Financial variables measuring the liquidity and level of indebtedness of companies as well as macroeconomic factors are traditionally included as determinants of firms' default. Further, a large literature has pointed to the importance of interfirm links in determining a company's performance. An important type of network is the web of interlocking directorates of companies, having profound effects on companies regarding a variety of decisions, strategies and structures, and may prove to be critical during periods of high uncertainty, such as the one induced by the COVID-19 pandemic. We estimate a spatial Probit model for firms default as a function of firms characteristics, its pre-crisis financial situation, macroeconomic variables and spatial effects, where board membership is used to construct the network. We exploit a novel large data set on about 2,343,103 companies based in the United Kingdom followed over the years 2016 and 2022. Results suggest that the network plays an important role in determining firm performance and well-connected firms are more resilient during the pandemic relative to those that are isolated.

EO297 Room Virtual R2 THEORIES AND METHODOLOGIES FOR STOCHASTIC PROCESSES Chair: Teppei Ogihara

E0165: **Existence in the inverse Shiryaev problem**
Presenter: **Yoann Potiron**, Keio University, Japan

The inverse first-passage Shiryaev problem is considered, i.e. for a standard Brownian motion $(W_t)_{t\geq 0}$, and any upper boundary continuous function $g:\mathbb{R}^+\to\mathbb{R}$ satisfying $g(0)\geq 0$, we define $\tau_g^W := \inf\{t\in\mathbb{R}^+ \text{ s.t. } W_t\geq g(t)\}$, and $f_g^W(t)$ its related density. For any target density function of the form $f:\mathbb{R}^+\to\mathbb{R}^+$ satisfying some smooth assumptions and any arbitrarily big horizon time $T>0$, we show the existence of a related boundary $g_{f,T}:\mathbb{R}^+\to[0,T]$, with $g_{f,T}(0)\geq 0$, which satisfies $f_{g_{f,T}}^W(t)=f(t)$ for $0\leq t\leq T$. As an example, the exponential distribution $f(t)=\lambda\exp(-\lambda t)$ for $\lambda>0$ satisfies the assumptions. As $g_{f,T}$ is exhibited as a limit boundary of a subsequence of a piecewise linear boundary, we do not obtain any explicit formula for $g_{f,T}$ as a function of f, nor the unicity of the solution. The results are also proved in the symmetrical two-dimensional boundary case.

E0230: **Local asymptotic normality for jump-diffusion processes**
Presenter: **Teppei Ogihara**, University of Tokyo, Japan
Co-authors: Yuma Uehara

When we try to show local asymptotic normality (LAN) of jump-diffusion processes with discrete observations, there are two problems. The first one is to control transition density ratios between two different values of the parameter. To solve this, we use the scheme with the so-called L^2 regularity condition. The original scheme cannot be applied for jump-diffusion processes because of their fat-tailed behaviors. Therefore, we extend the scheme so that it can be applied to jump-diffusion processes. The second problem is that the transition probability for no jump is quite different from that for the presence of jumps. This fact makes it difficult to identify the asymptotic behavior of the likelihood function. To deal with this problem, we approximate the original likelihood function by using a thresholding likelihood function that detects the existence of jumps. As a consequence of these techniques, we obtain LAN for jump-diffusion processes. Moreover, the quasi-maximum-likelihood and Bayes-type estimators are shown to be asymptotically efficient in this model.

E0457: **Model comparison for ergodic Levy driven SDEs in YUIMA**
Presenter: **Shoichi Eguchi**, Osaka Institute of Technology, Japan

There are several studies of model selection for stochastic differential equations (SDEs), which include the contrast-based information criterion for ergodic diffusion processes and the Schwarz type information criterion for ergodic SDEs. Based on these studies, in the R package yuima, the function for model selection for diffusion processes has been implemented. However, this function is not compatible with the model selection for Levy driven SDEs. We will overview the model selection methods for Levy driven SDEs and explain the improvements of the model selection function.

E0739: **Benign overfitting in stochastic regression**
Presenter: **Shogo Nakakita**, The University of Tokyo, Japan
Co-authors: Masaaki Imaizumi

The excess risks of overparameterized stochastic regression, that is, linear regression with covariates being stochastic processes and whose number is greater than that of samples, are considered. Recent studies show that overparameterized linear regression with i.i.d. samples can predict well even if they have fewer samples than parameters and no sparsity. We examine how time series structure can affect the performance of overparameterized regression without sparsity. One of the results is that even if the covariates have long-range dependence, the sufficiently fast decay of eigenvalues of the covariance operator can make the excess risk converge to zero.

EO303 Room Virtual R3 BAYESIAN ECONOMETRICS FOR EVIDENCE-BASED POLICY MAKING Chair: Thomas Zoerner

E0646: **Hawks vs. Doves: Monetary policy effectiveness in light of diverging national policy stances**
Presenter: **Thomas Zoerner**, Vienna University of Economics and Business, Austria
Co-authors: Florian Huber, Niko Hauzenberger

The secular increase in globalization led to substantial increases in the connectedness between global financial markets. This has important implications for the conduct of monetary policy, since central bank policies might diverge across countries, hampering key transmission channels of domestic policy actions. We develop a non-linear multivariate time series model to shed light on how US monetary policy affects the conduct of monetary policy in the Euro area (EA). Based on a smooth transition model, we assume that dynamic coefficients implicitly depend on proxies of unexpected US monetary policy shocks. This assumption allows us to investigate how dynamic responses of high-frequency quantities such as government bond yields and inflation swaps to EA monetary policy shocks change if the Federal Reserve unexpectedly changes its monetary policy stance. Carrying out scenario-specific impulse responses shows that EA monetary policy transmission strongly depends on the monetary policy stance of the Federal Reserve and has sizeable effects on a variety of euro area variables.

E0705: **Uncertain spillover effects and priors for spatial models**
Presenter: **Nikolas Kuschnig**, Vienna University of Economics and Business, Austria

In an ever more connected world spillover effects are at the centre of a wide range of applied research. Spatial econometric models are commonly used to analyse such spillovers empirically. However, these models suffer from rigid specifications and strong assumptions regarding connectivity between units. We address these issues by adopting a fully Bayesian approach. Assumptions such as known connectivities can be loosened, with their forms being learned from the data instead. Weakly informative priors provide the foundation for a flexible framework, imposing regularisation where appropriate and limiting assumptions otherwise. The result is more credible and extensible empirical tools that natively account for uncertainty. We dismantle the spatial econometric framework, discuss prior information in the context, and sketch out a Bayesian approach. We introduce general purpose and specific probability models that allow flexible treatment of connectivity. A simulation exercise and an empirical application show the merits of this approach. Bayesian methods present a great opportunity for once again raising the bar in spatial econometrics.

E0783: **Accounting for model uncertainty in Bayesian Poisson regression models**
Presenter: **Gregor Zens**, Bocconi University, Italy
Co-authors: Mark Steel

Variable selection in Poisson regression models is a standard task for applied researchers in various fields. While frequentist penalized likelihood methods are well established, Bayesian frameworks have received considerably less attention. We develop a novel, exact and computationally feasible hierarchical framework for model averaging and variable selection in Poisson regression models. Posterior simulation is based on automatic and efficient reversible jump Markov chain Monte Carlo algorithms. A simulation study demonstrates the strengths of the framework relative to a number of competitor models, and real data applications further illustrate the approach.

E0785: **Forecasting sectoral greenhouse gas emissions for a global sample**
Presenter: **Lukas Vashold**, Vienna University of Economics and Business, Austria
Co-authors: Jesus Crespo Cuaresma

Effectively tackling climate change requires sound knowledge about its driving forces, namely greenhouse gases (GHGs) emissions, and their sources. We use a hierarchical Bayesian multivariate time series model for Gross Domestic Product (GDP) per capita, population and sectoral emissions intensity across countries, in the spirit of a country- and sector-specific Kaya's identity. Conditioning on established projections in line with the Shared Socioeconomic Pathways (SSPs), we derive predictions for sectoral GHG emissions for the period until 2050. Results show that yearly GHG emissions are increasing strongly in the next three decades. Increases are mainly driven by emerging and developing economies where the reduction of emissions intensity is not fast enough to outweigh increases in affluence and population. Emissions related to the energy sector account for most of the growth, followed by industrial sources. The majority of already advanced economies show reductions in overall GHG emissions as well as in most sectors. However, we also document that emissions in the transport sector are still to rise in these countries given rather slow technological progress and uptake thereof. Our predictions show that limiting global warming to levels agreed to under the Paris agreement is unlikely in the absence of further efforts to decarbonise.

EO305 Room Virtual R4 CLIMATE ECONOMETRICS Chair: Marina Friedrich

E0262: **A state space representation of a two-component energy balance model**
Presenter: **Jingying Zhou Lykke**, Aarhus University, Denmark
Co-authors: Mikkel Bennedsen, Eric Hillebrand

A state-space representation (EBM-SS model) of the two-component energy balance model (EBM) is proposed. The EBM-SS model incorporates three extensions to the two-component EBM. First, we include ocean heat content (OHC) as a measurement of the temperature in the deep ocean layer. Second, we decompose the latent state of radiative forcing into a natural component and an anthropogenic component. Lastly, we use multiple GMST (Global Mean Surface Temperature) anomaly data sources from separate research groups as measurements for the latent states in the two-component EBM. We estimate the EBM-SS model using observations at the global level during 1955 - 2020 by maximum likelihood. We show in empirical estimation and in simulations that using multiple data sources for the latent process reduces parameter estimation uncertainty. When fitting eight observational GMST anomaly series, the physical parameter estimates are comparable to those obtained by using datasets from the Coupled Model Intercomparison Project 5 (CMIP 5) in other literature. We find that using this set of estimates, the GMST projection results under four Representative Concentration Pathway (RCP) scenarios considerably agree with the outputs from the climate emulator Model for the Assessment of Greenhouse Gas Induced Climate Change (MAGICC) 7.5 and CMIP 5 models. We show that utilizing a simple climate model and historical records alone can produce meaningful GMST projections.

E0356: **The dependence between income inequality and carbon emissions: A distributional copula analysis**
Presenter: **Franziska Dorn**, University of Goettingen, Germany
Co-authors: Thomas Kneib, Simone Maxand

High levels of carbon emissions and rising income inequality are interconnected challenges for the global society. Commonly-applied linear regression models fail to unravel the complexity of bidirectional transmission channels. In particular, consumption, energy sources, the structure of the economy and the political system are determinants of the strength and direction of the dependence between emissions and inequality. To capture their impact, the conditional dependence between income inequality and emissions is investigated by applying distributional copula models on an unbalanced panel data set of 154 countries from 1960 to 2019. A comparison of high-, middle- and low-income countries contradicts a linear relationship and sheds light on heterogeneous dependence structures implying synergies, trade-offs and decoupling between income inequality and carbon emissions. Based on the conditional distribution, we can identify determinants associated with higher/lower probabilities of a country falling in an area of potential social and environmental sustainability. The results imply that the joint activation of multiple channels opens the way for a sustainable future.

E0381: **Sieve Bootstrap inference for time-varying coefficient models**
Presenter: **Marina Friedrich**, VU Amsterdam, Netherlands
Co-authors: Yicong Lin

A sieve bootstrap framework is proposed to conduct pointwise and simultaneous inference for time-varying coefficient regression models based on a nonparametric local linear estimator. The asymptotic validity of the sieve bootstrap in the presence of autocorrelation is established. We find that it automatically produces a consistent estimation of nuisance parameters, both at the interior and boundary points. In addition, we develop a bootstrap test for parameter constancy and show that it is asymptotically correctly sized. An extensive simulation study supports our findings. The proposed methods are applied to assess the price development of CO2 certificates in the European Emissions Trading System (EU ETS). We find evidence of time variation in the relationship between allowance prices and their fundamental price drivers.

E0562: **Demand or supply: An empirical exploration of the effects of climate change on the macroeconomy**
Presenter: **Fulvia Marotta**, Queen Mary University of London, United Kingdom

Using an original panel data set for 24 OECD countries over the sample 1990-2019 and a standard empirical macroeconomic framework for business cycle analysis, the paper tests the combined macroeconomic effects of climate change, environmental policies and technology. Overall, we find evidence of significant macroeconomic effects over the business cycle: physical risks act as negative demand shocks while transition risks as downward supply movements. The disruptive effects on the economy are exacerbated for countries without carbon tax or with high exposure to natural disasters. In general, results support the need for a uniform policy mix to counteract climate change with a balance between demand-pull and technology-push policies.

EO089 Room Virtual R6 RECENT ADVANCES IN LARGE PANEL DATA MODELLING Chair: Bin Peng

E0334: **Binary response models for heterogeneous panel data with interactive fixed effects**
Presenter: **Bin Peng**, Monash University, Australia

Binary response models are investigated for heterogeneous panel data with interactive fixed effects by allowing both the cross-sectional dimension and the temporal dimension to diverge. From a practical point of view, the proposed framework can be applied to predict the probability of corporate failure, conduct credit rating analysis, etc. Theoretically and methodologically, we build a link between a maximum likelihood estimation and a least-squares approach, provide a simple information criterion to detect the number of factors, and establish the corresponding asymptotic theory. In addition, we conduct intensive simulations to examine the theoretical findings. In an empirical study, we focus on the sign prediction of stock returns, and then use the results of the sign forecast to conduct portfolio analysis.

E0336: **Functional-coefficient quantile regression for panel data with latent group structure**
Presenter: **Degui Li**, University of York, United Kingdom

The focus is on estimating functional-coefficient models in panel quantile regression with individual effects, allowing the cross-sectional and temporal dependence for large panel observations. A latent group structure is imposed on the heterogenous quantile regression models so that the number of nonparametric functional coefficients to be estimated can be reduced considerably. With the preliminary local linear quantile estimates of the subject-specific functional coefficients, a classic agglomerative clustering algorithm is used to estimate the unknown group structure and an easy-to-implement ratio criterion is proposed to determine the group number. The estimated group number and structure are shown to be consistent. Furthermore, a post- grouping local linear smoothing method is introduced to estimate the group-specific functional coefficients, and the relevant asymptotic normal distribution theory is derived with a normalisation rate comparable to that in the literature. The developed methodologies and theory are verified through a simulation study and showcased with an application to house price data from UK local authority districts, which reveals different homogeneity structures at different quantile levels.

E0470: **Estimating quantile-dependent networks on panel data**
Presenter: **Yutao Sun**, Dongbei University of Finance and Economics, China
Co-authors: Wendun Wang

Methods are proposed for the estimation of an unknown network (in particular, the corresponding adjacency matrix) from a panel data set in which the individuals are connected through the network. We consider two scenarios: a quantile-dependent network and a quantile-invariant network. A quantile-dependent network involves links that mutate across data quantiles. In such a case, our approach involves a nonlinear quantile regression model where the entries of the adjacency matrix are treated as model parameters. A quantile-invariant network possesses links that are constant and do not change over data quantiles. When the network is quantile-invariant, we consider a composite quantile estimation approach that estimates the entries of the adjacency matrix on multiple data quantile levels. Such an approach exploits the information at several quantile levels jointly and is efficient. We further impose a sparsity assumption on the network and invoke standard regularization techniques to improve the estimation efficiency. Our estimation procedures are computationally feasible in that we establish a derivative-based nonlinear programming algorithm for the underlying optimization problem. Simulation studies are conducted to investigate the performance of our methods.

E0520: **Factor-augmented nonstationary panels with multiple structural changes**
Presenter: **Qu Feng**, Nanyang Technological University, Singapore, Singapore

Nonstationary panels are widely used in empirical studies in economics and other related fields. Multiple structural changes are considered in nonstationary heterogeneous panels with common factors. Breaks could occur in slopes and error factor loadings. Unobserved error factors are treated as additional regressors. Thus, different breakpoints in slopes and error factor loadings are treated as multiple breaks in linear regression models. As previously, unobserved error factors can be proxied by cross-sectional averages of observable data. We show that the breakpoints in both slopes and error factor loadings can be consistently estimated by least squares in both cases of i) nonstationary factors and ii) nonstationary regressors. Monte Carlo simulations are conducted to verify the main results in finite samples.

EO171 Room Virtual R7 REGRESSION AND BIAS REDUCTION IN EXTREME VALUE THEORY Chair: Antoine Usseglio-Carleve

E0554: **Automatic threshold selection for extreme value regression models**
Presenter: **Julien Hambuckers**, University of Liege, Belgium
Co-authors: Marie Kratz, Antoine Usseglio-Carleve

The problem of threshold selection is investigated in the context of the extreme value regression model. In this regression context, the threshold choice is a non-trivial task since it should also depend on the covariates and can have important consequences on the final estimates. We propose an efficient and robust solution to automatically estimate these thresholds with the help of the distributional regression machinery. We illustrate its properties through several simulation studies. The method is later applied to the estimation of hedge funds tail risks.

E0648: **A refined Weissman estimator for extreme quantiles**
Presenter: **Jonathan El Methni**, Universite Paris Cite, France
Co-authors: Stephane Girard, Michael Allouche

Weissman's extrapolation methodology for estimating extreme quantiles from heavy-tailed distributions is based on two estimators: an order statistic to estimate an intermediate quantile and an estimator of the tail-index. The common practice is to select the same intermediate sequence for both estimators. We show how an adapted choice of two different intermediate sequences leads to a reduction of the asymptotic bias associated with the resulting refined Weissman estimator. The asymptotic normality of the latter estimator is established and a data-driven method is introduced for the practical selection of the intermediate sequences. Our approach is compared to the Weissman estimator and to six bias reduced estimators of extreme quantiles in a large scale simulation study. It appears that the refined Weissman estimator outperforms its competitors in a wide variety of situations, especially in challenging high bias cases. Finally, an illustration of an actuarial real data set is provided.

E0859: **Extreme partial least-squares**
Presenter: **Meryem Bousebata**, Inria, France
Co-authors: Stephane Girard, Geoffroy Enjolras

A new approach, called Extreme-PLS, is proposed for dimension reduction in conditional extreme values settings. The objective is to find linear combinations of covariates that best explain the extreme values of the response variable in a non-linear inverse regression model. The asymptotic normality of the Extreme-PLS estimator is established in the single-index framework and under mild assumptions. The performance of the method is assessed on simulated data. A statistical analysis of French farm income data, considering extreme cereal yields, is provided as an illustration.

E0893: **Extremal expectile regression**

Presenter: **Yasser Abbas**, Fondation Jean-Jacques Laffont, France

Co-authors: Abdelaati Daouia, Gilles Stupfler

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

EO239 **Room Virtual R8** RECENT ADVANCES IN BIG AND COMPLEX DATA ANALYSIS **Chair: Xiaojun Mao**

E0569: **Structure learning via unstructured kernel-based M-regression**

Presenter: **Xin He**, Shanghai University of Finance and Economics, China

In statistical learning, identifying underlying structures of true target functions based on observed data plays a crucial role to facilitate subsequent modeling and analysis. Unlike most of those existing methods that focus on some specific settings under certain model assumptions, a general and novel framework for recovering true structures of target functions is proposed by using unstructured M-regression in a reproducing kernel Hilbert space (RKHS). The proposed framework is inspired by the fact that gradient functions can be employed as a valid tool to learn underlying structures, including sparse learning, interaction selection and model identification, and it is easy to implement by taking advantage of the nice properties of the RKHS. More importantly, it admits a wide range of loss functions, and thus includes many commonly used methods, such as mean regression, quantile regression, likelihood-based classification, and margin-based classification, which is also computationally efficient for solving convex optimization tasks. The asymptotic results of the proposed framework are established within a rich family of loss functions without any explicit model specifications. The superior performance of the proposed framework is also demonstrated by a variety of simulated examples and a real case study.

E0655: **Large-scale importance selection of heteroscedastic units**

Presenter: **Bowen Gang**, Fudan University, China

Choosing candidates to whom a limited set of resources will be distributed is a pervasive dilemma. In multiple testing procedures that can be used to choose such candidates, power is traditionally defined as the number or proportion of correctly selected non-null hypotheses. We propose a generalized power that allows researchers to better select more desirable testing units and propose a specific formulation to capture not only if a unit has been correctly categorized as null or alternative but also to better reward the detection of larger effect sizes. Our new empirical Bayes multiple testing framework rewards discovering not just significant but large effects while controlling type I error. Hence, the selection process is better able to incorporate effect size into selection. We provide theoretical guarantees for FDR control and power optimization as well as numeric evidence for the utility of a generalized power.

E0693: **A two-stage model for high-risk prediction in insurance ratemaking**

Presenter: **Yanxi Hou**, Fudan University, China

In actuarial practice, modern statistical methodologies are one primary consideration for real actuarial problems, such as premium calculation, insurance preservation, marginal risk analysis, etc. The claim data usually possesses a complex data structure, so direct applications of statistical techniques will result in unstable predictions. For example, insurance losses are semicontinuous variables, where a positive mass on zero is often associated with an otherwise positive continuous outcome. Thus, the prediction of high-risk events of claim data needs additional treatment to avoid significant underestimation. We propose a new two-stage composite quantile regression model for the prediction of the value-at-risks of the aggregate insurance losses. As we are interested in the statistical properties of our method, the asymptotic results are established corresponding to different types of risk levels. Finally, some simulation studies and data analysis are implemented for the illustration of our method.

E0777: **Functional calibration under non-probability survey sampling**

Presenter: **Zhonglei Wang**, Xiamen University, China

Co-authors: Xiaojun Mao, Jae Kwang Kim

Non-probability sampling is prevailing in survey sampling, but ignoring its selection bias leads to erroneous inferences. Incorporating auxiliary information from an independent probability sample, we propose a unified nonparametric method to estimate the sampling weights for a non-probability sample by calibrating functions of auxiliaries in a reproducing kernel Hilbert space. The consistency and the limiting distribution of the proposed estimator are established under rejective sampling, and the corresponding variance estimator is also investigated. Compared with existing works, the proposed method is more robust since no parametric assumption is needed for the selection mechanism of the non-probability sample. Numerical results demonstrate that the proposed method outperforms its competitors, especially when the model is misspecified. The proposed method is applied to analyze the average total cholesterol of Korean citizens based on a non-probability sample from the National Health Insurance Sharing Service and a probability sample from the Korea National Health and Nutrition Examination Survey.

EC434 **Room 107 (Hybrid 7)** CONTRIBUTIONS IN APPLIED STATISTICS AND ECONOMETRICS **Chair: Makoto Takahashi**

E0863: **The impact of professional development on teacher job satisfaction: Evidence from a multilevel model**

Presenter: **Mike Smet**, KU Leuven, Belgium

Numerous studies in different countries find evidence for high rates of teacher turnover, leading to shortages and potential quality issues. Job satisfaction is found to be an important antecedent of turnover. We investigate the impact of various aspects of professional development for teachers (as well as interactions of these aspects) on their job satisfaction, using data from the 2018 wave of the Teaching and Learning International Survey (TALIS), which was conducted by OECD. Our empirical analysis consists of 49378 teachers, nested in 3128 schools, nested in 20 regions. The hierarchical structure of the data requires the use of an appropriate estimation technique: multilevel or hierarchical linear modelling (HLM) with three levels: teachers, schools and regions. We find a significant positive relationship between job satisfaction and the need for professional development for teaching diversity and special needs, which is (negatively) moderated by the number of professional development activities a teacher had participated in in the 12 months prior to the study. Another indicator, measuring the need for professional development in subject matter and pedagogy shows a significant negative relationship with job satisfaction and is (positively) moderated by the number of professional development activities.

E0549: **Long- and short-term impacts of COVID-19 on price volatility using MIDAS in Palestine**

Presenter: **Samir Safi**, United Arab Emirates University, United Arab Emirates

The aim is to study the impact of COVID-19 on price volatility in Palestine, and to forecast the extent of price volatility. We will show the challenges that are facing Palestine, diagnose the effects of COVID-19 on prices, and provide some remedy procedures to tackle the problems resulting from the continuous spread of novel COVID-19 in Palestine. This impact will be examined by using mixed data sampling, the MIDAS regression model. In the methodology of this technique, the selected best model will be tested through the validity of the estimated model based on information criteria. Issues in different forecasting accuracy of the MIDAS regression model approaches and criteria to choose the best model will be discussed in this research. This approach of using MIDAS will tackle the time series data with low and high frequencies. The MIDAS methodology addresses the situation where the response variable in the regression is sampled at a lower frequency than one or more of the regressors. The proposed research will be the first in this area to address long-run and short-run price related issues associated with COVID-19. Previous research has faced bottlenecks when combining data with different frequencies in this area.

E0687: **Component decomposition of the Nikkei stock average time series using moving linear model approach**
Presenter: **Koki Kyo**, Niigata University of Management, Japan

A moving linear model (MLM) approach is developed to decompose a time series into trend and fluctuation components. The features of the proposed approach are as follows: It is only necessary to set a local linear model for the trend component and is not necessary to introduce any model for the fluctuation component in which the trend and fluctuation components are uncorrelated. We can decompose a time series into several components by using the MLM approach so we can analyze each component using a simple model. As an illustration, we apply the MLM approach to analyze the daily time series of the Nikkei stock average in Japan.

E0856: **Value of non-traditional data sources for official statistics perspective from developing economy**
Presenter: **Syeda Rabab Mudakkar**, Lahore School of Economics, Pakistan

Web scraped data has recently been incorporated into the inflation estimates by organizations such as the Statistics Bureau of Japan, Australian Bureau of Statistics and Statistics New Zealand, among others, using the average price of a commodity over a specific period of time. However, still, the cases are few. Due to a lack of technology infrastructure, a restrictive legislative environment, data access and quality issues, and other factors, the majority of national statistical offices are wary of “big data” exploration. This is a first attempt to examine the challenges and opportunities of web scraped price data for a developing economy. The daily information of one hundred thousand durable and non-durable products for a time period of one year is collected. The results reveal that due to the low penetration of online retail, the market-clearing phenomenon works less efficient compared to brick and mortar retail. Further, the data exhibits the presence of features such as a higher increase in price size relative to reduction, skewness, kurtosis etc. Overall sticky-price behavior is observed but few “random-signals” light up in situations when the size of price increases is far from zero. This indicates the presence of Calvo price-setting behavior in internet retail in developing economies.

Monday 06.06.2022 07:40 - 09:20 Parallel Session K – EcoSta2022

EO315 Room Virtual R1 RECENT DEVELOPMENTS ON MICROBIOME DATA ANALYSIS Chair: Wodan Ling

E0171: **Resampling-based inferences for compositional regression when sample sizes are limited**
Presenter: **Sungkyu Jung**, Seoul National University, Korea, South
Co-authors: Sujin Lee, Jeongyoun Ahn

Gut microbiomes are increasingly found to be associated with many health-related characteristics of humans as well as animals. A regression with compositional microbiomes covariates is commonly used to identify important bacterial taxa that are related to various phenotype responses. Often the dimension of microbiome taxa easily exceeds the number of available samples, which creates a serious challenge in the estimation and inference of the model. We propose a new estimation and inference procedure for linear regression models with extremely low-sample sized compositional predictors. Under the compositional log-contrast regression framework, the proposed approach consists of two steps. The first step is to screen relevant predictors by fitting a log-contrast model with a sparse penalty. The screened-in variables are used as predictors in the non-sparse log-contrast model in the second step, where each of the regression coefficients is tested using nonparametric, resampling-based methods such as permutation and bootstrap. The performances of the proposed methods are evaluated by a simulation study, which shows they outperform traditional approaches based on normal assumptions or large sample asymptotics. Application to steer microbiomes data successfully identifies key bacterial taxa that are related to important cattle quality measures.

E0514: **Multiscale analysis of count data through topic alignment**
Presenter: **Julia Fukuyama**, Indiana University, United States
Co-authors: Laura Symul, Kris Sankaran

Topic modeling is a popular method used to describe biological count data. With topic models, the user must specify the number of topics K. Since there is no definitive way to choose K and since a true value might not exist, we develop techniques to study the relationships across models with different K. This can show how many topics are consistently present across different models, if a topic is only transiently present, or if a topic splits in two when K increases. This strategy gives more insight into the process of generating the data than choosing a single value of K would. We design a visual representation of these cross-model relationships, which we call a topic alignment, and present three diagnostics based on it. We show the effectiveness of these tools for interpreting the topics on simulated and real data, and we release an accompanying R package, alto.

E0720: **A Gaussian mixture model to integrate metagenome and metatranscriptome data**
Presenter: **Di Wu**, University of North Carolina at Chapel Hill, United States

Bacterial dysbiosis has been implicated in various clinical conditions, e.g., caries, gut diseases and cancer. Microbial composition and abundance can be captured by microbiome DNA sequencing for the question of what bacteria are there?, while metatranscriptomics through RNA sequencing identifies functional characterization of complex microbial communities to answer what the bacteria do there. The joint analyses of paired metagenomics and metatranscriptomics data are one way to study bacterial species and genes' functional roles in statuses of health and diseases but remain challenging due to the high dimension and sparsity of the data. To address this knowledge gap, we will study the differential transcriptional activity by investigating the RNA/DNA ratios at species. We propose a two-step differential expression analysis approach that includes testing at each of the two modalities and fitting the log-RNA/DNA ratio to a novel Gaussian Mixture statistical model. Our proposed method IntegRatio is flexible to control batch effects and accommodate multiple study covariates. It also comprehensively tests more microbiome-specific hypotheses simultaneously than the conventional method. Real data-inspired simulations show the controlled type I error and decent power. The proposed method has been applied in studying Early Childhood Caries (ECC) and Inflammatory Bowel Diseases (IBD) to identify species that have differential regulatory activities associated with diseases.

E0811: **Scalable and interpretable rare feature aggregation with microbiome data**
Presenter: **Kun Chen**, University of Connecticut, United States

Statistical learning with a large number of rare features is commonly encountered in modern applications, such as in analyzing the gun-brain axis with high-throughput microbiome features that are often given as compositions or presence/absence indicators. Properly balancing the features' rarity and specificity holds the key to harvesting valuable information from them. Fortunately, an inherited hierarchical tree structure often exists among the features, making it possible to perform interpretable feature aggregation. Two statistical learning approaches are introduced for rare feature aggregation and selection conforming to any given tree structure, one for compositional features and another for binary features. For compositional features, we propose Relative-Shift Regression, in which the compositions are aggregated based on whether shifting relative concentrations between them affects the outcome. For binary features, we propose Convex Logic Regression, in which feature reduction is achieved through both a sparsity pursuit and an aggregation promoter with the logic operator of "or". Equi-sparse convex regularization methods and efficient smoothing proximal gradient algorithms are developed with theoretical guarantees. Applications with microbiome data from a preterm infant study are discussed.

EO311 Room Virtual R10 MODERN STATISTICAL METHODS FOR LONGITUDINAL AND SURVIVAL DATA Chair: Esra Kurum

E0526: **Challenges of modeling longitudinal intensive care unit data**
Presenter: **Joel Dubin**, University of Waterloo, Canada

Prediction of health outcomes is an important component for determining how to make recommendations and treat individuals. Regarding treatment, the intensive care unit (ICU) is a place where many such decisions are made. A primary goal for ICU patients is treating them to achieve positive outcomes (e.g., hospital discharge alive, improvement from in-hospital ailments, extended survival). A major analytical issue is the preponderance of information available at ICU entry (e.g., age, sex, co-morbidities, prescriptions, vital signs), and especially longitudinally (e.g., vital sign changes, dynamic renal function, in-ICU treatment). We will present some interesting analytic challenges utilizing longitudinal data for predictive modeling from a large ICU database, and discuss a few remedies that we have investigated.

E0506: **Efficiency loss with binary pre-processing of continuous monitoring data**
Presenter: **Elizabeth Juarez-Colunga Juarez-Colunga**, University of Colorado Anschutz Medical Campus, United States
Co-authors: Paula Langner

In studies with a repeatedly measured recurrent event outcome, events may be captured as counts during subsequent intervals or follow-up times either by design or for ease of analysis. In many cases, recurrent events may be further coarsened such that only an indicator of one or more events in an interval is observed at the follow-up time, resulting in a loss of information relative to a record of all events. We examine efficiency loss when coarsening longitudinally observed counts to binary indicators and aspects of the design which impact the ability to estimate a treatment effect of interest. The investigation is motivated by a study of patients with Cardiac Implantable Electronic Devices (CIEDs) in which investigators aimed to examine the effect of treatment on events detected by the devices over time. In order to study components of such a recurrent event process impacted by data coarsening, we derive the asymptotic relative efficiency (ARE) of a treatment effect estimator utilizing a count outcome, which represents a longitudinal recurrent event process, relative to a coarsened binary outcome. We compare the efficiencies and consider conditions where the binary process maintains good efficiency in estimating a treatment effect.

E0546: **Optimal cut-points for screening for pre-clinical disease based on various criteria**
Presenter: **Cuiling Wang**, Albert Einstein College of Medicine, United States

The selection of a cut-point for a marker of disease to identify those at high risk of developing the disease in the future is crucial in clinical practice and research. Although optimal thresholds based on various criteria can be obtained through time-dependent receiver operating characteristic (ROC) analysis, the properties of the optimal cut-points are not well known. Recently, the properties and estimation methods for the time-dependent optimal cuts based on the Youdens index have been studied. However, optimal cuts based on other criteria are rarely studied. We investigate the properties of the time-dependent optimal cut-points based on various criteria including setting the target level for sensitivity or specificity alone, the weighted sum of sensitivity and specificity which includes Youden's index as a special case, and the average overall cost. For the weighted sum and average overall cost criteria, we provide formulae to estimate the optimal cuts using survival models. The methods are applied to screening for pre-clinical Alzheimer's dementia using a well-established memory test in the Einstein Aging Study. Simulation studies are performed to evaluate the direct estimates as well as those obtained from time-dependent ROC analysis.

E0532: **A Bayesian multilevel time-varying framework for joint modeling of hospitalization and survival in patients on dialysis**
Presenter: **Esra Kurum**, University of California, Riverside, United States

Over 782,000 individuals in the U.S. have end-stage kidney disease with about 72% of patients on dialysis, a life-sustaining treatment. Dialysis patients experience high mortality and frequent hospitalizations, about twice per year. These poor outcomes are exacerbated at key time periods, such as the fragile period after the transition to dialysis. In order to study the time-varying effects of modifiable patient and dialysis facility risk factors on hospitalization and mortality, we propose a novel Bayesian multilevel time-varying joint model. Efficient estimation and inference are achieved within the Bayesian framework using Markov Chain Monte Carlo, where multilevel (patient- and dialysis facility-level) varying coefficient functions are targeted via Bayesian P-splines. Applications to the United States Renal Data System, a national database that contains data on nearly all patients on dialysis in the U.S., highlight significant time-varying effects of patient- and facility-level risk factors on hospitalization risk and mortality. The finite sample performance of the proposed methodology is studied through simulations.

EO121 Room Virtual R11 RECENT ADVANCES IN MODELS WITH COMPLEX DEPENDENCE Chair: Cheng Li

E0449: **Classification trees for imbalanced data: Surface-to-volume regularization**
Presenter: **Yichen Zhu**, Duke University, United States
Co-authors: Cheng Li, David Dunson

Classification algorithms face difficulties when one or more classes have limited training data. We are particularly interested in classification trees, due to their interpretability and flexibility. When data are limited in one or more of the classes, the estimated decision boundaries are often irregularly shaped due to the limited sample size, leading to poor generalization error. We propose a novel approach that penalizes the Surface-to-Volume Ratio (SVR) of the decision set, obtaining a new class of SVR-Tree algorithms. We develop a simple and computationally efficient implementation while proving estimation consistency for SVR-Tree and rate of convergence for an idealized empirical risk minimizer of SVR-Tree. SVR-Tree is compared with multiple algorithms that are designed to deal with imbalance through real data applications.

E0575: **Consistent Bayesian community detection for assortative networks**
Presenter: **Sheng Jiang**, University of California Santa Cruz, United States
Co-authors: Surya Tokdar

Stochastic Block Models (SBMs) are a fundamental tool for community detection in network analysis. But little theoretical work exists on the statistical performance of Bayesian SBMs, especially when the community count is unknown. This paper studies weakly assortative SBMs whose members of the same community are more likely to connect with one another than with members from other communities. The weak assortativity constraint is embedded within an otherwise weak prior, and under mild regularity conditions, the resulting posterior distribution is shown to concentrate on the true community count and membership allocation as the network size grows to infinity. An efficient Gibbs sampler is developed to sample from the posterior distribution. Finite sample properties are examined via simulation studies in which the proposed method offers competitive estimation accuracy relative to existing methods under a variety of challenging scenarios.

E0732: **Probabilistic contrastive principal component analysis**
Presenter: **Didong Li**, Princeton University; University of California, Los Angeles, United States
Co-authors: Andy Jones, Barbara Engelhardt

Dimension reduction is useful for exploratory data analysis. In many applications, it is of interest to discover a variation that is enriched in a "foreground" dataset relative to a "background" dataset. Recently, contrastive principal component analysis (CPCA) was proposed for this setting. However, the lack of a formal probabilistic model makes it difficult to reason about CPCA and tune its hyperparameter. We propose probabilistic contrastive principal component analysis (PCPCA), a model-based alternative to CPCA. We discuss how to set the hyperparameter in theory and in practice, and we show several of PCPCA's advantages over CPCA, including greater interpretability, uncertainty quantification and principled inference, robustness to noise and missing data, and the ability to generate data from the model. We demonstrate PCPCA's performance through a series of simulations and case-control experiments with datasets of gene expression, protein expression, and images.

E0749: **Intrinsic and extrinsic deep learning on manifolds**
Presenter: **Lizhen Lin**, The University of Notre Dame, United States

Both intrinsic and extrinsic deep neural network (DNN) models on the manifolds are discussed. An intrinsic DNN employs a Riemannian structure of the manifold while an extrinsic DNN relies on embedding a manifold onto a high-dimensional Euclidean space. The excessive risk of the DNN estimators will be derived and extensive numerical studies have been carried out to demonstrate the utilities of the models and illustrate the role of the geometry in developing the DNN models.

EO163 Room Virtual R12 STATISTICAL MACHINE LEARNING WITH NETWORKS, MATRICES, AND MANIFOLDS Chair: Joshua Cape

E0202: **Limit theorems for out-of-sample extensions of spectral graph embeddings**
Presenter: **Keith Levin**, University of Wisconsin, United States

Graph embeddings, a class of dimensionality reduction techniques designed for relational data, have proven useful in exploring and modeling network structure. Most dimensionality reduction methods allow out-of-sample extensions, by which an embedding can be applied to observations not present in the training set. Applied to graphs, the out-of-sample extension problem concerns how to compute the embedding of a vertex that is added to the graph after an embedding has already been computed. We consider the out-of-sample extension problem for two graph embedding procedures: the adjacency spectral embedding and the Laplacian spectral embedding. In both cases, we prove that when the underlying graph is generated according to a latent space model called the random dot product graph, which includes the popular stochastic block model as a special case, an out-of-sample extension based on a least-squares objective obeys a central limit theorem. In addition, we prove a concentration inequality for the out-of-sample extension of the adjacency spectral embedding based on a maximum-likelihood objective. Our results also yield a convenient framework in which to analyze trade-offs between estimation accuracy and computational expenses, which we will explore briefly if time allows in the context of both simulated and real-world data.

E0663: **Clustering and dimension reduction via Fermat distances**
Presenter: **Anna Little**, Univeristy of Utah, United States

Fermat distances are an optimal path metric which balances density-based and geometric information present in data. Graph Laplacians constructed with Fermat distance provide a useful tool for obtaining sparse graph cuts, dealing with elongated data structures, and automatically detecting the number of clusters. As the sample size converges to infinity, the spectrum of the discrete Fermat Graph Laplacian converges to the spectrum of a continuum Kolmogorov operator which generates a diffusion on the associated manifold. Unlike the Euclidean case where the diffusion occurs at a constant speed, the resulting diffusion is accelerated in regions of high density, allowing for the rapid exploration of elongated data structures. The purpose is to discuss some of the desirable properties of Fermat Graph Laplacians and highlight how the continuum limit provides a theoretical framework for understanding these properties. In addition, these metrics can be efficiently computed by restricting them to a sparse graph.

E0702: **Network estimation by adaptive mixing**
Presenter: **Can Minh Le**, University of California, Davis, United States
Co-authors: Tianxi Li

Networks analysis has been commonly used to study the interactions between units of complex systems. One problem of particular interest in learning the networks' underlying connection pattern given a single and noisy instantiation. While many methods have been proposed to address this problem in recent years, they usually assume that the true model belongs to a known class, which is not verifiable in most real-world applications. Consequently, network modeling based on these methods either suffers from model misspecification or relies on additional model selection procedures that are not well understood in theory and can potentially be unstable in practice. To address this difficulty, we propose a mixing strategy that leverages available arbitrary models to improve their individual performances. The proposed method is computationally efficient and almost tuning-free; thus, it can be used as an off-the-shelf method for network modelling. We show that the proposed method performs equally well as the oracle estimate when the true model is included as individual candidates. More importantly, the method remains robust and outperforms all current estimates even when the models are misspecified. Extensive simulation examples are used to verify the advantage of the proposed mixing method. Evaluation of link prediction performance on 385 real-world networks from six domains also demonstrates the universal competitiveness of the mixing method across multiple domains.

E0741: **Graph matching between bipartite and unipartite networks**
Presenter: **Jesus Arroyo**, Texas AM University, United States
Co-authors: Carey Priebe, Vince Lyzinski

Graph matching is the problem of aligning the vertices of two unlabeled graphs in order to maximize the shared structure across the networks; when the graphs are unipartite, this is commonly formulated as minimizing their edge disagreements. The common setting in which one of the graphs matches is a bipartite network and one is unipartite is addressed. Commonly, the bipartite networks are collapsed or projected into a unipartite graph, which potentially leads to noisy edge estimates and loss of information. A novel formulation of the graph matching problem between a bipartite and a unipartite graph is introduced, as well as methods to find the alignment. Theoretical performance is studied, providing non-asymptotic conditions that ensure the exact recovery of the matching solution. The method is illustrated in simulations and real networks, including a co-authorship-citation network pair, and brain structural and functional data.

EO385 Room Virtual R13 DEPENDENCY IN NETWORK DATA Chair: Moo K Chung

E0298: **Introduction to persistent homology for graph analysis**
Presenter: **Shizuo Kaji**, Kyushu University, Japan

Topological data analysis (TDA) is an emerging field in the intersection of mathematics and data science that utilises the power of algebraic topology to analyse data given in the form of point clouds, time-series, images, and graphs. TDA focuses on the shape of the data by looking at the local-global structures, quantifying the characteristics of data complementary to the ones obtained by conventional methods. Persistent homology (PH) is one of the main tools of TDA, and it provides quantification of holes and cliques together with their scales in a mathematically rigorous and computable way. We discuss the basic idea of PH and demonstrate its usability through examples of simple graph analysis. In particular, we see how similarity metrics and features of graphs are defined by PH and used for downstream tasks such as classification and regression.

E0982: **Modeling abnormal brain dynamics using statistical physics and MRI**
Presenter: **Alex Leow**, University of Illinois, United States

While statistical physics has been successfully used to model a wide range of complex systems and phenomena in nature, it has been underutilized in the field of computational neuroimaging. The human brain is governed by fundamental principles of functional segregation and integration; with intrinsic and induced integration directed by a balance of excitatory and inhibitory neural activities. Through the lens of statistical physics, we show how to leverage the mixed-spin ferromagnetic/antiferromagnetic Ising model to reveal patterns of excitation-inhibition (E/I) balance in brain dynamics using multi-modal magnetic resonance imaging (MRI) data (diffusion-weighted MRI and resting-state functional MRI). Consistent with findings from mouse models of Alzheimer's disease (AD) coming out of our own lab and many others, we demonstrated abnormal E/I balance towards hyper-excitation in middle-aged cognitively-normal subjects carrying the Apolipoprotein (APOE-4) allele (a genetic risk for AD) that occurs in the hippocampus.

E0999: **Spectral non-linear Granger causality for multivariate time series**
Presenter: **Hernando Ombao**, KAUST, Saudi Arabia

One of the key goals in analyzing multivariate time series is to characterize and estimate the cross-dependence structure among the components. Traditional approaches (e.g., coherence and correlation) capture only linear dependence. This serious limitation could lead to false conclusions under non-linearity. Keeping this as motivation, we propose a procedure for identifying non-linear and frequency-band-specific Granger causality (Spec NLGC) connections. The advantages of the Spec NLGC approach over traditionally used VAR-based models will be demonstrated using simulations and in the analysis of epileptic seizure EEG data. It was able to uncover non-linear dynamics and yielded novel and insightful findings. The time-evolving Spec NLGC connections give more meaningful insights regarding the frequency-specific connectivity changes at the onset of epileptic seizures as compared to VAR-based PDC connections. These confirm the viability of the proposed algorithm as a good connectivity exploration tool.

E1011: **Functional-coefficient models for multivariate time series in designed experiments: Applications to brain signals**
Presenter: **Paolo Victor Redondo**, King Abdullah University of Science and Technology, Saudi Arabia
Co-authors: Raphael Huser, Hernando Ombao

To study the neurophysiological dynamics of attention deficit hyperactivity disorder (ADHD), clinicians use multichannel electroencephalography (EEG) which records neuronal electrical activity in the cortex. The most commonly-used metric in ADHD is the theta-to-beta power ratio (TBR) which is derived from the spectrum of the EEGs. However, initial findings for this measure have not yet been replicated in other studies. Instead of focusing on spectral power, this paper develops a novel model for investigating dependence between channels in the entire network. Although dependence measures such as coherence and partial directed coherence (PDC) are well explored in studying brain connectivity, these measures only capture linear dependence. Moreover, in designed clinical experiments, it is observed that these dependence measures can vary across subjects even within a homogeneous group. Hence, to address these limitations, we propose the mixed-effects functional-coefficient autoregressive (MX-FAR) model. The advantages of MX-FAR are the following: (1.) it captures non-linear dependence between channels; (2.) it is nonparametric and hence flexible; (3.) it can capture differences between groups; (4.) it accounts for variation across subjects; (5.) the framework easily incorporates well-known inference methods from mixed-effects models. Finally, we showcase the MX-FAR model through numerical experiments and report novel findings from the analysis of EEG signals in ADHD.

EO033 **Room Virtual R2** NEW DEVELOPMENTS IN DESIGN AND ANALYSIS OF EXPERIMENTS **Chair: John Stufken**

E0194: **A convex approach to optimum design of experiments with correlated observations**
Presenter: **Werner Mueller**, Johannes Kepler University Linz, Austria
Co-authors: Andrej Pazman, Markus Hainy

Optimal design of experiments for correlated processes is an increasingly relevant and active research topic. Until now only heuristic methods were available without a possibility to judge their quality. We complement the virtual noise approach by a convex formulation and an equivalence theorem comparable to the uncorrelated case. Hence, it is now possible to provide an upper performance bound against which alternative design methods can be judged. We provide a comparison using some classical examples from the literature.

E0255: **Design selection for 2-level supersaturated designs**
Presenter: **Rakhi Singh**, UNC Greensboro, United States
Co-authors: John Stufken

The commonly used design optimality criteria are inadequate for selecting supersaturated designs. As a result, there is extensive literature on alternative optimality criteria within this context. Most of these criteria are rather ad hoc and are not directly related to the primary goal of experiments that use supersaturated designs, which is factor screening. Especially, unlike almost any other optimal design problem, the criteria are not directly related to the method of analysis. An assumption needed for the analysis of supersaturated designs is the assumption of effect sparsity. Under this assumption, a popular method of analysis for 2-level supersaturated designs is the Gauss-Dantzig Selector (GDS), which shrinks many estimates to 0. We develop new design selection criteria inspired by the GDS and establish that designs that are better under these criteria tend to perform better as screening designs than designs obtained using existing criteria.

E0621: **Particle swarm exchange algorithms with applications in generating optimal model-discrimination designs**
Presenter: **Ray-Bing Chen**, National Cheng Kung University, Taiwan

Exchange-type algorithms have been commonly used to construct optimal designs. As these algorithms may converge to a local optimum, the typical procedure requires the use of several randomly chosen initial designs. Thus, the search for the optimal design can be conducted by performing several independent optimizations. We propose a general framework that combines exchange algorithms with particle swarm intelligence techniques. The main strategy is to represent each initial design as a particle and make the algorithm share information from various converging paths from those initial designs. This amounts to conducting one coordinated optimization instead of several independent optimizations. The proposed general algorithm is called the particle swarm exchange (PSE) algorithm. We compare the performance of PSE with those of two commonly used exchange algorithms – the columnwise-pairwise (CP) exchange algorithm for designs with structural requirements and the coordinate exchange algorithm for designs without such requirements. In the context of model-robust discriminating designs, we demonstrate that PSE typically performs as well as or, very often, better than the corresponding pure exchange algorithms.

E0805: **Experimental designs for functional modeling of longitudinal data**
Presenter: **MingHung Kao**, Arizona State University, United States

Functional data analysis (FDA) has gained much popularity in extracting useful information from repeated measurements collected at various points on a domain, such as time. A crucial step for rendering a precise and valid inference is to have a high-quality sampling schedule to sample informative data from the underlying function. We are concerned with this design problem of FDA, and propose efficient computational approaches for obtaining good designs to rein in cost. Our proposed approaches generate high-quality designs to allow a precise recovery of the underlying function, as well as precise prediction with functional linear and quadratic regressions.

EO119 **Room Virtual R3** ADVANCES IN BAYESIAN METHODS AND COMPUTATION **Chair: Jouchi Nakajima**

E0198: **Tree boosting for learning probability measures**
Presenter: **Naoki Awaya**, Duke University, United States
Co-authors: Li Ma

Learning probability measures based on an i.i.d. sample is a fundamental inference task, but is challenging when the sample space is high-dimensional. Inspired by the success of tree boosting in high-dimensional classification and regression, we propose a tree boosting method for learning high-dimensional probability distributions. We formulate concepts of “addition” and “residuals” on probability distributions in terms of compositions of a new, more general notion of multivariate cumulative distribution functions (CDFs) than classical CDFs. This then gives rise to a simple boosting algorithm based on forward-stagewise (FS) fitting of an additive ensemble of measures, which sequentially minimizes the entropy loss. The output of the FS algorithm allows the analytic computation of the probability density function for the fitted distribution. It also provides an exact simulator for drawing independent Monte Carlo samples from the fitted measure. Typical considerations in applying boosting–namely choosing the number of trees, setting the appropriate level of shrinkage/regularization in the weak learner, and the evaluation of variable importance–can all be accomplished in an analogous fashion to traditional boosting in supervised learning. Numerical experiments confirm that boosting can substantially improve the fit to multivariate distributions compared to the state-of-the-art single-tree learner and is computationally efficient.

E0961: **Simultaneous graphical dynamic linear models for macroeconomic policy**
Presenter: **Meng Xie**, Duke University, United States
Co-authors: Knut Are Aastveit, Kaoru Irie, Mike West

There is abundant interest in incorporating information from many economic time series to improve forecasts and inform policy. However, traditional methods, including time-varying vector autoregression models, can become over-parameterized as the number of series increases. Simultaneous graphical dynamic linear models (SGDLMs) provide a flexible and computationally efficient approach for modeling macroeconomic series, and also enable novel order-free structural analyses for policy decision-making. In SGDLMs, each data series is modeled with its own specialized and limited set of sparse predictors, with simultaneous relationships represented through contemporaneous predictors. At each time point, the posterior distributions of states and volatilities in each of the univariate models are independently updated for the new observation. Then, the series-specific models are recoupled to account for cross-series dependencies, and decoupled again to continue sequential estimation and forecasting. We apply SGDLMs to forecast macroeconomic series from the Federal Reserve Economic Data database, using interventions to sequential updating for dynamic variable selection.

E0295: **A realized multi-factor regression using a multivariate stochastic volatility model**
Presenter: **Tsunehiro Ishihara**, Takasaki City University of Economics, Japan

A multifactor model is a standard tool in financial econometrics. We introduce information from high-frequency data into the multifactor model using realized measures. We calculate market, size, value quasi-factors, and their realized covariance matrix from intra-daily announced market indices. We propose a time-varying coefficient factor regression model and transform it into the multivariate stochastic volatility model with realized covariance. Bayesian estimation using the Markov chain Monte Carlo method is also proposed. We present empirical illustrations using several sector indices of the Japanese stock market. It is shown that the proposed factors behave similarly to the Fama-French three factors. In addition, the coefficients of the factors were found to change over time. The model is also compared with the realized stochastic volatility model without the factor. The Bayesian predictive likelihood loss function shows that factor models perform well, especially for longer forecast horizons.

E0256: **Realized stochastic volatility models with skew-t distributions**
Presenter: **Makoto Takahashi**, Hosei University, Japan
Co-authors: Yasuhiro Omori, Toshiaki Watanabe, Yuta Yamauchi

Predicting volatility and quantiles of financial returns is essential to measure the financial tail risk such as value-at-risk and expected shortfall. There are two important aspects of volatility and quantile forecasts: the distribution of financial returns and the estimation of the volatility. Building on the traditional stochastic volatility model, the realized stochastic volatility model incorporates realized volatility as the precise estimator of the volatility. Using three types of skew-*t* distributions, the model is extended to capture the well-known characteristics of the return distribution, namely skewness and heavy tails. In addition to the normal and Student's *t* distributions, included as the special cases of the skew-*t* distributions, two of them contain the skew-normal, and hence allows more flexible modeling of the return distribution. The Bayesian estimation scheme via a Markov chain Monte Carlo method is developed and applied to major stock indices. The estimation results show that the negative skewness is evident for both indices whereas the heavy tail is largely captured by the realized stochastic volatility, and thus demonstrate that the model with the skew-normal distribution performs well. On the other hand, the prediction results suggest that incorporating both skewness and heavy tail to daily returns is important for volatility and quantile forecasts, especially in a high-volatility period.

EO101 Room Virtual R4 STATISTICAL MODELING OF CHALLENGING DATA Chair: Yuedong Wang

E0757: **Functional mixed effects clustering with application to longitudinal urologic chronic pelvic pain syndrome symptom data**
Presenter: **Wensheng Guo**, University of Pennsylvania., United States

By clustering patients with the urologic chronic pelvic pain syndromes (UCPPS) into homogeneous subgroups and associating these subgroups with baseline covariates and other clinical outcomes, we provide opportunities to investigate different potential elements of pathogenesis, which may also guide us in the selection of appropriate therapeutic targets. Motivated by the longitudinal urologic symptom data with extensive subject heterogeneity and differential variability of trajectories, we propose a functional clustering procedure where each subgroup is modeled by a functional mixed-effects model, and the posterior probability is used to iteratively classify each subject into different subgroups. The classification takes into account both group-average trajectories and between-subject variabilities. We develop an equivalent state-space model for efficient computation. We also propose a cross-validation-based Kullback-Leibler information criterion to choose the optimal number of subgroups. We apply our methods to longitudinal bi-weekly measures of a primary urological urinary symptoms score from a UCPPS longitudinal cohort study and identify four subgroups ranging from moderate decline, mild decline, stable and mild increasing.

E0192: **Linear models for doubly multivariate data with exchangeably distributed errors and site-dependent covariates**
Presenter: **Anuradha Roy**, The University of Texas at San Antonio, United States
Co-authors: Timothy Opheim

Doubly multivariate repeated measures data, where observations are made on p response variables and each response variable is measured over n sites or time points, construct matrix-valued response variable, and arise across a wide range of disciplines, including medical, environmental and agricultural studies. In many practical situations, response variables are affected by several explanatory variables, and these explanatory variables may vary over sites or time points too. In this case, we say that the data have site-dependent covariates, which construct a matrix-valued explanatory variable. Rao's score test (RST) for testing the intercept and slope parameters for doubly multivariate linear models with site-dependent covariates is developed and applied to an agricultural dataset. Monte Carlo simulations indicate that the RST statistic is much more accurate than its counterpart likelihood ratio test (LRT) statistic and it takes significantly less computation time than the LRT statistic.

E0179: **Optimal-k sequence for difference-based methods in nonparametric regression**
Presenter: **Tiejun Tong**, Hong Kong Baptist University, Hong Kong

Difference-based methods have been attracting increasing attention in nonparametric regression, in particular for estimating the residual variance. To implement the estimation, one needs to choose an appropriate difference sequence, mainly between the optimal difference sequence and the ordinary difference sequence. This difference sequence selection is a fundamental problem in nonparametric regression, and it remains unresolved until recently. We propose to further advance the difference sequence selection from another unique perspective, which creates a new family of difference sequences called the optimal-k sequence. Our proposed difference sequence not only provides a better bias-variance trade-off but also includes the optimal and the ordinary difference sequences as two important special cases. Through theoretical and numerical studies, we demonstrate that the optimal-k sequence has been pushing the boundaries of our knowledge in difference-based methods in nonparametric regression, and more importantly, it always performs the best in practical situations.

E0207: **A sumsampling method for regression problems based on minimum energy criterion**
Presenter: **Wenlin Dai**, Renmin University of China, China

The extraordinary amounts of data generated in science today pose heavy demands on computational resources and time, which hinders the implementation of various statistical methods. An efficient and popular strategy of downsizing data volumes and hence alleviating these challenges is subsampling. However, the existing methods either rely on specific assumptions for the underlying models or acquire only partial information from the available data. We propose a novel approach, termed adaptive subsampling, that is based on the minimum energy criterion (ASMEC). The proposed method requires no explicit model assumptions and 'smartly' incorporates information on covariates and responses. ASMEC subsamples possess two desirable properties: space-filling and spatial adaptiveness to the full data. We investigate the theoretical properties of the ASMEC estimator under the smoothing spline regression model and show that it converges at an identical rate to two recently proposed basis selection methods. The effectiveness and robustness of the ASMEC approach are also supported by a variety of simulated examples and two real-life examples.

EO129 **Room Virtual R5** **RECENT ADVANCES IN MACHINE LEARNING** **Chair: Yiming Ying**

E0364: **Total stability of SVMs and localized SVMs**
Presenter: **Hannes Koehler**, University of Bayreuth, Germany
Co-authors: Andreas Christmann

Regularized kernel-based methods such as support vector machines (SVMs) typically depend on the underlying probability measure (respectively data set) as well as on the regularization parameter and the kernel that are used. Whereas classical statistical robustness only considers the effect of small perturbations in the probability measure alone, we investigate how the resulting predictor is influenced by simultaneous slight variations in the whole triple of probability measure, regularization parameter and kernel. Existing results from the literature are considerably generalized and improved. In order to also make them applicable to big data, where regular SVMs suffer from their super-linear computational requirements, the results are transferred to the context of localized learning.

E0640: **Approximation of nonlinear functionals using deep ReLU networks**
Presenter: **Jun Fan**, Hong Kong Baptist University, Hong Kong

In recent years, functional neural networks have been proposed and studied in order to approximate nonlinear continuous functionals. However, their theoretical properties are largely unknown beyond the universality of approximation or the existing analysis does not apply to the rectified linear unit (ReLU) activation function. To fill in this void, we investigate here the approximation power of functional deep neural networks associated with the ReLU activation function by constructing a piecewise linear interpolation under a simple triangulation. In addition, we establish rates of approximation of the proposed functional deep ReLU networks under mild regularity conditions.

E0672: **Learning theory of stochastic gradient descent**
Presenter: **Yunwen Lei**, Southern University of Science and Technology, China

Stochastic Gradient Descent (SGD) has become the workhorse behind many machine learning problems. Despite its promising success in applications, the theoretical analysis is still not satisfactory. We will discuss the learning theory of SGD. We will introduce new algorithmic stability concepts to relax the existing restrictive assumptions and improve the existing learning rates. Our results show new connections between generalization and optimization, which illustrate how a better learning performance can be achieved by early stopping.

E0678: **A statistical learning assessment of Huber regression**
Presenter: **Yunlong Feng**, The State University of New York at Albany, United States
Co-authors: Qiang Wu

Some theoretical understanding of Huber regression from a statistical learning viewpoint will be reported. The focus will be on the following two aspects: (1) how Huber regression estimators learn the conditional mean function and (2) why they work in the absence of light-tailed noise assumptions. To answer the two questions, we will report the following efforts we made. First, the usual risk consistency property of Huber regression estimators, which is usually pursued in learning, cannot guarantee their learnability in mean regression; second, it is argued that Huber regression should be implemented in an adaptive way to perform mean regression, implying that one needs to tune the scale parameter in accordance to the sample size and the moment condition of the noise; third, with an adaptive choice of the scale parameter, Huber regression estimators can be mean regression calibrated under $(1+\varepsilon)$-moment conditions ($\varepsilon > 0$), and exponential-type convergence rates for Huber regression estimators can be established.

EO131 **Room Virtual R6** **RECENT DEVELOPMENTS IN DIMENSION REDUCTION AND MULTIVARIATE ANALYSIS** **Chair: Yeonhee Park**

E0797: **A comprehensive Bayesian framework for envelope models**
Presenter: **Saptarshi Chakraborty**, State University of New York at Buffalo, United States
Co-authors: Zhihua Su

The envelope model aims to increase efficiency in multivariate analysis by utilizing dimension reduction techniques. It has been used in many contexts including linear regression, generalized linear models, matrix/tensor variate regression, reduced rank regression, and quantile regression, and has shown the potential to provide substantial efficiency gains. Most of these advances have been made from a frequentist perspective, and the literature addressing envelope models from a Bayesian point of view is still sparse. The Bayesian paradigm provides unique flexibility in terms of incorporating prior knowledge if available, and coherent quantification of all modeling uncertainties, including in envelope dimension selection. The objective is to propose a computationally feasible comprehensive Bayesian framework that is applicable across various envelope model contexts. We provide a simple block Metropolis-within-Gibbs MCMC sampler for efficient practical implementations of our method. Simulations and data examples are included for illustration.

E0827: **A Bayesian approach to envelope quantile regression**
Presenter: **Minji Lee**, Edwards Lifesciences, United States
Co-authors: Saptarshi Chakraborty, Zhihua Su

The enveloping approach employs sufficient dimension reduction techniques to gain estimation efficiency and has been used in several multivariate analysis contexts. However, its Bayesian development has been sparse, and the only Bayesian envelope construction is in the context of linear regression. We propose a Bayesian envelope approach to quantile regression, using a general framework that may potentially aid enveloping in other contexts as well. The proposed approach is also extended to accommodate censored data. Data augmentation Markov chain Monte Carlo algorithms are derived for approximate sampling from the posterior distributions. Simulations and data examples are included for illustration.

E0530: **Bayesian inference for multivariate probit model with latent envelope**
Presenter: **Kwangmin Lee**, University of Wisconsin-Madison, United States

The response envelope model is known to be an efficient method to estimate the regression coefficient under the context of the multivariate linear regression model. It identifies material and immaterial parts of responses to improve estimation efficiency. The response envelope model has been investigated only for continuous response variables. We suggest the multivariate probit model with latent envelope, in short, the probit envelope model, to apply the idea of the response envelope models to multivariate binary response data. In the probit envelope model, we employ the Gaussian latent vector formulation of the multivariate probit model. We assume that the latent vector follows the assumption by the response envelope models, i.e., the latent vector is assumed to have a covariate-invariant part called the immaterial part. Then, the response vector of the probit envelope model is derived by thresholding the latent vector. We address the identifiability of the probit envelope model by reparametrizing the model, and we suggest an MCMC algorithm for the Bayesian inference. We illustrate the probit envelope model via simulation studies and real data analysis, in which we also apply the probit envelope model to the multilabel classification problem.

E0612: **Envelope model for function-on-function linear regression**
Presenter: **Zhihua Su**, University of Florida, United States
Co-authors: Bing Li, Dennis Cook

The envelope model is a recently developed methodology for multivariate analysis that enhances estimation accuracy by exploiting the relation between the mean and eigenstructure of the covariance matrix. We extend the envelope model to function-on-function linear regression, where the response and the predictor are assumed to be random functions in Hilbert spaces. We use a double envelope structure to accommodate the eigenstructures of the covariance operators for both the predictor and the response. The central idea is to establish a one-to-one relation between the functional envelope model and the multivariate envelope model and estimate the latter by the existing method. We also developed the asymptotic theories, confidence and prediction bands, an order determination method along with its consistency, and a characterization of the efficiency gained by the proposed model. Simulation comparisons with the standard function-on-function regression and data applications show significant improvement by our method in terms of cross-validated prediction error.

EO241 Room Virtual R7 ESTIMATION AND INFERENCE OF HIGH DIMENSIONAL TIME SERIES Chair: Danna Zhang

E0493: **Tensor methods for high-dimensional time series modeling**
Presenter: **Yao Zheng**, University of Connecticut, United States

Tensor decomposition is a powerful dimensionality reduction tool that has gained much interest in modern machine learning applications. However, its development in the areas of time series analysis and econometrics is still in its infancy. We will present some recent work on high-dimensional time series modeling via tensor methods. Specifically, we will discuss the use of Tucker decomposition in high-dimensional vector autoregressive modeling, tensor-valued autoregressive time series modeling, and high-dimensional mixed-data sampling (MIDAS) regression. The focus will be more on motivations, model formulations, interpretations and empirical examples. Estimation methods and theoretical properties will be mentioned briefly.

E0531: **Factorized binary search: Change point detection in the network structure of multivariate high-dimensional time series**
Presenter: **Ivor Cribben**, Alberta School of Business, Canada

The purpose is to introduce factorized binary search (FaBiSearch), a novel change point detection method in the network structure of multivariate high-dimensional time series. FaBiSearch uses non-negative matrix factorization, an unsupervised dimension reduction technique, and a new binary search algorithm to identify multiple change points. In addition, we propose a new method for network estimation for data between change points. We show that FaBiSearch outperforms another state-of-the-art method on simulated data sets and we apply FaBiSearch to a resting-state and to a task-based fMRI data set.

E0561: **CP factor model for dynamic tensors**
Presenter: **Yuefeng Han**, Rutgers University, United States

Observations in various applications are frequently represented as a time series of multidimensional arrays, called tensor time series, preserving the inherent multidimensional structure. We present a factor model approach, in a form similar to tensor CP decomposition, to the analysis of high-dimensional dynamic tensor time series. As the loading vectors are uniquely defined but not necessarily orthogonal, it is significantly different from the existing tensor factor models based on Tucker-type tensor decomposition. The model structure allows for a set of uncorrelated one-dimensional latent dynamic factor processes, making it much more convenient to study the underlying dynamics of the time series. A new high order projection estimator is proposed for such a factor model, utilizing the special structure and the idea of the higher-order orthogonal iteration procedures commonly used in the Tucker-type tensor factor model and general tensor CP decomposition procedures. Theoretic al investigation provides statistical error bounds for the proposed methods, which shows the significant advantage of utilizing the special model structure.

E0909: **Frequency-domain graphical models for multivariate time series**
Presenter: **Sumanta Basu**, Cornell University, United States

Graphical models offer a powerful framework to capture intertemporal and contemporaneous relationships among the components of a multivariate time series. For stationary time series, these relationships are encoded in the multivariate spectral density matrix and its inverse. We will present adaptive thresholding and penalization methods for the estimation of these objects under suitable sparsity assumptions. We will discuss new optimization algorithms and investigate the consistency of estimation under a double-asymptotic regime where the dimension of the time series increases with sample size. If time permits, we will introduce a frequency-domain graphical modeling framework for multivariate nonstationary time series that captures a new property called conditional stationarity.

EO207 Room Virtual R8 MODERN STATISTICAL METHODS FOR ENVIRONMENTAL DATA ANALYSIS Chair: Whitney Huang

E0478: **Distributional validation of precipitation data products with spatially varying mixture models**
Presenter: **Lynsie Warr**, University of California Irvine, United States
Co-authors: Matthew Heaton, William Christensen, Philip White, Summer Rupper

The high mountain regions of Asia contain more glacial ice than anywhere on the planet outside the polar regions. Because the large populations living in the Indus watershed region are reliant on glacial melt for freshwater, understanding the factors that affect glacial melt and the impacts of climate change on the region is important for managing these natural resources. While there are multiple climate data products (e.g. reanalysis and global climate models) available to study these factors and impacts, each product has a different amount of skill in projecting a given climate variable, such as precipitation. We develop a spatially varying mixture model to compare the distribution of precipitation in the High Mountain Asia region as produced by climate models with the corresponding distribution from in situ observations from the Asian Precipitation Highly Resolved Observational Data Integration Towards Evaluation (APHRODITE) data product. Parameter estimation is carried out via an efficient Markov chain Monte Carlo algorithm. Each estimated distribution from each climate data product is validated against APHRODITE using a spatially varying Kullback-Leibler divergence measure.

E0487: **Calibration of spatio-temporal forecasts from urban air pollution data with sparse recurrent neural networks**
Presenter: **Matthew Bonas**, University of Notre Dame, United States
Co-authors: Stefano Castruccio

Data collected from personal air quality monitors have become an increasingly valuable tool to complement existing public health monitoring systems in urban areas. The potential of using such 'citizen science data' for automatic early warning systems is hampered by the lack of models able to capture the high resolution, nonlinear spatio-temporal features stemming from local emission sources such as traffic, residential heating and commercial activities. A machine learning approach si proposed to forecast high-frequency spatial fields which has two advantages from standard methods in time: 1) sparsity of the neural network via a spike-and-slab prior, and 2) a small parametric space. The introduction of stochastic neural networks generates additional uncertainty, and we propose a fast approach to ensure that the forecast is correctly assessed (calibration), both marginally and spatially. We focus on assessing exposure to urban air pollution in San Francisco, and our results suggest an improvement of over 30% in the mean squared error over the standard time series approach with a calibrated forecast for up to 5 days.

E0502: **Joint modeling of wind speed and wind direction through a conditional approach**
Presenter: **Eva Murphy**, Clemson University, United States

Atmospheric near-surface wind speed and wind direction play an important role in many applications, ranging from air quality modeling, building design, and wind turbine placement to climate change research. It is, therefore, crucial to accurately estimate the joint probability distribution of wind speed and direction. We develop a conditional approach to model the two variables, where the joint distribution is decomposed into the product of the marginal distribution of wind direction and the conditional distribution of wind speed given wind direction. To accommodate the circular nature of wind direction a von Mises mixture distributions are used; the conditional wind speed distribution is modeled as a directional dependent Weibull distribution via a two-stage procedure, consisting of a binned Weibull parameter estimation, followed by a harmonic regression used to model the dependence of the Weibull parameters on wind direction. A Monte Carlo simulation study suggests that our method outperforms an alternative method that uses periodic quantile regression in terms of estimation efficiency and bias. We illustrate our method by using the outputs of climate model simulations to investigate how the joint distribution of wind speed and direction may change under some future climate scenarios.

E0559: **Accounting for the spatial structure of weather systems in detected changes in precipitation extremes**
Presenter: **Likun Zhang**, Lawrence Berkeley National Lab, United States

The detection of changes over time in the distribution of precipitation extremes is complicated by noise at the spatial scale of weather systems. Traditional approaches for quantifying observed changes in extreme precipitation return values are often based on single-station analyses, which fail to account for the spatial coherence of individual storms and hence yield unrealistic and potentially misleading estimates of the true underlying changes in extremes. We demonstrate how the use of a flexible statistical method that robustly accounts for the so-called "storm dependence" in measurements of daily precipitation removes a challenging source of noise and results in improved estimates of changes in the distribution of precipitation extremes. Furthermore, the analysis provides important insights into the spatial structure of seasonal extreme precipitation across increasing event rarity. Applying the methodology to long-term in situ records of daily precipitation from the central United States, we find that properly accounting for storm dependence leads to increased detection of statistically significant changes in return values as compared with existing approaches. We also find that simultaneous precipitation extremes in this region tend to organize on scales of 100-200 km for high quantile levels, which is consistent with observed spatial patterns in the NEXRAD Stage IV radar-based data set.

EO275 Room Virtual R9 SEQUENTIAL ANALYSIS AND ONLINE UPDATING Chair: Yan Zhuang

E0981: **Fixed-accuracy big data estimation of population Gini income inequality index: Practical distribution-free strategies**
Presenter: **Nitis Mukhopadhyay**, University of Connecticut-Storrs, United States

Recently, elegant sequential fixed-width confidence interval (FWCI) and minimum risk point estimation (MRPE) methodologies for $G(F)$ have been developed. $G(F)$ is the celebrated Gini income inequality index in a population associated with an unknown distribution function F having its support on positive real numbers. We revisit both problems from the vantage point of big data science by proposing newly designed easy-to-implement sequential estimation strategies with nearly minimal computational complexities and technical difficulties. Inference techniques recently introduced will be emphasized. We show that these new sequential estimation strategies have a wide range of appealing asymptotic properties including both first-order and second-order approximations. The proposed approaches are flexible enough to embrace other non-standard inference problems in the future.

E0217: **An adaptive Monte Carlo method to estimate confidence interval for population sizes under mark-recapture-mark sampling**
Presenter: **Debanjan Bhattacharjee**, Utah Valley University, United States
Co-authors: Ivair Silva, Yan Zhuang

The conventional mark-recapture strategy is modified to estimate the size (N) of a finite population. In this new procedure non-marked, resampled items are marked before they are released back into the population. A sequential adaptive stopping rule for fixed-length-interval-estimation of N is proposed. A Monte Carlo solution is derived and compared with the accelerated sequential method. Estimating sizes of finite populations can become particularly relevant in knowing the total number of patients infected with a disease at a particular time in a geographical region. The new method is illustrated with a simulation that estimates the number of infected COVID-19 individuals in a near-closed population. In addition, we present a numeric application inspired by the problem of estimating the population size of endangered monkeys of the Atlantic Forest in Brazil.

E0363: **Online updating of survival analysis**
Presenter: **Jing Wu**, University of Rhode Island, United States

When large amounts of survival data arrive in streams, conventional estimation methods become computationally infeasible since they require access to all observations at each accumulation point. We develop online updating methods for carrying out survival analysis under the Cox proportional hazards model in an online-update framework. Our methods are also applicable with time-dependent covariates. Specifically, we propose online-updating estimators as well as their standard errors for both the regression coefficients and the baseline hazard function. Extensive simulation studies are conducted to investigate the empirical performance of the proposed estimators. A large colon cancer data set from the Surveillance, Epidemiology, and End Results (SEER) program and a large venture capital (VC) data set with time-dependent covariates are analyzed to demonstrate the utility of the proposed methodologies.

E0427: **Maximum precision estimation for a step-stress model using two-stage methodologies**
Presenter: **Sudeep Bapat**, Indian Institute of Management Indore, India

A two-stage sequential procedure to estimate the parameters of a cumulative exposure model under an accelerated testing scenario is discussed. We focus on a step-stress model where the stress level is updated after a pre-specified number of failures occur, which is also random. This is termed as the "random stress change time" in the literature. To obtain maximum precision, a certain variance optimality criterion is applied. A pseudo-real data example from reliability studies is also analysed to outline the performance of the proposed methodology.

Monday 06.06.2022 09:30 - 11:10 Parallel Session L – EcoSta2022

EO085 Room Live Theater (Hybrid 2) VARIATIONAL INFERENCE IN STATISTICS AND ECONOMETRICS I Chair: Pierre Alquier

E0494: **Gibbs posterior distributions: Construction, concentration, and calibration**
Presenter: **Ryan Martin**, North Carolina State University, United States

Bayesian inference has certain advantages, but a fully (and generally correctly) specified statistical model is needed to realize those. What if the quantity of interest is not naturally described as a "model parameter"? Then there is no sense in which a specified statistical model could be "correct" and, hence, there is a risk of model misspecification bias. To avoid this bias, one can construct a so-called Gibbs posterior that directly targets the quantity of interest, compared to a Bayesian posterior that does so only indirectly through a (possibly misspecified) statistical model and marginalization. First, we will discuss the Gibbs posterior construction; second, we will present asymptotic concentration properties of Gibbs posteriors, with a focus on specific examples; and, finally, we will discuss the need to properly calibrate the Gibbs posterior so that inferences (or predictions) are valid, and present an algorithm that achieves this.

E0533: **On the robustness to misspecification of -posteriorsand their variational approximations**
Presenter: **Jose Luis Montiel Olea**, Columbia University, United States

Alpha-posteriors and their variational approximations distort standard posterior inference by downweighting the likelihood and introducing variational approximation errors. We show that such distortions, if tuned appropriately, reduce the Kullback-Leibler (KL) divergence from the true, but perhaps infeasible, posterior distribution when there is potential parametric model misspecification. To make this point, we derive a Bernstein-von Mises theorem showing convergence in total variation distance of alpha-posteriors and their variational approximations to limiting Gaussian distributions. We use these limiting distributions to evaluate the KL divergence between true and reported posteriors. We show the KL divergence is minimized by choosing alpha strictly smaller than one, assuming there is a vanishingly small probability of model misspecification. The optimized value of alpha becomes smaller as the misspecification becomes more severe. The optimized KL divergence increases logarithmically in the magnitude of misspecification and not linearly as with the usual posterior. Moreover, the optimized variational approximations of alpha-posteriors can induce additional robustness to model misspecification, beyond that obtained by optimally downweighting the likelihood.

E0988: **On statistical and algorithmic aspects of variational inference**
Presenter: **Anirban Bhattacharya**, Texas AM University, United States

Statistical as well as algorithmic aspects of variational inference in Bayesian hierarchical models are investigated. The focus is beyond the mean-field setup. We present recent findings in a number of examples such as logit models as well as dynamic models.

EO261 Room Main Theater (Hybrid 1) EXTERNAL VALIDITY AND DATA FUSION IN CAUSAL INFERENCE Chair: Caleb Miles

E0668: **Generalizing trial evidence to target populations in non-nested designs: Applications to AIDS clinical trials**
Presenter: **Ashley Buchanan**, University of Rhode Island, United States
Co-authors: Fan Li, Stephen Cole

Comparative effectiveness evidence from randomized trials may not be directly generalizable to a target population of substantive interest when, as in most cases, trial participants are not randomly sampled from the target population. Motivated by the need to generalize evidence from two trials conducted in the AIDS Clinical Trials Group (ACTG), we consider weighting, regression, and doubly robust estimators to estimate the causal effects of HIV interventions in a specified population of people living with HIV in the USA. We focus on a non-nested trial design and discuss strategies for both point and variance estimation of the target population's average treatment effect. Specifically, in the generalizability context, we demonstrate both analytically and empirically that estimating the known propensity score in trials does not increase the variance for each of the weighting, regression, and doubly robust estimators. We apply these methods to generalize the average treatment effects from two ACTG trials to specified target populations and operationalize key practical considerations. Finally, we report on a simulation study that investigates the finite-sample operating characteristics of the generalizability estimators and their sandwich variance estimators.

E0942: **Causally interpretable meta-analysis: Transporting inferences from multiple randomized trials to a target population**
Presenter: **Issa Dahabreh**, Harvard University, United States

Methods are presented for causally interpretable meta-analyses that combine information from multiple randomized trials to draw causal inferences for a target population of substantive interest. We consider identifiability conditions, derive implications of the conditions for the law of the observed data, and obtain identification results for transporting causal inferences from a collection of independent randomized trials to a new target population in which experimental data may not be available. We propose an estimator for the potential outcome mean in the target population under each treatment studied in the trials. The estimator uses covariate, treatment, and outcome data from the collection of trials, but only covariate data from the target population sample. We show that it is doubly robust, in the sense that it is consistent and asymptotically normal when at least one of the models it relies on is correctly specified. We study the finite sample properties of the estimator in simulation studies and demonstrate its implementation using data from a multi-center randomized trial.

E0735: **Robust and efficient nonparametric estimation of transported total causal effects and mediation causal effects**
Presenter: **Kara Rudolph**, Columbia University, United States
Co-authors: Ivan Diaz

The focus is on identifying and estimating transported causal effects in the context of encouragement-design interventions—those that encourage uptake of exposure of interest. We consider types of transported total effects, like the intent-to-treat average treatment effect, the average treatment effect of the exposure, and the complier average treatment effect. We also consider transported interventional indirect and direct effects of the encouragement on the outcome that operate through mediators or not, respectively. We describe robust and efficient estimators for each type of transported effect and apply them to motivating research questions from the Moving to Opportunity Study (MTO). MTO is a large-scale encouragement-design intervention in which Section 8 housing vouchers, which encourage families in public housing to move by subsidizing rents on the private market, were randomly assigned. In the context of MTO, we use the transport estimators in service of understanding the reasons for differences in site-specific effect estimates. These transport estimators may also be useful in other scenarios-including, to generate place-specific intervention effect estimates, in problems related to surrogacy, or in other data-fusion-related problems. We end with current work on extending these identification results and estimators to accommodate more general data structures.

E0905: **Efficient estimation under data fusion**
Presenter: **Alex Luedtke**, University of Washington, United States

The aim is to make inferences about a smooth, finite-dimensional parameter by fusing data from multiple sources together. Previous works have studied the estimation of a variety of parameters in similar data fusion settings, including the estimation of the average treatment effect, optimal treatment rule, and average reward, with the majority of them merging one historical data source with covariates, actions, and rewards and one data source of the same covariates. We consider the general case where one or more data sources align with each part of the distribution of the target population, for example, the conditional distribution of the reward given actions and covariates. We describe potential gains in efficiency that can arise from fusing these data sources together in a single analysis, which we characterize by a reduction in the semiparametric efficiency bound. We also provide a general means to construct estimators that achieve these bounds. In numerical experiments, we show marked improvements in efficiency from using our proposed estimators rather than their natural alternatives. Finally, we illustrate the magnitude of efficiency gains that can be realized in vaccine immunogenicity studies by fusing data from two HIV vaccine trials.

EO103 Room Virtual R1 RECENT DEVELOPMENTS IN GRAPHICAL MODELS Chair: Kuang-Yao Lee

E0572: **Simultaneous inference in multiple matrix-variate graphs for high-dimensional neural recordings**
Presenter: **Zhao Ren**, University of Pittsburgh, United States

As large-scale neural recordings become common, many neuroscientific investigations are focused on identifying functional connectivity from spatio-temporal measurements in two or more brain areas across multiple sessions. Spatial-temporal data in neural recording can be viewed as matrix-variate data, where the first dimension is time and the second dimension is space. We exploit the multiple matrix-variate Gaussian Graphical model (MGGM) to encode the common underlying spatial functional connectivity across multiple sessions of neural recordings. By effectively integrating information across multiple graphs, we develop a novel inferential framework that allows simultaneous testing to detect meaningful connectivity for a target edge subset of arbitrary size. The test statistics are based on a group penalized regression approach and a high-dimensional Gaussian approximation technique. The validity of simultaneous testing is demonstrated theoretically under very mild assumptions on sample size and non-stationary autoregressive temporal dependence. We demonstrate the efficacy of the new method through both simulations and an experimental study with multiple local field potential (LFP) recordings in Prefrontal Cortex (PFC) and visual area V4 during a memory-guided saccade task.

E0971: **Nonparametric assessment of conditional dependence using a restricted score test**
Presenter: **Aaron Hudson**, Unviersity of California, Berkeley, United States

Infinite-dimensional parameters that can be defined as the minimizer of a population risk arise naturally in many applications. Classic examples include the conditional mean function and the density function. Though there is extensive literature on constructing consistent estimators for infinite-dimensional risk minimizers, there is limited work on quantifying the uncertainty associated with such estimates via, e.g., hypothesis testing and construction of confidence regions. We propose a general inferential framework for infinite-dimensional risk minimizers as a nonparametric extension of the score test. We illustrate that our framework requires only mild assumptions and is applicable to a variety of estimation problems. As an example, we apply our proposed methodology to test for conditional dependence in a graphical model for which the conditional mean of any node given all remaining nodes takes an arbitrary additive form.

E0991: **Functional causal modeling via Karhunen-Loeve expansions**
Presenter: **Kuang-Yao Lee**, Temple University, United States
Co-authors: Lexin Li, Bing Li

A new method is introduced to estimate directed acyclic graphs from multivariate functional data, based on the notion of faithfulness that relates a directed acyclic graph with a set of conditional independence relations among the random functions. To characterize and evaluate these relations, we develop two linear operators, the conditional covariance operator and the partial correlation operator. Based on these operators, we adapt and extend the PC-algorithm to estimate the functional directed graph, so that the computation time depends on the sparsity rather than the full size of the graph. We study the asymptotic properties of the two operators, derive their uniform convergence rates, and establish the uniform consistency of the estimated graph, all of which are obtained while allowing the graph size to diverge to infinity with the sample size. We demonstrate the efficacy of our method through both simulations and an application to a time-course proteomic dataset.

E0995: **A Bayesian subset specific approach to joint selection of multiple graphical models**
Presenter: **Kshitij Khare**, University of Florida, United States
Co-authors: Peyman Jalali, George Michailidis

The problem of joint estimation of multiple graphical models from high dimensional data has been studied in statistics and machine learning, due to its importance in diverse fields including molecular biology, neuroscience and the social sciences. A Bayesian approach is developed which decomposes the model parameters across the multiple graphical models into shared components across subsets of models and edges, and idiosyncratic ones. Further, it leverages a novel multivariate prior distribution, coupled with a jointly convex regression-based pseudo-likelihood that enables fast computations through a robust and efficient Gibbs sampling scheme. We establish strong posterior consistency for model selection under high dimensional scaling, with the number of variables growing exponentially as a function of the sample size.

EO075 Room Virtual R10 RECENT DEVELOPMENT IN FUNCTIONAL DATA ANALYSIS AND CLUSTER ANALYSIS Chair: Guanqun Cao

E0229: **Optimal classification for functional data using deep neural network**
Presenter: **Guanqun Cao**, Auburn University, United States

The optimal functional data classification problem is exploited via deep neural networks. A sharp non-asymptotic estimation error bound on the excess misclassification risk is established which achieves the minimax rates of convergence. In contrast to existing literature, the proposed deep neural network classifier is proven to achieve optimality without the knowledge of likelihood functions. This framework is further extended to accommodate general multi-dimensional functional data classification problems. We demonstrate the favourable finite sample performance of the proposed classifiers in various simulations and two real data applications, including the speech recognition data and the brain imaging data.

E0313: **SPF: A spatial and functional data analytic approach to cell imaging data**
Presenter: **Thao Vu**, University of Colorado, United States

The tumor microenvironment (TME), which characterizes the tumor and its surroundings, plays a critical role in understanding cancer development and progression. Recent advances in imaging techniques enable researchers to study the spatial structure of the TME at a single-cell level. Investigating spatial patterns and interactions of cell subtypes within the TME provides useful insights into how cells with different biological purposes behave, which may consequentially impact a subjects clinical outcomes. We utilize a class of well-known spatial summary statistics, the K-function and its variants, to explore inter-cell dependence as a function of distances between cells. Using techniques from functional data analysis, we introduce an approach to model the association between these summary spatial functions and subject-level outcomes, while controlling for other clinical scalar predictors such as age and disease stage. In particular, we leverage the additive functional Cox regression model (AFCM) to study the nonlinear impact of spatial interaction between tumor and stromal cells on overall survival in patients with non-small cell lung cancer, using multiplex immunohistochemistry (mIHC) data. The applicability of our approach is further validated using a publicly available Multiplexed Ion beam Imaging (MIBI) triple-negative breast cancer dataset.

E0426: **Multiple change point clustering of count processes**

Presenter: **Shuchismita Sarkar**, Bowling Green State University, United States

A model-based clustering algorithm relying on a finite mixture of negative binomial Levy processes is proposed. The algorithm models heterogeneous stochastic count process data and automatically estimates multiple change points upon fitting the mixture model. Such a change point estimation identifies time points when deviation from the standard process has occurred and serves as an important diagnostic tool for analyzing temporal data. The proposed model is applied to the COVID-positive ICU cases in the state of California with very interesting results.

E0723: **Modeling spiky functional data with derivatives of smooth functions in function-on-function regression**

Presenter: **Ruiyan Luo**, Georgia State University, United States

Smoothness penalties are efficient regularization and dimension reduction tools for functional regressions. However, for spiky functional data observed on a dense grid, the coefficient function in a functional regression can be spiky and, hence, the smoothness regularization is inefficient and leads to over-smoothing. We propose a novel approach to fit the function-on-function regression model by viewing the spiky coefficient functions as derivatives of smooth auxiliary functions. Compared with the smoothness regularization or sparsity regularization imposed directly on the spiky coefficient function in existing methods, imposing smoothness regularization on the smooth auxiliary functions can more efficiently reduce the dimension and improve the performance of the fitted model. Using the estimated smooth auxiliary functions and taking derivatives, we can fit the model and make predictions. Simulation studies and real-data applications show that compared with existing methods, the new method can greatly improve model performance when the coefficient function is spiky and performs similarly well when the coefficient function is smooth.

EO047 Room Virtual R11 DESIGN AND ANALYSIS OF COMPLEX EXPERIMENTS: THEORY AND APPLICATIONS Chair: MingHung Kao

E0288: **Optimal designs for generalized linear mixed models**

Presenter: **John Stufken**, University of North Carolina at Greensboro, United States

Co-authors: Yao Shi, Wanchunzi Yu

While generalized linear mixed models are useful, optimal design questions for such models are challenging due to the complexity of the information matrices. For longitudinal data, we propose an approximation of the information matrix based on the penalized quasi-likelihood method. We evaluate this approximation for logistic mixed models with time as the single predictor variable. Assuming that the experimenter controls at which time observations are to be made, the approximation is used to identify locally optimal designs based on commonly used optimality criteria. The method can also be used for random block effects models.

E0755: **Modeling and active learning for experiments with quantitative-sequence factors**

Presenter: **Abhyuday Mandal**, University of Georgia, United States

A new type of experiment which targets finding optimal quantities of a sequence of factors is drawing much attention in medical science, bioengineering and many other disciplines. Such studies require simultaneous optimization for both quantities and sequence orders of several components, which is defined as a new type of factors: quantitative-sequence (QS) factors. Due to the large and semi-discrete solution spaces in such experiments, it is non-trivial to efficiently identify the optimal (or near-optimal) solutions using only a few experimental trials. To address this challenge, we propose a novel active learning approach, named QS-learning, to enable effective modeling and efficient optimization for experiments with QS factors. The QS-learning consists of three parts: a novel mapping-based additive Gaussian process (MaGP) model, an efficient global optimization scheme (QS-EGO), and a new class of optimal designs (QS-design) for collecting initial data. Theoretical properties of the proposed method are investigated and techniques for optimization using analytical gradients are developed. The performance of the proposed method is demonstrated via a real drug experiment on lymphoma treatment and several simulation studies.

E0819: **Experimental design and active learning**

Presenter: **Rong Pan**, Arizona State University, United States

In machine learning or artificial intelligence, supervised learning methods such as classification and regression are so important that almost 80% ML/AI practice is about supervised learning. To perform supervised learning, one must have labeled data to build and train the learning model. However, labeling data are often expensive, while unlabeled data are cheap to obtain. Also, in specific tasks, not all available data are equally useful. The question is how to find the good, useful data to label them at minimal cost, while receiving maximum benefit, so as to learn the system more efficiently. Parallel to this notion, statistical experimental design is about deriving a strategy of selecting experimental conditions to conduct experiments such that the expected experimental results can best achieve the experimenter's objective. Therefore, both AL andDOE concern how to take samples from a population. We will present some recent developments of active learning (AL) in the ML/AI field and draw the connection of AL to traditional experimental design methodologies, particularly optimal design and sequential design. We will discuss how optimal design and sequential design theories can provide some theoretical enhancements to AL as well as practical improvements of AL algorithms.

E0824: **The summary of effect aliasing structure for supersaturated and factorial designs**

Presenter: **Frederick Kin Hing Phoa**, Academia Sinica, Taiwan

Co-authors: Yi-Hua Liao

In the assessment and selection of supersaturated designs, the aliasing structure of interaction effects is usually ignored in traditional criteria such as $E(s^2)$ optimality. We introduce the Summary of Effect Aliasing Structure (SEAS) for assessing the aliasing structure of supersaturated designs and other nonregular fractional factorial designs that takes account of interaction terms and provides detailed summaries such as (generalized) resolution and wordlength patterns. The new summary consists of three criteria, abbreviated as MAP: (1) the Maximum dependency aliasing pattern; (2) the Average square aliasing pattern; and (3) the Pairwise dependency ratio. These criteria provide insight when traditional criteria fail to differentiate between designs. We theoretically study the relationship between the MAP criteria and traditional quantities and demonstrate the use of SEAS for comparing some examples of supersaturated designs.

EO193 Room Virtual R12 RECENT METHODS IN FINANCE Chair: Soohun Kim

E0333: **The nature of ownership and stock returns**

Presenter: **Soohun Kim**, KAIST, Korea, South

Co-authors: Daniel Weagley, John Kim

Investment strategies vary in their reliance on characteristics versus intangible information to make investment decisions. We propose a simple machine learning methodology for estimating a strategy's reliance on characteristics or intangible information in making stock selection decisions. Using this methodology, we find stocks held by active mutual funds employing strategies more reliant on intangible information outperform stocks held by funds more reliant on characteristics. A long-short strategy based on the nature of stock ownership earns a CAPM-alpha of 4.8 pps per year (t-statistic of 3.54), and exhibits similar outperformance relative to other factor models. We find especially strong out-performance (19.4 pps per year, t-statistic > 9) within those stocks experiencing relative increases in mutual fund ownership breadth.

E0528: **Missing data in asset pricing panels**
Presenter: **Andreas Neuhierl**, Washington University in St. Louis, United States
Co-authors: Michael Weber, Joachim Freyberger, Bjoern Hoeppner

Missing data for return predictors is a common problem in cross-sectional asset pricing studies. Most papers do not explicitly discuss how they treat missing data but conventional treatments focus on complete cases for all predictors or impute the unconditional mean for the missing predictor. Both methods have undesirable properties - they are either inefficient or lead to biased estimators and incorrect inference. We propose a simple and computationally attractive alternative approach using conditional mean imputations and weighted least squares. This method allows us to use all sample points with observed returns, it results in valid inference, and it can be applied in non-linear and high-dimensional settings. We map our estimator into a GMM framework to study its relative efficiency and find that it performs almost as well as the efficient but computationally costly GMM estimator in many cases. We apply our procedure to a large panel of return predictors and find that it leads to improved out-of-sample predictability.

E0529: **Structural deep learning in conditional asset pricing**
Presenter: **Yuan Liao**, Rutgers University, United States
Co-authors: Andreas Neuhierl, Jianqing Fan, Tracy Ke

New structural nonparametric methods are developed for estimating conditional asset pricing models using deep neural networks, by employing time-varying conditional information on alphas and betas carried by firm-specific characteristics. Contrary to many applications of neural networks in economics, we can open the black box of machine learning predictions by incorporating financial economics theory into the learning, and provide an economic interpretation of the successful predictions obtained from neural networks, by decomposing the neural predictors as risk-related and mispricing components. The estimation method starts with period-by-period cross-sectional deep learning, followed by local PCAs to capture time-varying features such as latent factors of the model. We formally establish the asymptotic theory of the structural deep-learning estimators, which apply to both in-sample fit and out-of-sample predictions. We also illustrate the double-descent-risk phenomena associated with over-parametrized predictions, which justifies the use of over-fitting machine learning methods.

E0701: **Idiosyncratic volatility and the consistency of the ICAPM**
Presenter: **Gang Li**, The Chinese University of Hong Kong, Hong Kong
Co-authors: Bing Han

The average stock return idiosyncratic volatility is shown to contain useful information about the hedge portfolio under the ICAPM. In time series, two different weighted averages of individual stock idiosyncratic volatility together can significantly predict stock market returns over both short- and long-term horizons, both in-sample and out-of-sample. In cross-section, we propose a new method to estimate individual stock exposure to the unobserved hedge portfolio using aggregate idiosyncratic volatilities and find that the estimated beta is significantly related to the cross-section of expected stock returns. Finally, we show both theoretically and empirically that the return predictability of a previous tail index can be explained by idiosyncratic volatility under the ICAPM. Our results support the ICAPM pricing linkage between the time series and cross-section of stock returns.

EO231 Room Virtual R13 RECENT ADVANCES IN EVENT HISTORY STUDIES Chair: Jianguo Sun

E0226: **A new robust approach for regression analysis of panel count data with time-varying covariates**
Presenter: **Dayu Sun**, Emory University, United States

The focus is on robust inference for panel count data with time-varying covariates, which often occur in real-life applications such as medical studies and reliability experiments. Several mean model-based methods that are robust to within-subject correlation structures have been proposed and widely used. However, the robust mean model-based approach usually requires the mean function to be monotone, which may not be realistic when covariate values fluctuate over time. To address these issues, we propose a robust rate model-based procedure with rigorous theoretical justification. Under the rate model, another challenge is to develop variance estimators because the asymptotic variance usually has no closed forms. To our knowledge, the existing literature with similar problems all resorted to computationally intensive numerical methods that may not be robust. We develop novel computationally efficient robust variance estimators with closed forms based on Expectation-Maximization (EM) algorithm. We rigorously show that the variance estimators are consistent regardless of the underlying distribution assumption. An extensive simulation study is performed to demonstrate the superiority of the proposed approach and a real-life study of sexually transmitted infections is used to demonstrate the applicability of this newly proposed approach.

E0599: **A class of additive transformation models for recurrent gap times**
Presenter: **Ling Chen**, Washington University in St. Louis, United States
Co-authors: Yanqin Feng, Jianguo Sun

The gap time between recurrent events is often of primary interest in many fields such as medical studies. We discuss regression analysis of the gap times arising from a general class of additive transformation models. For the problem, we propose two estimation procedures, the modified within-cluster resampling(MWCR) method and the weighted risk-set (WRS) method, and the proposed estimators are shown to be consistent and asymptotically follow the normal distribution. In particular, the estimators have closed forms and can be easily determined, and the methods have the advantage of leaving the correlation among gap times arbitrary. A simulation study is conducted for assessing the finite sample performance of the presented methods and suggests that they work well in practical situations. Also, the methods are applied to a set of real data from a chronic granulomatous disease (CGD) clinical trial.

E1032: **High-dimensional variable selection for partially functional Cox regression with interval-censored data**
Presenter: **Yuanyuan Guo**, Duke University, United States
Co-authors: Tian Tian, Jianguo Sun

Variable selection is considered for interval-censored data with partially functional covariates and potential nonlinear effects. A flexible additive Cox model is proposed to incorporate the functional principal component analysis for modeling the functional predictors and Bernstein polynomials to approximate the nonlinear effects. We develop a penalized sieve maximum likelihood approach with an efficient group coordinate descent algorithm to allow for both low- and high-dimensional scenarios. The performance of the presented approach is assessed via a simulation study and the analysis of the Alzheimers Disease Neuroimaging Initiative (ADNI) data.

EO211 Room Virtual R2 META-ANALYSIS, NETWORK META-ANALYSIS AND IPD META-ANALYSIS Chair: Tiejun Tong

E0215: **A robust and computational-efficient method for multiple-outcome network meta-analysis**
Presenter: **Yong Chen**, Univ. of Pennsylvania, United States

In many biomedical settings, there is an increasing number of interventions available for a disease condition. It is critical for clinical decision-making to accurately evaluate and compare the relative efficacy and safety, as well as other patient centered outcomes of these interventions. We propose a network meta-analysis model for multiple clinical outcomes. Inspired by the idea of composite likelihood, the proposed method only requires the specification of the marginal distribution of each outcome, and a pseudolikelihood is then constructed under a working independence assumption. We also develop a novel inferential procedure with an associated efficient computational algorithm, which is statistically robust (i.e., requires minimal distributional assumptions) and computational stable and fast. We will illustrate our method through multiple case studies including a network meta-analysis of comparing 12 labor induction methods. The proposed composite likelihood-based multivariate network meta-analysis method leads to a computationally efficient algorithm with robust statistical inference, while being able to take multiple outcomes into consideration.

E0222: **A guide to estimating the reference range from a meta-analysis using aggregate or individual participant data**
Presenter: **Haitao Chu**, University of Minnesota School of Public Health, United States
Co-authors: Lianne Siegel, Hassan Murad, Richard Riley, Fateh Bazerbachi, Zhen Wang

Clinicians frequently must decide whether a patients measurement reflects that of a healthy normal individual. Thus, the reference range is defined as the interval in which some proportion (frequently 95%) of measurements from a healthy population is expected to fall. One can estimate it from a single study, or preferably from a meta-analysis of multiple studies to increase generalizability. This range differs from the confidence interval for the pooled mean or the prediction interval for a new study mean in a meta-analysis, which does not capture natural variation across healthy individuals. Methods for estimating the reference range from a meta-analysis of aggregate data that incorporate both within and between-study variations were recently proposed. We present three approaches for estimating the reference range: a frequentist, a Bayesian, and an empirical method. Each method can be applied to either aggregate or individual participant data (IPD) meta-analysis, with the latter being the gold standard when available. We illustrate these approaches using a clinical scenario about the normal range of a liver stiffness test.

E0639: **Bayesian estimation and testing in random effect meta-analysis of rare binary adverse events with flexible variability**
Presenter: **Johan Lim**, Seoul National University, Korea, South
Co-authors: Ming Zhang, Jackson Barth, Johan Lim, Xinlei Wang

Meta-analysis of rare binary events has become a routine in the pharmaceutical industry to access the safety of healthcare intervention since people can hardly make a quantitative and decisive conclusion from an individual study alone. Various frequentist or Bayesian methods have been proposed to attempt to report the accurate estimated treatment effect or the inter-study heterogeneity. However, almost all approaches pre-defined a direction of variance between the control and the treatment group, which might be more appropriate to be decided by data. Recently, a new flexible binomial-normal hierarchical model has been proposed by assuming no direction of variance. However, they mainly focus on comparing the current widely-used methods rather than proposing a new estimator. Therefore, we adopt a Bayesian hierarchical approach and develop our estimator (FlexB) and Bayesian hypothesis testing process using the flexible random-effects model, and we compare our method with existing frequentist and Bayesian competitors via extensive simulation. As for Bayesian calculation, we creatively incorporate the new Polya-Gamma data-augmentation technique into our sampling process, which brings some computational convenience and stability for estimation. Two data examples, updated rosiglitazone data and glutathione S-transferase P1(GSTP1) GG genotype data, are analyzed by our approach as well.

E0213: **A unified framework for meta-analysis with the five-number summary**
Presenter: **Jiandong Shi**, Hong Kong University of Science and Technology, Hong Kong

For clinical studies with continuous outcomes, if the data are skewed, researchers often report the whole or part of the five-number summary rather than the sample mean and standard deviation. Most existing methods for meta-analysis, however, cannot handle the normal and skewed data simultaneously. By incorporating the recent advances in data transformation, we develop a unified framework for meta-analysis when some studies are reported with the five-number summary. Specifically, we first develop a new testing method, using only the five-number summary, to check whether or not the underlying distribution of the data is skewed away from normality. If the skewness test is not rejected, we then apply the transformation methods to recover the sample mean and standard deviation from the five-number summary. Otherwise, it is suggested to either exclude the skewed studies from the meta-analysis for normal data, or apply a subgroup analysis that separates the normal and skewed studies.

EO213 Room Virtual R3 RECENT ADVANCES IN STATISTICAL MODELING AND COMPUTING FOR COMPLEX DATA Chair: Weixin Yao

E0518: **Two post-survey imaginary random mechanisms with implications**
Presenter: **Xiaogang Duan**, Beijing Normal University, Beijing, China, China

For Neyman's design-based inference, we introduce and compare ideas and implications of two virtual random mechanisms after actual sampling practice. One is named "virtual permutation". The other has been proposed recently and is named "imaginary census", for which a key element is the imaginary census matrix, recording the exact sampling trajectory of each draw without replacement until all population units were sampled out. Both provide an interesting supplement to Neyman's conventional framework. In particular, both provide insight into the relationship of sampling with and without replacement.

E0865: **Adversarially robust subspace learning in the spiked covariance model**
Presenter: **Ruizhi Zhang**, University of Nebraska-Lincoln, United States
Co-authors: Fei Sha

The problem of robust subspace learning is studied when there is an adversary who can attack the data to increase the projection error. By deriving the adversarial projection risk when data follows the multivariate Gaussian distribution with the spiked covariance or so-called Spiked Covariance model, we propose to use the empirical risk minimization method to obtain the optimal robust subspace. We then find a non-asymptotic upper bound of the adversarial excess risk, which implies the empirical risk minimization estimator is close to the optimal robust adversarial subspace. The optimization problem can be solved easily by the projected gradient descent algorithm for the rank-one spiked covariance model. However, in general, it is computationally intractable to solve the empirical risk minimization problem. Thus, we propose to minimize the upper bound of the empirical risk to find the robust subspace for the general spiked covariance model.

E0912: **Wasserstein regression**
Presenter: **Yaqing Chen**, University of California, Davis, United States
Co-authors: Zhenhua Lin, Hans-Georg Mueller

The analysis of samples of random objects that do not lie in a vector space has found increasing attention in statistics in recent years. An important class of such object data is univariate probability distributions. Adopting the Wasserstein geometry, we develop a class of regression models, for which the predictor and response are both random distributions. The proposed distribution-to-distribution regression model provides an extension of multiple linear regression for Euclidean data and function-to-function regression for Hilbert space valued data in functional data analysis. We derive asymptotic rates of convergence for the estimates of the regression operator and illustrate the proposed method with human mortality data. We also consider an extension to autoregressive modeling of distributional time series and a nonparametric approach when predictors are scalars and responses are distributions.

E0916: **Exact permutation/randomization tests algorithms**
Presenter: **Subir Ghosh**, University of California, United States

R.A. Fisher described the exact permutation and randomization tests for comparative experiments without assuming normality or probability distribution. While having this as an attractive feature, the computational challenge was a disadvantage at that time but not now with modern computers. The aim is to introduce a permutation/randomization data algorithm to generate the permutation/randomization distributions under the null hypotheses for calculating the P-values. The properties of permutation/randomization data matrices developed by algorithms following the proposed mathematical processes are derived. Two illustrative examples demonstrate the usefulness of the proposed computational methods.

EO355 Room Virtual R4 NEW DIRECTIONS IN HIGH-DIMENSIONAL AND FUNCTIONAL DATA ANALYSIS Chair: Alexander Petersen

E0509: **Functional data analysis for longitudinal data with informative observation times**
Presenter: **Luo Xiao**, North Carolina State University, United States

In functional data analysis for longitudinal data, the observation process is typically assumed to be noninformative, which is often violated in real applications. Thus, methods that fail to account for the dependence between observation times and longitudinal outcomes may result in biased estimation. For longitudinal data with informative observation times, we find that under a general class of shared random effect models, a commonly used functional data method may lead to inconsistent model estimation while another functional data method results in consistent and even rate optimal estimation. Indeed, we show that the mean function can be estimated appropriately via penalized splines and that the covariance function can be estimated appropriately via penalized tensor product splines, both with specific choices of parameters. For the proposed method, theoretical results are provided, and simulation studies and a real data analysis are conducted to demonstrate its performance.

E0890: **Regression modeling for distributional response data**
Presenter: **Alexander Petersen**, Brigham Young University, United States
Co-authors: Wendy Meiring, Xi Liu, Aritra Ghosal

Data consisting of samples of probability density functions are increasingly prevalent, necessitating the development of methodologies for their analysis that respect the inherent nonlinearities associated with densities. In many applications, density curves appear as functional response objects in a regression model with vector predictors. We consider two such models in which the regression function takes the form of conditional Fréchet means under the Wasserstein geometry of optimal transport. The first model, known as global Fréchet regression, is developed as a generalization of multiple linear regression, for which we demonstrate the use of hypothesis testing of global and partial effects, as well as simultaneous confidence bands for estimated conditional mean densities. In the second, greater flexibility in the predictor-response relationship is achieved by a generalization of single-index models, fitted by local Fréchet regression techniques. These methods are illustrated through regression analyses of post-intracerebral hemorrhage hematoma densities and distributions of age-at-death for various countries.

E0686: **Low-rank latent matrix factor-analysis modeling for generalized linear regression with imaging biomarkers**
Presenter: **Catherine Liu**, The Hong Kong Polytechnic University, Hong Kong

Medical imaging has been recognized as a phenotype associated with various clinical traits in diagnostics and prognosis of clinical trials and cancer studies. Motivated by the cutting-edge matrix factor analysis modeling, we propose a new latent matrix factor generalized regression tool named FamGLM, which relates a scalar treatment outcome with predictors including imaging variate. The FamGLM enjoys high prediction capability since the extracted matrix factor score refines the structural effect of the matrix-valued predictor and circumvents over-dimension reduction. Inspired by 2nd-order tensor principal component analysis, we develop a matrix SVD-based estimation procedure and algorithm through generalized low-rank approximation of matrices, which has a much lower computational cost compared with existing statistical approaches. The proposed FamGLM also achieves higher prediction capability than existing methods. In numerical analysis, we evaluate the finite sample performance of FamGLM in classification and prediction compared with existing statistical approaches under various GLM scenarios. The FamGLM outperforms in discriminant power in the analysis of a COVID-CT image data set.

E0958: **Factor-augmented smoothing model for functional data**
Presenter: **Yanrong Yang**, The Australian National University, Australia
Co-authors: Han Lin Shang, Yuan Gao

Modeling raw functional data as a mixture of a smooth function and a high-dimensional factor component is proposed. The conventional approach to retrieving the smooth function from the raw data is through various smoothing techniques. However, the smoothing model is not adequate to recover the smooth curve or capture the data variation in some situations. These include cases where there is a large amount of measurement error, the smoothing basis functions are incorrectly identified, or the step jumps in the functional mean levels are neglected. To address these challenges, a factor-augmented smoothing model is proposed, and an iterative numerical estimation approach is implemented in practice. Including the factor model component in the proposed method solves the aforementioned problems since a few common factors often drive the variation that cannot be captured by the smoothing model. Asymptotic theorems are also established to demonstrate the effects of including factor structures on the smoothing results. As a byproduct of independent interest, an estimator for the population covariance matrix of the raw data is presented based on the proposed model. Extensive simulation studies illustrate that these factor adjustments are essential in improving estimation accuracy and avoiding the curse of dimensionality. The superiority of our model is also shown in modeling Canadian weather data and Australian temperature data.

EO027 Room Virtual R5 RECENT ADVANCES IN MATRIX AND TENSOR LEARNING Chair: Kejun He

E0206: **Two-level monotonic multistage recommender systems**
Presenter: **Ben Dai**, The Chinese University of Hong Kong, China
Co-authors: Xiaotong Shen, Wei Pan

A recommender system learns to predict the user-specific preference over items, making personalized recommendations based on a relatively small number of observations. One challenging issue is how to leverage three-way interactions, referred to as user-item-stage dependencies on a monotonic chain of events, to enhance the prediction accuracy. A monotonic chain of events occurs, for instance, in an article sharing dataset, where a “follow” action implies a “like” action, which in turn implies a “view” action. We develop a multistage recommender system utilizing a two-level monotonic property for personalized prediction. Particularly, we derive a large-margin classifier based on a nonnegative additive latent factor model, reducing the number of model parameters for personalized prediction while guaranteeing prediction consistency. On this ground, we derive a regularized cost function to learn user-specific behaviors at different stages, linking decision functions to numerical and categorical covariates to model user-item-stage interactions. Computationally, we derive an algorithm based on blockwise coordinate descent. Theoretically, we show that the two-level monotonic property enhances the accuracy of learning as compared to a standard method treating each stage individually and an ordinal method utilizing only one-level monotonicity. Finally, the proposed method compares favorably with existing methods in simulations and an article sharing dataset.

E0588: **Correlation tensor decomposition and its application in spatial imaging data**
Presenter: **Xiwei Tang**, University of Virginia, United States

Multi-dimensional tensor data have gained increasing attention in recent years, especially in biomedical imaging analyses. However, most existing tensor models are only based on the mean information of imaging pixels. Motivated by multimodal optical imaging data in a breast cancer study, we develop a new tensor learning approach to use pixel-wise correlation information, which is represented through the higher-order correlation tensor. We proposed a novel semi-symmetric correlation tensor decomposition method that effectively captures the informative spatial patterns of pixel-wise correlations to facilitate cancer diagnosis. We establish the theoretical properties for recovering structure and for classification consistency. In addition, we develop an efficient algorithm to achieve computational scalability. Our simulation studies and an application on breast cancer imaging data all indicate that the proposed method outperforms other competing methods in terms of pattern recognition and prediction accuracy.

E0593: **Exact clustering in tensor block model: Statistical optimality and computational limit**
Presenter: **Rungang Han**, Duke University, United States
Co-authors: Yuetian Luo, Miaoyan Wang, Anru Zhang

High-order clustering aims to identify heterogeneous substructures in multiway datasets that arise commonly in neuroimaging, genomics, social network studies, etc. The non-convex and discontinuous nature of this problem pose significant challenges in both statistics and computation. We propose a tensor block model and the computationally efficient methods, high-order Lloyd algorithm (HLloyd), and high-order spectral clustering (HSC), for high-order clustering. The convergence guarantees and statistical optimality are established for the proposed procedure under a mild sub-Gaussian noise assumption. Under the Gaussian tensor block model, we completely characterize the statistical-computational trade-off for achieving high-order exact clustering based on three different signal-to-noise ratio regimes. The analysis relies on new techniques of high-order spectral perturbation analysis and a "singular-value-gap-free" error bound in tensor estimation, which are substantially different from the matrix spectral analyses in the literature. Finally, we show the merits of the proposed procedures via extensive experiments on both synthetic and real datasets.

E0665: **Matrix completion with model-free weighting**
Presenter: **Raymond Ka Wai Wong**, Texas AM University, United States
Co-authors: Jiayi Wang, 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 the 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.

EO415 **Room Virtual R6** **GAUSSIAN PROCESS REGRESSION MODELS** **Chair: Xia Wang**

E0355: **Bayesian variable selection in double generalized linear Tweedie spatial process models**
Presenter: **Aritra Halder**, University of Virginia, United States
Co-authors: Shariq Mohammed, Dipak Dey

Double generalized linear models provide a flexible framework for modeling data by allowing the mean and the dispersion to vary across observations. Common members of the exponential dispersion family including Gaussian, compound Poisson-gamma, Gamma, and inverse-Gaussian, are known to admit such models. However, the lack of their use can be attributed to ambiguities that exist in the model specification under a large number of covariates and, complications that arise when data from a chosen application displays dependence. We consider a hierarchical specification for these models with a spatial random effect. The spatial effect is targeted at performing uncertainty quantification by modeling dependence within the data arising from location-based indexing of the response. We focus on a Gaussian process specification for the spatial effect. Simultaneously we tackle the problem of the model specification under such hierarchical spatial process models using Bayesian variable selection, which is effected through a continuous spike and slab prior on the model parameters (or fixed effects). The novelty lies in the Bayesian frameworks developed for such models which have not been explored previously. We perform various synthetic experiments to showcase the accuracy of our frameworks. These developed frameworks are then applied to analyze automobile insurance claims.

E0607: **Flexible link functions in a joint hierarchical Gaussian process model**
Presenter: **Xia Wang**, University of Cincinnati, United States

Many longitudinal studies often require jointly modeling a biomarker and an event outcome, in order to provide more accurate inference and dynamic prediction of disease progression. Cystic fibrosis (CF) studies have illustrated the benefits of these models, primarily examining the joint evolution of lung-function decline and survival. We propose a novel joint model within the shared parameter framework that accommodates non-linear lung-function trajectories, in order to provide more accurate inference on the lung-function decline over time and to examine the association between the evolution of lung function and the risk of a pulmonary exacerbation event recurrence. Specifically, a two-level Gaussian process is used to estimate the nonlinear longitudinal trajectories and a flexible link function is introduced for a more accurate depiction of the binary process on the event outcome. Bayesian model assessment is used to evaluate each component of the joint model in simulation studies and an application to longitudinal data on patients receiving care from a CF center. A nonlinear structure is suggested by both the longitudinal continuous and binary evaluations. Including a flexible link function improves model fit to these data.

E0715: **Airflow recovery using synchrosqueezing transform and locally stationary Gaussian process regression**
Presenter: **Yu-Bo Wang**, Clemson University, United States
Co-authors: Whitney Huang, Yu-Min Chung, Jeff Mandel, Hau-Tieng Wu

A wealth of information about the respiratory system is encoded in the airflow signal. While direct measurement of airflow via spirometer with an occlusive seal is the gold standard, this may not be practical for ambulatory monitoring of patients. Advances in sensor technology have made measurement of the motion of the thorax and abdomen feasible with small inexpensive devices, but estimating airflow from these time series is challenging due to the presence of complicated nonstationary oscillatory signals. To properly extract the relevant oscillatory features from thoracic and abdominal movement, a nonlinear-type time-frequency analysis tool, the synchrosqueezing transform, is employed; these features are then used to estimate the airflow by a locally stationary Gaussian process regression. It is shown that using a dataset that contains respiratory signals under normal sleep conditions, accurate airflow out-of-sample predictions, and hence the precise estimation of an important physiological quantity, inspiration respiration ratio, can be achieved by fitting the proposed model both in the intra- and inter-subject setups. The method is also applied to a more challenging case, where subjects under general anaesthesia underwent transitions from pressure support to unassisted ventilation to further demonstrate the utility of the proposed method.

E0746: **Joint modeling with integrated fractional Brownian motion**
Presenter: **Seongho Song**, University of Cincinnati, United States
Co-authors: Anushka Palipana, Rhonda Szczesniak , Nishant Gupta

Biomarker data are often used to understand a disease's progression over time and characterize the relationship between biomarker data and the event outcome simultaneously using joint models. Motivated by being unable to effectively capture a biological process' variations using conventional random effects longitudinal sub-model, we propose a five-component longitudinal sub-model for a joint model. The most novel development is the scaled integrated fractional Brownian motion (IFBM) which has shown to reasonably depict biological processes measured with error. Other model components are the random intercept, fixed effects, and measurement error. Cox proportional hazards model serves as the event sub-model, which includes a time-dependent true longitudinal trajectory, and a set of baseline covariates. We use Markov chain Monte Carlo (MCMC) methods for Bayesian posterior computation and inference. We perform a simulation study and a comparative study of our joint model with IOU process from literature, and a joint model without a stochastic process. Our novel approach is then applied to the National Heart, Lung, and Blood Institute (NHLBI) lymphangioleiomyomatosis (LAM) registry data set and the Cystic Fibrosis (CF) data from 2 selected CF centers recorded in the US CF foundation patient registry (CFF-PR). We use forced expiratory volume in one second (FEV1) in liters and FEV1 pct-predicted as longitudinal biomarkers in LAM, and CF applications respectively.

EO371 Room Virtual R7 ADVANCES IN STATISTICAL METHODS FOR HANDLING COMPLEX DATA Chair: MinJae Lee

E0613: **Data augmentation using aggregate statistics from big data and survey: 2nd order delta-method and bootstrap inference**
Presenter: **Ryung Kim**, Albert Einstein College of Medicine, United States

It is often useful to analyze large but potentially biased big data jointly with smaller gold-standard surveys. Health surveys with higher quality and standardized measurements can benefit from the augmentation of electronic health records (EHR) originally collected for administrative and billing purposes. We recently showed the efficiency of an estimator that pools aggregate statistics from two sources. However, it remains unknown how to perform statistical inference based on the estimator. We develop two methods for statistical inference based on the Mosteller estimator that employ correction of bias and skewness: the second-order delta method, and a modified version of the biased-corrected and accelerated bootstrap approach. The methods are based on aggregated statistics obtained from two sources, one of which is potentially biased. In the numerical study, these methods provide valid coverage rates while the nave plug-in method and the first order delta method do not. Finally, the methods are demonstrated with two databases in South Korea: the Korea National Health and Nutrition Examination Survey and the National Health Insurance Service Sample Cohort. The prevalence of uncontrolled diabetes in the senior population was estimated typically be lower in the EHR database of health examinations compared to the gold-standard health survey. The proposed confidence intervals almost always were shorter than the interval solely based on the health survey.

E0571: **Comparing and combining data from immune assays with different limits of detection**
Presenter: **Ying Huang**, Fred Hutchinson Cancer Research Center, United States

In vaccine research towards the prevention of infectious diseases, immune response biomarkers serve as an important tool for comparing and ranking vaccine candidates based on their immunogenicity and predicted protective effect. However, analyses of immune response outcomes can be complicated by differences across assays when immune response data are acquired from multiple groups/laboratories. Motivated by a real-world problem to accommodate the use of two different neutralization assays in COVID-19 vaccine trials, we propose methods that integrate external paired-sample data with bridging assumptions to achieve two objectives, both using pooled data acquired from different assays: i) comparing immunogenicity between vaccine regimens, and ii) evaluating correlates of risk. Our methods adjust for differences between assays with respect to measurement error and the lower limit of detection. Simulation studies were conducted to demonstrate the satisfactory performance of the proposed methods and their advantage over alternative approaches. We apply the proposed methods to SARS-CoV-2 spike-pseudotyped virus neutralization assay data generated in vaccine and convalescent samples by two different laboratories.

E0745: **Testing calibration of risk models at extremes of disease risk**
Presenter: **Minsun Song**, Sookmyung Women's University, Korea, South

Risk-prediction models need careful calibration to ensure they produce unbiased estimates of risk for subjects in the underlying population given their risk-factor profiles. As subjects with extreme high or low risk may be the most affected by knowledge of their risk estimates, checking the adequacy of risk models at the extremes of risk is very important for clinical applications. We propose a new approach to test model calibration targeted toward extremes of disease risk distribution where standard goodness-of-fit tests may lack power due to the sparseness of data. We construct a test statistic based on model residuals summed over only those individuals who pass high and/or low-risk thresholds and then maximize the test statistic over different risk thresholds. We derive an asymptotic distribution for the max-test statistic based on the analytic derivation of the variance-covariance function of the underlying Gaussian process. The method is applied to a large case-control study of breast cancer to examine the joint effects of common single nucleotide polymorphisms (SNPs) discovered through recent genome-wide association studies. The analysis clearly indicates a non-additive effect of the SNPs on the scale of absolute risk, but an excellent fit for the linear-logistic model even at the extremes of risks.

E0740: **Transporting randomized trial results to estimate counterfactual survival functions in target populations**
Presenter: **Youngjoo Cho**, Konkuk University, Korea, South
Co-authors: Zhiqiang Cao, Fan Li

Generalizability and transportability have been studied extensively for uncensored data. However, scarce literature focuses on survival data with censoring. Motivated by controversial results from two clinical trials of blood pressure, we study the transportability of survival outcome findings from randomized clinical trials to an external target population. Based on four assumptions, we propose inverse probability weighting estimators and doubly robust estimators and show that when both the sampling score model and censoring model are correctly specified, the proposed estimators are consistent. Furthermore, the doubly robust estimators are still consistent if the survival time model is correct no matter sampling score model and censoring model are misspecified or not. We derive the influence functions of the proposed estimators and conduct simulation studies to examine their finite-sample performances. We finally apply our proposed estimators to assess the transportability of survival difference between treatment and control groups found in ACCORD-BP trial to the adults with Diabetes mellitus in the U.S. population.

EO201 Room Virtual R8 STATISTICAL MODELS & MACHINE LEARNING FOR OFFICIAL STATISTICS AND SURVEYS Chair: Scott Holan

E0223: **Tackling the overabundance of options in survey estimation**
Presenter: **Kelly McConville**, Harvard University, United States

With the increased availability of non-survey data and the plethora of exciting, new estimators, survey practitioners may start feeling paralyzed by all these choices. How should we combine the survey and non-survey data? Should we use a model- or design-based approach? Bayesian or frequentist? Do we borrow strength outside the domain or leverage more auxiliary data within the domain? If borrowing from outside, what similar domains should we utilize? We will tackle these questions using examples from the US Forest Inventory and Analysis Program, and, in a true statistical fashion, we will find that the answers vary depending on the context.

E0279: **On the use of auxiliary variables in multilevel regression and poststratification**
Presenter: **Yajuan Si**, University of Michigan, United States

Multilevel regression and poststratification (MRP) have become a popular approach for selection bias adjustment in subgroup estimation, with widespread applications from social sciences to public health. We examine the statistical properties of MRP in connection with poststratification and hierarchical models. The success of MRP prominently depends on the availability of auxiliary information strongly related to the outcome. We present a framework for statistical data integration and robust inferences of probability and nonprobability surveys, providing solutions to various challenges in practical applications. The simulation studies indicate the statistical validity of MRP with a tradeoff between bias and variance, and the improvement over alternative methods is mainly on subgroup estimates with small sample sizes. Our development is motivated by the Adolescent Brain Cognitive Development (ABCD) Study that has collected children across 21 U.S. geographic locations for national representation but is subject to selection bias as a nonprobability sample. We apply the methods for population inferences to evaluate cognition performances of diverse groups of children in the ABCD study and demonstrate that the use of auxiliary variables affects the inferential findings.

E0277: **Design consistent Bayesian tree models**
Presenter: **Daniell Toth**, US Bureau of Labor Statistics, United States
Co-authors: Scott Holan, Diya Bhaduri

Tree models provide a method for analyzing survey data because of the easy way they can handle a large number of variables with many interactions often found in this type of data. However, until recently design consistent tree modeling algorithms have not been available for use on data collected from a complex sample design. Design consistent algorithms are very desirable due to the many potential applications of these methods to survey data. As these applications have become more complex, interest in modeling the conditional distribution at each node using more sophisticated models has grown. Bayesian tree modeling approaches with a prior distribution on the set of all possible tree models and then selecting the optimal model using a stochastic search have been developed for independent data, but there are no methods for incorporating survey weights to produce design consistent models. Since the Bayesian framework allows for easily incorporating more complex models, we propose extending the Bayesian tree algorithm research to obtain a design consistent Bayesian tree model. The methods are illustrated through empirical simulation and an application to the Consumer Expenditure Survey.

E0287: **Computationally efficient Bayesian unit-level models for non-Gaussian data under informative sampling**
Presenter: **Scott Holan**, University of Missouri, United States
Co-authors: Ryan Janicki, Paul Parker

Statistical estimates from survey samples have traditionally been obtained via design-based estimators. In many cases, these estimators tend to work well for quantities such as population totals or means, but can fall short as sample sizes become small. In today's information age, there is a strong demand for more granular estimates. To meet this demand, using a Bayesian pseudo-likelihood, we propose a computationally efficient unit-level modeling approach for non-Gaussian data collected under informative sampling designs. Specifically, we focus on binary and multinomial data. Our approach is both multivariate and multiscale, incorporating spatial dependence at the area-level. We illustrate our approach through an empirical simulation study and through a motivating application to health insurance estimates using the American Community Survey.

EO063 Room Virtual R9 MODERN STATISTICAL METHODS IN DATA SCIENCE Chair: Yichuan Zhao

E0661: **Confidence intervals of mean residual life function in length-biased sampling based on modified empirical likelihood**
Presenter: **Wei Ning**, Bowling Green State University, United States
Co-authors: Suthakaran Ratnasingam

The mean residual life (MRL) function is one of the basic parameters of interest in survival analysis. We develop three procedures based on modified versions of empirical likelihood (EL) to construct confidence intervals of the MRL function with length-biased data. The asymptotic results corresponding to the procedures have been established. The proposed methods exhibit better finite sample performance over other existing procedures, especially in small sample sizes. Simulations are conducted to compare coverage probabilities and the average lengths of confidence intervals under different scenarios for the proposed methods and some existing methods. Two real data applications are provided to illustrate the methods of constructing confidence intervals.

E0855: **Optimal subsampling in a massive data linear regression**
Presenter: **Fei Tan**, Indiana University-Purdue University Indianapolis, United States
Co-authors: Hanxiang Peng

To fast approximate the least-squares estimate efficiently in a massive data linear regression by a subsampling estimate, we give numerous optimal sampling distributions based on the criteria of minimum bias and maximum information. We show that the statistical leverage scores-based distribution minimizes the bias, the A-optimal distribution minimizes the trace norm of the covariance matrix, and a distribution with the likelihood ratio to the uniform bounded away from zero attains the optimal convergence rate. We exhibit the necessity of truncating the sampling distributions, provide relative error bounds, and discuss subsample size determination. We construct an algorithm with a running time $o(n \times p^2)$, and report a large simulation study and a massive real data application.

E1039: **Robust method for optimal treatment decision making based on survival data**
Presenter: **Min Zhang**, University of Michigan, United States

Methods are developed for estimating the optimal treatment decision rule based on data with survival time as the primary endpoint. The methods are based on a flexible semiparametric accelerated failure time model, where only the treatment contrast (i.e., the difference in means between treatments) is parameterized and all other aspects are unspecified. An individual's treatment contrast is firstly estimated robustly by an augmented inverse probability weighted estimator (AIPWE). Then, the optimal decision rule is estimated by minimizing the loss between the treatment contrast and the AIPWE contrast. Two loss functions with different strategies to account for censoring are proposed. The proposed loss functions distinguish from existing ones in that they are based on treatment contrasts, which completely determine the optimal treatment rule. Our methods can further incorporate a penalty term to select variables that are only important for treatment decision-making, while taking advantage of all covariates predictive of outcomes to improve performance. Comprehensive simulation studies have been conducted to evaluate the performances of the proposed methods relative to existing methods. The proposed methods are illustrated with an application to the ACTG 175 clinical trialon HIV-infected patients.

E0765: **Novel empirical likelihood inference for the mean difference with right-censored data**

Presenter: **Yichuan Zhao**, Georgia State University, United States

Co-authors: Kangni Alemdjrodo

The focus is on comparing two means and finding a confidence interval for the difference of two means with right-censored data using the empirical likelihood method combined with the i.i.d. random functions representation. In the literature, some early researchers proposed empirical likelihood-based confidence intervals for the mean difference based on right-censored data using the synthetic data approach. We propose an empirical likelihood method based on the i.i.d. representation of Kaplan-Meier weights involved in the empirical likelihood ratio. We obtain the standard chi-squared distribution. We also apply the adjusted empirical likelihood to improve coverage accuracy for small samples. In addition, we investigate a new empirical likelihood method, the mean empirical likelihood, within the framework of our study. The performances of all the empirical likelihood methods are compared via extensive simulations. The proposed empirical likelihood-based confidence interval has better coverage accuracy than those from existing methods. Finally, our findings are illustrated with a real data set.

Monday 06.06.2022 13:30 - 14:45 Parallel Session N – EcoSta2022

EO301 Room Live Theater (Hybrid 2) VARIATIONAL INFERENCE IN STATISTICS AND ECONOMETRICS II Chair: Pierre Alquier

E0536: **Variational Bayes for models with nuisance parameters**
Presenter: **Minh-Ngoc Tran**, University of Sydney, Australia
Co-authors: Paco Tseng, Robert Kohn

Statistical models with nuisance parameters are ubiquitous in statistics and its related fields. Elimination of nuisance parameters, for reliable inference about the main parameters of interest, is an important problem for statistical inference in those models. We revisit this problem and present a range of Variational Bayes approaches for eliminating nuisance parameters and efficient inference about the main parameters.

E0537: **Inference in stochastic volatility models with variational sequential Monte Carlo**
Presenter: **Yuliya Shapovalova**, Radboud University, Netherlands

Applications of stochastic volatility models, in particular in the multivariate case, are limited due to high computational time requirements by state-of-the-art methods such as particle Markov chain Monte Carlo methods. When the number of latent states in stochastic volatility models is large, inference with this class of methods becomes practically infeasible. Variational methods in recent years have shown great potential in large-scale applications. However, those relying on linearization techniques in the case of stochastic volatility models may still result in large biases even in the univariate case. We consider a recently published work on a new approximating family of distributions, the variational sequential Monte Carlo (VSMC), with application to stochastic volatility models. The advantage of this new inference method is in its flexibility: the posterior can be approximated arbitrarily well while maintaining efficient optimization of the parameters. Using a case study of stochastic volatility models, we evaluate the potential of VSMC in terms of scalability and precision and place it in the literature in comparison to other methods varying from particle MCMC to INLA.

E1002: **Bayesian bilinear neural network for predicting the mid-price dynamics in limit-order book markets**
Presenter: **Martin Magris**, Aarhus University, Denmark
Co-authors: Alexandros Iosifidis, Mostafa Shabani

Recent advances in Variational Bayes (VB) for deep learning led to fast and efficient methods for scalable Bayesian inference applicable to complex modelling and predictive tasks, yet their use in econometric modelling is not widespread. In modern financial markets, traditional time-series forecasting models are often incapable of capturing the underlying complexity and interacting nature of the patterns driving price dynamics. On the other hand, data-driven Machine Learning (ML) methods have been proved effective. However traditional ML approaches are incapable of addressing parameters uncertainty and confidence interval for the predictions. Bayesian ML methods provide a natural remedy to bring together the predictive ability of ML methods and the probabilistic dimension typical of econometric research. Within the VB framework, we train a Bayesian bilinear network with temporal attention by adopting a state-of-the-art Bayesian optimizer to address the challenging prediction of mid-price movements in high-frequency limit-order book markets. Our results underline the feasibility of Bayesian methods in complex econometric modelling and their predictive and interpretative gains over several ML alternatives. We address the use of the Bayesian framework to analyse errors and uncertainties associated with forecasts and deliver insights for actionable trading decisions.

EO387 Room Main Theater (Hybrid 1) DISCRIMINATION AND PRINCIPAL COMPONENT ANALYSIS OF SIGNALS Chair: Yan Liu

E0243: **Discriminant analysis based on binary time series**
Presenter: **Yuichi Goto**, Kyushu University, Japan
Co-authors: Masanobu Taniguchi

Binary time series is the time series taking values 0 and 1. We discuss discriminant analysis and propose a new classification method based on binary time series. First, we show that the misclassification probability tends to zero when the number of observations tends to infinity. Next, we evaluate the asymptotic misclassification probability when two categories are contiguous. Finally, we show that our classification method based on binary time series has good robustness when the process is contaminated by an outlier, that is, our classification method is insensitive to the outlier. However, the classical method based on smoothed periodogram is sensitive to the outlier.

E0250: **Sparse principal component analysis for high-dimensional stationary time series**
Presenter: **Kou Fujimori**, Shinshu University, Japan
Co-authors: Yuichi Goto, Yan Liu

The sparse principal component analysis for high-dimensional stationary processes is discussed. The standard principal component analysis performs poorly when the dimension of the process is large. We establish the oracle inequalities for penalized principal component estimators for the processes including heavy-tailed time series. The rate of convergence of the estimators is established. We also elucidate the theoretical rate for choosing the tuning parameter in penalized estimators. The performance of the sparse principal component analysis is demonstrated by numerical simulations.

E0768: **Functional threshold autoregressive model**
Presenter: **Kun Chen**, Southwestern University of Finance and Economics, China
Co-authors: Chun Yip Yau, Yuanbo Li

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 according to different values of a threshold variable at a past time point. Sufficient conditions for strict stationarity and ergodicity of the functional threshold autoregressive process are investigated. A novel criterion-based method is developed to simultaneously conduct dimension reduction and estimation of thresholds, autoregressive orders, and model parameters. The consistency and asymptotic distributions of the estimators of both thresholds and underlying autoregressive models are established. Simulation studies and a real application of U.S. Treasury zero-coupon yield rates are provided to illustrate the effectiveness and usefulness of the proposed methodology.

EO057 Room Virtual R1 INNOVATIVE APPROACHES IN ORDINAL AND MIXED-TYPE DATA MODELLING Chair: Cristina Mollica

E0573: **A mixture model for ordinal variables measured on semantic differential scales**
Presenter: **Marica Manisera**, Universita' degli Studi di Brescia, Italy
Co-authors: Paola Zuccolotto

In surveys aimed at measuring subjective perceptions toward latent traits, questions asking to rate opinions on ordered response scales are usually exploited. The most popular ordered response scale is the Likert-type, but semantic differential scales (SMD) are also commonly used. In SMD, the respondent is asked to rate his/her position between two opposite adjectives. We present the CUM model, a mixture of a linearly transformed Multinomial and a Uniform random variable, suited to fit rating data expressed on SMD. The formulation of the model, belonging to the CUB class, is derived from specific hypotheses about the decision process in the mind of the respondent, who is assumed to start their reasoning from the center of the SMD and move upward/downward according to two specific feeling parameters, measuring the attitudes toward the two opposite adjectives. A specific representation of the parameter space is proposed with a triangular plot. Parameter estimation is carried out via the EM algorithm and a case study is examined, also with the comparison to the results obtained using other models of the CUB class.

E0615: **Convex clustering of mixed numerical and categorical data**
Presenter: **Carlo Cavicchia**, Erasmus University Rotterdam, Netherlands

Clustering analysis is an unsupervised learning technique widely used for information extraction. Current clustering algorithms often face instabilities due to the non-convex nature of their objective function. The class of convex clustering methods does not suffer from such instabilities and finds a global optimum for the clustering objective. Whereas convex clustering has previously been established for single-type data, real-life data sets usually comprise both numerical and categorical, or mixed, data. Therefore, we introduce the mixed data convex clustering (MIDACC) framework. We implement this framework by developing a dedicated subgradient descent algorithm. Through numerical experiments, we show that, in contrast to baseline methods, MIDACC achieves near-perfect recovery of both spherical and non-spherical clusters, is able to capture information from mixed data while distinguishing signal from noise, and has the ability to recover the true number of clusters present in the data. Furthermore, MIDACC outperforms all baseline methods on a real-life data set.

E0936: **Energy trees: Regression and classification with structured and mixed-type covariates**
Presenter: **Riccardo Giubilei**, Luiss Guido Carli, Italy
Co-authors: Tullia Padellini, Pierpaolo Brutti

The continuous growth of data complexity requires methods and models that adequately account for non-trivial structures, as any simplification may induce a loss of information. Many analytical tools have been introduced to work with complex data objects in their original form, but they can typically deal with single-type variables only. We introduce Energy Trees as a model for regression and classification where covariates are potentially both structured and of different types. Energy Trees incorporate Energy Statistics to generalize Conditional Trees, from which they inherit statistically sound foundations, interpretability, scale invariance, and lack of distributional assumptions. We consider the cases of functions, graphs, and persistence diagrams as structured covariates, besides showing that the model can be easily adapted to work with almost any other type of variable. Finally, we employ an extensive simulation study and some empirical analyses with human biological data to confirm the desirable properties and the predictive ability of our proposal.

EO199 Room Virtual R2 ADVANCES IN MATHEMATICAL DATA SCIENCE Chair: Xin Guo

E0538: **Robustness of kernel-based pairwise learning**
Presenter: **Patrick Gensler**, University of Bayreuth, Germany
Co-authors: Andreas Christmann

Pairwise learning can be applied in a variety of situations such as ranking, which is an important topic in machine learning and information retrieval, and also similarity learning and distance metric learning. Many results on the statistical robustness of kernel-based pairwise learning can be derived under basically no assumptions on the input and output spaces. In particular, neither moment conditions on the conditional distribution of Y given $X = x$ nor the boundedness of the output space is needed. Results on the existence and boundedness of the influence function have been obtained and show the qualitative robustness of the kernel-based estimator.

E0628: **Error analysis of OWL algorithms with varying Gaussians and convex loss**
Presenter: **Daohong Xiang**, Zhejiang Normal University, China

The goal of precision medicine is to determine the optimal individualized treatment rules by considering the heterogeneity of patients, so as to maximize the expected clinical outcome. Outcome weighted learning (OWL) is one of the algorithms to estimate the optimal individualized treatment rules. We mainly study the convergence theory of OWL associated with varying Gaussians and general convex loss. Fisher's consistency of OWL with convex loss is proved by making full use of the convexity of the loss function. Under some noise conditions on distributions, a quantitative relationship between weighted misclassification error and weighted generalization error is proved. The sample error is estimated by using a projection operator and a tight bound for the covering numbers of reproducing kernel Hilbert spaces generated by Gaussian kernels. Fast learning rates of OWL associated with least square loss, exponential-hinge loss and r-norm SVM loss are derived explicitly.

E0810: **Regularized Kaczmarz algorithm**
Presenter: **Xuemei Chen**, University of North Carolina Wilmington, United States

A novel algorithmic framework based on the Kaczmarz algorithm is proposed for tensor recovery. This is an iterative algorithm for solving a linear constrained optimization problem (with a regularizer). We provide a thorough convergence analysis and its applications from the vector case to the tensor one. We also show numerical results on a variety of tensor recovery applications to illustrate the enormous potential of the proposed methods.

EO227 Room Virtual R3 TOPICS ON HIGH-DIMENSIONAL AND COMPLEX MODELS Chair: Eugen Pircalabelu

E0263: **Node aggregation in large-scale graphical models**
Presenter: **Ines Wilms**, Maastricht University, Netherlands
Co-authors: Jacob Bien

High-dimensional graphical models are often estimated using regularization by relying on edge sparsity as a simplifying assumption. We aggregate the nodes of the graphical model to produce parsimonious graphical models that provide an even simpler description of the dependence structure than would otherwise be possible. We develop a convex regularizer that estimates graphical models that are both edge-sparse and node-aggregated. The aggregation is performed in a data-driven fashion by leveraging side information in the form of a tree that encodes node similarity and facilitates the interpretation of the resulting aggregated nodes. We illustrate our proposal's practical advantages in simulations and applications.

E0660: **On variables selection Type-I and type-II error tradeoff for high dimensional logistic regression**
Presenter: **Jing Zhou**, KU Leuven, Belgium
Co-authors: Gerda Claeskens

In recent years, controlling false discovery rate (FDR), also known as type-I error, has gradually attracted attention to improving the reproducibility of variable selection. We focus on the variable selection problem for l_1-regularized logistic regression with p variables and n samples. In addition, we assume n, p follow a linear growth rate including both $n > p$ and $n \leq p$ cases. Since the l_1-regularizer performs variable selection by nature, we show that the corresponding selection type-I and type-II errors satisfy a tradeoff. This tradeoff is characterized asymptotically by describing type-I error rate (FDR) as a function of 1 - Type-II error rate (power) using a system of equations with six parameters. Further, we propose two applications of this tradeoff curve: (1) a sample size calculation procedure to achieve certain power under prespecified FDR level using the FDR-power tradeoff; (2) FDR level calibration for variable selection taking power into consideration. Similar asymptotic analysis for the model-X knockoff, which provides FDR controlled selection, is also investigated. We illustrate the type-I and type-II error tradeoff analysis using simulated and real data.

E0654: **Residual-based estimation of parametric copulas under regression**
Presenter: **Yue Zhao**, University of York, United Kingdom

A multivariate response regression model is studied where each coordinate is described by a location-scale regression, and where the dependence structure of the “noise” terms in the regression is described by a parametric copula. The goal is to estimate the associated Euclidean copula parameter given a sample of the response and the covariate. In the absence of the copula sample, the oracle ranks in the usual pseudo-likelihood estimation procedure are no longer computable. Instead, we base our estimation on the residual ranks calculated from some preliminary estimators of the regression functions. We show that the residual-based estimators are asymptotically equivalent to their oracle counterparts, even under severe divergence of the criterion functions in pseudo-likelihood estimation and when the dimension of the covariate in the regression is moderately diverging. Partially to serve this objective, we also study the weighted convergence of the residual empirical processes.

EO389 **Room Virtual R4** NEW FRONTIERS IN ECONOMETRICS **Chair: Namhyun Kim**

E0680: **Multiway empirical likelihood**
Presenter: **Harold Chiang**, University of Wisconsin-Madison, United States
Co-authors: Yukitoshi Matsushita, Taisuke Otsu

A general methodology is developed to conduct statistical inference for observations indexed by multiple sets of entities. We propose a novel multiway empirical likelihood statistic that converges to a chi-square distribution under the non-degenerate case, where corresponding Hoeffding type decomposition is dominated by linear terms. Our methodology is related to the notion of jackknife empirical likelihood but the leave-out pseudo values are constructed by leaving out columns or rows. We further develop a modified version of our multiway empirical likelihood statistic, which converges to a chi-square distribution regardless of the degeneracy and discover its desirable higher-order property compared to the t-ratio by the conventional Eickerr-White type variance estimator. The proposed methodology is illustrated by several important statistical problems, such as bipartite network, two-stage sampling, generalized estimating equations, and three-way observations.

E0924: **Biased 2SLS estimation in the presence of weak and/or many instruments**
Presenter: **Namhyun Kim**, University of Exeter, United Kingdom
Co-authors: Winfried Pohlmeier, Patrick Wongsaart

The aim is to address the severe finite sample bias of the two-stage least square estimator in the presence of many and/or weak instruments in the control function (CF) framework. The CF framework enables us to translate the weak and many instruments issues into the near singularity issues in the first and second stage of estimation steps of the two-stage least squares estimation (TSLSE), respectively. Therefore, we propose L2-norm regularization in the presence of weak and/or many instruments with the innovative penalty parameters in order to address the severe finite sample bias of the two-stage least squares estimator in the presence of weak and/or many instruments. The improvement of the bias is analytically shown by using the conventional higher-order approximation.

E0955: **Comparing survey based forecasts**
Presenter: **Yang Zu**, University of Nottingham, United Kingdom

New strategies are discussed to evaluate survey based macroeconomic forecasts by forecast evaluation tests.

EO145 **Room Virtual R5** RECENT DEVELOPMENT ON CHANGE POINT DETECTION **Chair: Weichi Wu**

E0166: **Multiple change point detection for high-dimensional data**
Presenter: **Wenbiao Zhao**, Hong Kong Baptist University, China
Co-authors: Lixing Zhu

Simultaneously detecting multiple change points are investigated for high-dimensional data with dimensions that can be of an exponential rate of the sample size. The proposed estimation approach utilizes a signal statistic that is based on a sequence of local U-statistics, no matter whether the data have a sparse or dense structure. Both expensive computations that exhaustive search algorithms need and false positives that hypothesis testing-based approaches have to control can be avoided. The estimation consistency can hold for the locations and number of change points even when the number of change points diverges at a certain rate as the sample size goes to infinity. Further, because of its visualization nature, in practice, plotting the signal statistic can greatly help identify the locations in contrast to existing methods. The numerical studies are conducted to examine its performance in finite sample scenarios and a real data example is analyzed for illustration.

E0331: **Multiscale jump testing and estimation under complex temporal dynamics**
Presenter: **Weichi Wu**, Tsinghua University, China
Co-authors: Zhou Zhou

The focus is on the problem of detecting jumps in an otherwise smoothly evolving trend whilst the covariance and higher-order structures of the system can experience both smooth and abrupt changes over time. The number of jump points is allowed to diverge to infinity with the jump sizes possibly shrinking to zero. The method is based on a multiscale application of an optimal jump-pass filter to the time series, where the scales are dense between admissible lower and upper bounds. For a wide class of non-stationary time series models and trend functions, the proposed method is shown to be able to detect all jump points within a nearly optimal range with a prescribed probability asymptotically under mild conditions. For a time series of length n, the computational complexity of the proposed method is $O(n)$ for each scale and $O(nlog1+n)$ overall, where is an arbitrarily small positive constant. Numerical studies show that the proposed jump testing and estimation method performs robustly and accurately under complex temporal dynamics.

E0830: **An adaptive-to-change ridge-ratio criterion for multiple change points in high-dimensional tensors**
Presenter: **Jiaqi Huang**, Beijing Normal University, China
Co-authors: Junhui Wang, Lixing Zhu

Two criteria are proposed for detecting change structure in tensor data, which include vector and matrix as order-one and two tensors. The first criterion is based on the Euclidean norm of the moving sums of tensor data, the second is based on the Euclidean norm of the moving sums of slices. To handle both dense and sparse scenarios, the norms defined are signal-adaptive to screen out those non-signal elements. Two signal statistics are respectively the ratios and the minimum of ratios of two moving sums in consecutive segments with a data-driven ridge function. The latter is going to take care of the scenarios where a fiber could have a very high dimension. The estimation consistency of the number of changes and their locations is derived. The results can still hold when the dimensions of fibers and the number of changes diverge at certain rates. Numerical studies are conducted to examine the finite sample performances of the proposed method and compare it with some existing competitors. We also analyse two real data examples for illustration.

EO079 **Room Virtual R6** **RECENT STATISTICAL ANALYSIS OF MICROBIOME DATA** **Chair: Sangwook Kang**

E1022: **Association test for longitudinal microbiome data**
Presenter: **Taesung Park**, Seoul National University, Korea, South
Co-authors: Nayeon Kang, Hyunwook Koh

High-throughput technologies allow a new era of metagenomics studies to explore microbial communities sampled directly. The main goal of human microbial studies is to detect associations between microbiota and subject grouping phenotype. However, the microbiome data has several issues to overcome such as count compositional structure, various total sequence reads per sample, over-dispersion and zero-inflation. Several tools have been developed to handle these characteristics. We propose a permutation approach to identifying differentially abundant markers between two groups. The proposed method is based on the logistic regression model and has the advantage of handling multiple markers easily. Compared to existing methods, the proposed approach shows better performance in empirical studies including simulations and real data studies.

E1035: **Using taxanomic ranks improves the prediction of case-control analysis of microbiome data**
Presenter: **Yujin Chung**, Kyonggi University, Korea, South

Recent studies reveal that microbial traits are differentially correlated in a phylogenetic tree. These results suggest that microbiome-based predictive models are improved by incorporating phylogenetic trees through the cophenetic distance. We propose a new way to use taxonomic ranks when a phylogenetic tree is not provided. We modified two phylogenetic tree-based predictive models to employ taxonomic ranks rather than phylogenetic trees. These predictive models were applied to microbiome data from patients with cirrhosis and hepatocellular carcinoma (HCC) and controls. The analysis shows that the taxonomic ranks improve predictive models as much as phylogenetic trees.

E1029: **Phylogenetic tree-based microbiome association test**
Presenter: **Sungho Won**, Seoul National University, Korea, South

Ecological patterns of the human microbiota exhibit high inter-subject variation, with few operational taxonomic units (OTUs) shared across individuals. To overcome these issues, non-parametric approaches, such as the Mann-Whitney U-test and Wilcoxon rank-sum test, have often been used to identify OTUs associated with host diseases. However, these approaches only use the ranks of observed relative abundances, leading to information loss, and are associated with high false-negative rates. We propose a phylogenetic tree-based microbiome association test (TMAT) to analyze the associations between microbiome OTU abundances and disease phenotypes. Phylogenetic trees illustrate patterns of similarity among different OTUs, and TMAT provides an efficient method for utilizing such information for association analyses. The proposed TMAT provides test statistics for each node, which are combined to identify mutations associated with host diseases. Power estimates of TMAT were compared with existing methods using extensive simulations based on real absolute abundances. Simulation studies showed that TMAT preserves the nominal type-1 error rate, and estimates of its statistical power generally outperformed existing methods in the considered scenarios. Furthermore, TMAT can be used to detect phylogenetic mutations associated with host diseases, providing more in-depth insight into bacterial pathology.

EO135 **Room Virtual R7** **ADVANCES IN TIME SERIES, RANDOM FORESTS AND CAUSAL INFERENCE** **Chair: Hiroshi Shiraishi**

E0560: **Generalized random forests for dependent data**
Presenter: **Hiroshi Shiraishi**, Keio University, Japan
Co-authors: Tomoshige Nakamura

The generalized random forests (GRF) is a nonparametric statistical estimation method based on random forests. We consider the asymptotic property of GRF under a time-series setting.

E0650: **Causal trees and forest with sufficient dimension reduction**
Presenter: **Tomoshige Nakamura**, Keio University, Japan
Co-authors: Hiroshi Shiraishi

The causal trees and forests are one of the methods for nonparametric statistical estimation method for individual causal effects based on random forests. We consider causal trees and causal forests that use sufficient dimension reduction (SDR) techniques to approximate a locally adaptive kernel.

E0170: **Study of a well-known importance measure computed via decision trees**
Presenter: **Erwan Scornet**, Polytechnique, France

Nowadays, machine learning procedures are used in many fields with the notable exception of so-called sensitive areas (health, justice, defense, to name a few) in which the decisions to be taken are fraught with consequences. In these fields, it is necessary to obtain a precise decision but, to be effectively applied, these algorithms must provide an explanation of the mechanisms that lead to the decision and, in this sense, be interpretable. Unfortunately, the most accurate algorithms today are often the most complex. A classic technique to try to explain their predictions is to calculate indicators corresponding to the strength of the dependence between each input variable and the output to be predicted. We will focus on one measure of importance created for decision trees and we will see how the theoretical study provides explanations on its practical use.

EP001 **Room Poster Room** **POSTER SESSION (ONLY VIRTUAL)** **Chair: Cristian Gatu**

E0445: **Graphical and numerical diagnostic tools to assess multiple imputation models by posterior predictive checking**

Presenter: **Mingyang Cai**, Utrecht University, China

A method is proposed to diagnose imputation models based on posterior predictive checking. To assess the congeniality of imputation models, we compare the observed data with their replicates generated under corresponding posterior predictive distributions. The idea is that if the imputation model is congenial with the substantive model, the observed data is expected to locate in the centre of corresponding predictive posterior distributions. We investigate the proposed diagnostic method for parametric and non-parametric imputation approaches, continuous and discrete incomplete variables, univariate and multivariate missingness patterns. The results show the validity of the proposed diagnostic method.

E0556: **Variogram modeling for spatial correlation in structural MRI images**

Presenter: **Saed MaraBeh**, Qatar University, Qatar

Co-authors: Esam Mahdi

The current research studies in medical sciences involve data that are randomly collected during the time over specified locations. We use geostatistical techniques such as the variogram and kriging approaches to uncover the spatial correlation in structural magnetic resonance imaging (sMRI) data and to predict the effect of a brain tumor on brain regions. We propose various variogram model approaches for some brain slices containing a brain tumor and the best model is selected by using the cross-validation method in kriging. A bootstrap resampling method is used to estimate the parameters of the selected models.

E0832: **Matching quantiles estimation for discrete distribution**

Presenter: **Hyungjun Lim**, Korea University, Korea, South

Co-authors: Arlene Kyoung Hee Kim

Analysis of independently collected data requires unpaired data analysis to account for the missing correspondence between the response variable and explanatory variables. Quantile matching estimation (QME) is one such method, which is built to find a linear combination of explanatory variables such that its distribution best matches the distribution of the response variable. Despite active research in the unpaired data analysis, no prior studies have been conducted for the case where the response variable follows a discrete distribution. We introduce a novel Poisson quantile matching estimation (PQME) as the first unpaired data analysis method designed for the discrete-count response variable. A simple yet effective algorithm of PQME is presented and its theoretical properties are proved. Simulation studies and real data applications are included to demonstrate both the practicality and the effectiveness of PQME compared to conventional methods such as GLM.

Monday 06.06.2022 15:15 - 16:55 Parallel Session O – EcoSta2022

EO321 Room Live Theater (Hybrid 2) ADVANCEMENTS IN BAYESIAN MIXTURE MODELS Chair: Raffaele Argiento

E0181: **Model selection for maternal hypertensive disorders with symmetric hierarchical Dirichlet processes**
Presenter: **Beatrice Franzolini**, Bocconi University, Italy
Co-authors: Antonio Lijoi, Igor Pruenster

Hypertensive disorders of pregnancy occur in about 10% of pregnant women around the world. Though there is evidence that hypertension impacts maternal cardiac functions, the relation between hypertension and cardiac dysfunctions is only partially understood. The study of this relationship can be framed as a joint inferential problem on multiple populations, each corresponding to a different hypertensive disorder diagnosis, that combines multivariate information provided by a collection of cardiac function indexes. A Bayesian nonparametric approach seems particularly suited for this setup and we demonstrate it on a dataset consisting of transthoracic echocardiography results of a cohort of Indian pregnant women. We are able to perform model selection, provide density estimates of cardiac function indexes and a latent clustering of patients: these readily interpretable inferential outputs allow us to single out modified cardiac functions in hypertensive patients compared to healthy subjects and progressively increased alterations with the severity of the disorder. The analysis is based on a Bayesian nonparametric model that relies on a novel hierarchical structure, called symmetric hierarchical Dirichlet process. This is suitably designed so that the mean parameters are identified and used for model selection across populations, a penalization for multiplicity is enforced, and the presence of unobserved relevant factors is investigated through a latent clustering of subjects.

E0273: **Mixtures of finite mixtures and the telescoping sampler**
Presenter: **Gertraud Malsiner-Walli**, WU Vienna University of Economics and Business, Austria
Co-authors: Sylvia Fruhwirth-Schnatter, Bettina Gruen

Within a Bayesian framework, a comprehensive investigation of the model class of mixtures of finite mixtures (MFMs) where a prior on the number of components is specified is performed. This model class has applications in model-based clustering as well as for semi-parametric density approximation but requires suitable prior specifications and inference methods to exploit its full potential. We contribute to the Bayesian analysis of MFMs by (1) considering static and dynamic MFMs where the Dirichlet parameter of the component weights either is fixed or depends on the number of components, (2) proposing a flexible prior distribution class for the number of components K, (3) characterizing the implicit prior on the number of clusters as well as partitions by deriving computationally feasible formulas, (4) linking MFMs to Bayesian non-parametric mixtures and (5) finally proposing a novel sampling scheme for MFMs called the telescoping sampler which allows Bayesian inference for mixtures with arbitrary component distributions. The telescoping sampler explicitly samples the number of components, but otherwise requires only the usual MCMC steps for estimating a finite mixture model. The ease of its application using different component distributions is demonstrated on real data sets.

E0851: **Learning the number of clusters: Conjugate prior for the Dirichlet process precision parameter**
Presenter: **Tommaso Rigon**, University of Milano-Bicocca, Italy
Co-authors: Alessandro Zito, David Dunson

A new and flexible prior distribution for the precision parameter of a Dirichlet process is introduced. We show how this prior is conjugate to the distribution of the number of distinct values arising from the process, thus admitting a posterior within that same family. Moments, properties and hyperparameters interpretation of the distribution are extensively studied, as well as its relationship with the class of exponential families. Interestingly, certain choices for the hyperparameters allow computing the normalizing constant explicitly. We show how this allows drawing a parallel with common Bayesian nonparametric models within the class of Gibbs-type processes. We illustrate the computational and practical advantages of using this prior over common alternatives proposed in the literature when adopting a Dirichlet process-based clustering algorithm.

E1003: **Clustering categorical data via Hammingd istance**
Presenter: **Raffaele Argiento**, Universita degli Studi di Bergamo, Italy
Co-authors: Lucia Paci, Edoardo Filippi-Mazzola

Clustering methods have typically found their application when dealing with continuous data. However, in many modern applications data consist of multiple categorical variables with no natural ordering. In the heuristic framework, the problem of clustering these data is tackled by introducing suitable distances. We develop a model-based approach for clustering categorical data with a nominal scale. The approach is based on a mixture of distributions defined via the Hamming distance between categorical vectors. Maximum likelihood inference is delivered through an expectation-maximization algorithm. A simulation study is carried out to illustrate the proposed approach.

EO323 Room Main Theater (Hybrid 1) MONETARY AND FISCAL POLICIES IN DSGE MODELS Chair: Takeki Sunakawa

E0211: **Forward guidance as a monetary policy rule**
Presenter: **Takeki Sunakawa**, Hitotsubashi University, Japan
Co-authors: Mitsuru Katagiri

Many central banks implement forward guidance as a state-contingent policy rather than an exogenous policy action in practice. This paper investigates the effects of forward guidance by formulating it as a systematic part of the monetary policy rule in a non-linear new Keynesian model, and shows that rule-based forward guidance can significantly mitigate adverse effects on inflation by changing the way of forming expectations about what the central bank can do in a crisis. A quantitative analysis shows that rule-based forward guidance provides new insights about controversial issues including the forward guidance puzzle and the missing deflation puzzle during the Great Recession.

E0212: **Systematic foreign exchange intervention and macroeconomic stability: A Bayesian DSGE approach**
Presenter: **Mitsuru Katagiri**, Hosei University, Japan

The role of foreign exchange interventions (FXIs) is studied by introducing a systematic FXI policy into a small open economy DSGE model. While the systematic FXI policy can either stabilize or destabilize the economy depending on the type of shocks (productivity, external, or monetary), a quantitative analysis of Vietnamese data using a Bayesian method reveals that FXIs significantly contribute to macroeconomic stability there. With FXIs that sufficiently insulate an economy from the external shock, the real FX rate is mainly accounted for by productivity shocks, in contrast with the exchange rate disconnect but consistent with the Balassa–Samuelson relationship.

E0214: **Habit persistence and zero lower bound risk under optimal discretionary policy**
Presenter: **Kohei Hasui**, Aichi University, Japan
Co-authors: Satoshi Hoshino

Previous studies have shown that the risk of nominal interest rates hitting the zero lower bound (ZLB) has profound implications for monetary policy. We show that habit persistence is a non-negligible deep parameter under the optimal discretionary policy when the risk of the ZLB is taken into account. The uncertainty effect, which is defined as a difference between risky steady-state (RSS) and deterministic steady-state (DSS), increases as the habit persistence increases. Under empirically reasonable values of habit persistence, we show that the RSS of the nominal interest rate would reach the ZLB under the optimal discretionary policy. Moreover, the uncertainty effect of the ZLB worsens welfare more as the habit persistence increases.

E0450: **The signalling effects of fiscal announcements**
Presenter: **Hiroshi Morita**, Hosei University, Japan
Co-authors: Francesco Zanetti, Leonardo Melosi

Fiscal announcements may transfer information about the government's view of the macroeconomic outlook to the private sector, diminishing the effectiveness of fiscal policy as a stabilization tool. We construct a novel data set that combines daily data on Japanese stock prices with narrative records from press releases about a set of extraordinary fiscal packages introduced by the Japanese government from 2011-2020. We use local projections to show that these fiscal stimuli were often interpreted as negative news by the stock market whereas exogenous fiscal interventions do not convey any information about the business cycle (e.g., the successful bids to host the Olympics on September 8, 2013) fostered bullish reactions. In addition, these negative effects on stock prices arose more commonly when fiscal stimuli were announced against a backdrop of heightened macroeconomic uncertainty. Both findings are shown to be consistent with the theory of signaling effects.

EO175 Room Virtual R1 COMPOSITIONAL DATA ANALYSIS Chair: Christine Thomas-Agnan

E0512: **Contributions of the compositional data methodology to constrained optimization in economics**
Presenter: **Jordi Saperas Riera**, Universitat de Girona, Spain
Co-authors: Josep Antoni Martin-Fernandez

Compositional data provide a specific geometry to the simplex that allows us to study the relationship between the parts of a whole. This geometry is known as Aitchison geometry. The basic operations of Aitchison's geometry defined on the simplex are perturbation and powering. Consequently, the concepts and statistical techniques that are part of the analysis of compositional data must be consistent with Aitchison's geometry. In economics, it is a common problem to look for the distribution of resources that optimizes an indicator or a function. In addition, in many cases, this optimization of resources is conditioned because the variables are constrained. In the modelling of compositional data, we want to make some contributions that allow us to formulate and solve constrained optimization problems, especially convex optimization problems, in a compatible way with Aitchison geometry. To this end, we will define the convex set and convex function in the simplex. Some examples will be presented for illustrating this approach.

E0659: **Climate change and rice yield: Compositional scalar-on-function regression approach**
Presenter: **Thi Huong Trinh**, Thuongmai University, Vietnam

Climate change has a significant impact on crop yields, especially in an agricultural country like Vietnam. Climate change is measured by changes in the maximum and minimum daily temperature for 30 years, from 1987 to 2016. We address the impact of weather, here the maximum (and minimum) daily temperature on rice yield per year in each province through a compositional scalar-on-function regression. A total of 3780 samples, i.e. province's daily temperature per year, are expressed as density functions in the Bayes space B^2. The functional centered log-ratio transformation, clr, converts the density function from B^2 space to L^2 space. We discretize the observed temperature densities and then smooth them using splines in L^2 with a zero integral constraint, which are adapted to our problem. Smoothing splines of the clr temperature function and a scalar dependent variable are treated as a functional linear regression model in L^2. The estimated function, represented in a $ZB-$spline basis, is obtained by minimizing the sum of squared errors and then transferred back to B^2 with the functional clr inverse transformation. The results in B^2 directly provide insight into the impact of climate on rice yield.

E0750: **Detecting outliers in compositional data using invariant coordinate selection**
Presenter: **Christine Thomas-Agnan**, Universite Toulouse, France
Co-authors: Anne Ruiz-Gazen, Thibault Laurent, Camille Mondon

Invariant Coordinate Selection (ICS) is a multivariate statistical method based on the simultaneous diagonalization of two scatter matrices. A model-based approach of ICS, called Invariant Coordinate Analysis, has already been adapted for compositional data. In a model-free context, ICS is also helpful in identifying outliers. We propose to develop a version of ICS for outlier detection in compositional data. This version is first introduced in coordinate space for a specific choice of ilr coordinate system associated with a contrast matrix and follows an existing outlier detection procedure. We then show that the procedure is independent of the choice of contrast matrix and can be defined directly in the simplex. To do so, we first establish some properties of the set of matrices satisfying the zero-sum property and introduce a simplex definition of the Mahalanobis distance and the one-step M-estimators class of scatter matrices. We also need to define the family of elliptical distributions in the simplex. We then show how to interpret the results directly in the simplex using two artificial datasets and a real dataset of market shares in the automobile industry.

E0763: **Location powered quotient: A compositional data analysis-based approach**
Presenter: **Takahiro Yoshida**, The University of Tokyo, Japan
Co-authors: Daisuke Murakami, Hajime Seya

A typical measure of industrial concentration is the Location Quotient (LQ), which is simply calculated as the regional and national ratios of employment in each industrial sector. However, its calculation focuses on a single sector and thus ignores relationships with other sectors. Therefore, we propose an alternative version of LQ based on compositional data analysis, which is commonly used and developed in geosciences. The proposed index, Location Powered Quotient (LPQ), has the following properties. (1) LPQ is derived from the powering operator in Aitchison's vector space structure, (2) LPQ considers not only specialization but also the balance of composition, and (3) LPQ has a sign. We apply this LPQ to an analysis of Japanese industry data to examine how the LPQ is interpreted.

EO025 Room Virtual R2 INNOVATIVE TECHNOLOGIES FOR BIG DATA Chair: Philip Yu

E0630: **Return correlation and volatility spillover among NFT, NFT-related coin, and cryptocurrency markets**
Presenter: **Kin Hon Ho**, The Hang Seng University of Hong Kong, Hong Kong
Co-authors: Tse-Tin Chan, Philip Yu

Since early 2021, non-fungible tokens (NFTs) have received tremendous attention, and their prices have boosted dramatically. There are various types of NFTs, such as collectibles, artworks, and digital characters in metaverses. With the success of NFTs, NFT-related coins and traditional cryptocurrencies play an essential role in the NFT ecosystem. One may find it interesting to explore the relationships among NFT, NFT-related coins, and traditional cryptocurrency markets regarding their pricing. We investigate these relationships by exploring their return correlation using wavelet coherence analysis to measure the co-movement of assets in terms of both time and frequency. We also use connectedness (spillover) indices to examine the volatility shock transmission from one market to another markets. Our results reveal moderate correlations between NFT-related coins and traditional cryptocurrencies, as well as strong correlations between NFT-related coins. Nevertheless, there are weak return correlations and low volatility transmission between NFT and cryptocurrency markets as well as between NFTs. More interestingly, there is no significant correlation and volatility transmission between NFTs and their related coins in most cases, especially in short cycles, where they are generally perceived to be closely associated. Therefore, insights into facilitating risk management through portfolio diversification are provided.

E0696: **A distributional perspective on autoencoder asset pricing models**
Presenter: **Zhoufan Zhu**, Shanghai University of Finance and Economics, China
Co-authors: Ke Zhu, Dong Li, Xuanling Yang

Quantitative trading and investment decision making are intricate financial tasks that rely on accurate asset selection. Despite advances in deep learning that have made significant progress in the complex and highly stochastic asset return prediction problem, the previous conditional asset pricing models face two major limitations. One is that they only focus on the mean of asset returns, and the other is that they ignore the potential randomness in latent factors. To get rid of these limitations, we consider a Dirac mixture model to represent the distribution over asset returns and employ the Variational Autoencoder to measure the randomness in latent factors. The key novelty is modeling the asset return distribution with a parametric model, which allows us to consider the expectation and variance(risk) of asset return simultaneously. Through simulations and an application to the US market spanning over sixty years of data, we show that our proposed method significantly outperforms previous relevant ones.

E0327: **Bayesian robust tensor completion via CP decomposition**
Presenter: **Xiaohang Wang**, Zhuhai Fundan Innovation Institute, China
Co-authors: Philip Yu, Weidong Yang, Jun Su

The real-world tensor data are inevitably missing and corrupted with noise. Some models of the low-rank tensor factorization (LRTF) add an L1 norm or L2 norm to deal with the sparse or Gaussian noise. However, the real noise is usually complex. We propose a robust Bayesian tensor completion method, called MoG BTC-CP, which could impute the missing data and remove the complex noise simultaneously. The observed tensor is assumed to be the summation of a low-rank tensor and the noise. CP decomposition is proposed to extract the low-rank structure of the tensor. We assume that the noise follows a Mixture of Gaussian (MoG) distribution. A full Bayesian framework together with a Gibbs sampling algorithm is designed to estimate the model. Extensive experiments including synthetic data and real-life applications show that MoG BTC-CP outperforms the recently published leading tensor completion and denoising methods.

E0596: **Preference learning across social networks for recommendations**
Presenter: **Philip Yu**, The Education University of Hong Kong, Hong Kong
Co-authors: Yipeng Zhuang

Preference learning refers to the problem of learning from preference data, which can ultimately understand individuals' preference behaviors. A typical problem is personalized item recommendation where users in a social media platform rated a set of items and their preferences may be influenced by their peers or friends in a social network. However, not all items were rated and many of these ratings are missing. We propose novel models for learning incomplete preferences of items across social networks. Finally, we apply our models to various big datasets on personalized movie recommendations with the goal of better prediction of the ratings of unrated movies for possible recommendation.

EO263 Room Virtual R3 BAYESIAN MODELING IN SOCIAL SCIENCES Chair: Kazuhiko Kakamu

E0690: **Bayesian dynamic modeling of Gini coefficient from grouped data**
Presenter: **Kazuhiko Kakamu**, Nagoya City University, Japan

A dynamic model is proposed for income distribution, which enables us to examine the Gini coefficient directly. In the analysis of income distribution, the choice of the hypothetical distribution is crucial and reflects on the results. It means that the dynamics of the Gini coefficients may be different depending on the choice of the hypothetical income distribution. Our approach also enables us to examine several kinds of hypothetical income distributions. This model is applied to several simulated and real datasets. The results of a real dataset, which comes from a Family Income and Expenditure Survey, show the importance of the choice of the hypothetical income distribution.

E0901: **Estimation of area-wise income distributions based on household-level grouped data**
Presenter: **Yuki Kawakubo**, Chiba University, Japan
Co-authors: Genya Kobayashi

From the household-level grouped income data, various characteristics in the area-wise income distributions are estimated. We observe which of the mutually exclusive intervals, separated by pre-specified thresholds, the sampled households' incomes belong. When estimating area-wise income distribution from such data, we face two problems: one is how to recover the underlying continuous variable (income) from the grouped data, and the other is that the estimation efficiency becomes poor when the sample size in each area is not sufficiently large. To address these issues, we treat the underlying household income as a latent variable, which is assumed to follow a mixed-effects model that incorporates household-level and area-level auxiliary variables and random effects as area effects. The effectiveness of using mixed-effects models has been actively studied on small area estimation. By predicting the latent variable for each household based on the observations as grouped data and auxiliary variables, we estimate various characteristics in the area-wise income distribution. This method is applied to Japanese income data to estimate not only the mean or median income, but also the Gini coefficient and several poverty indices for each area.

E0525: **Spatio-temporal smoothing, interpolation and prediction of income distributions based on grouped data**
Presenter: **Genya Kobayashi**, Meiji University, Japan
Co-authors: Shonosuke Sugasawa, Yuki Kawakubo

In science, especially in social science, exact values of some characteristics of individuals are not directly observed, but values of interest are grouped or collapsed in such a way that only numbers of individuals who belong to groups are observed. This type of data is typically called grouped data and any analysis should address this grouped nature. A new methodology of mixture modelling for grouped data observed over multiple spatial units or clusters and time periods is developed. The main idea is that all clusters share the common latent distributions and potential cluster-wise heterogeneity is captured by the cluster-wise mixing proportions. To model the unknown cluster-wise mixing proportions, we employ the multinomial logistic functions that include spatial and temporal effects. The inclusion of these effects enables smoothing of quantities of interest over time and space, imputation of missing values and prediction of future values. Using Polya-gamma data augmentation, an efficient posterior computational algorithm via Gibbs sampling is developed. As a specific application of the proposed method, modelling of cluster-wise income distributions based on longitudinal grouped data is considered. The usefulness of the method is demonstrated through the simulated data and income survey data of Japan.

E0534: **Spatially-varying bayesian predictive synthesis for flexible and interpretable spatial prediction**
Presenter: **Shonosuke Sugasawa**, University of Tokyo, Japan
Co-authors: Danielle Cabel, Masahiro Kato, Kenichiro McAlinn, Kosaku Takanashi

Spatial data are characterized by their spatial dependence, which is often complex, non-linear, and difficult to capture with a single model. Significant levels of model uncertainty – arising from these characteristics – cannot be resolved by model selection or simple ensemble methods, as performances are not homogeneous. We address this issue by proposing a novel methodology that captures spatially-varying model uncertainty, which we call spatial Bayesian predictive synthesis. The proposal is defined by specifying a latent factor spatially-varying coefficient model as the synthesis function, which enables model coefficients to vary over the region to achieve flexible spatial model ensembling. Two MCMC strategies are implemented for full uncertainty quantification, as well as a variational inference strategy for fast point inference. We also extend the estimations strategy for general responses. A finite sample theoretical guarantee is given for the predictive performance of our methodology, showing that the predictions are exact minimax. Through simulation examples and two real data applications, we demonstrate that our proposed spatial Bayesian predictive synthesis outperforms standard spatial models and advanced machine learning methods, in terms of predictive accuracy, while maintaining interpretability of the prediction mechanism.

EO105 Room Virtual R4 RECENT ADVANCES OF TIME SERIES ANALYSIS Chair: Wai-keung Li

E0359: **SARMA: A computationally scalable high-dimensional time series model**
Presenter: **Feiqing Huang**, University of Hong Kong, Hong Kong
Co-authors: Yao Zheng, Kexin Lu, Guodong Li

A novel parametric infinite-order vector autoregressive model is introduced. As a variant of the vector autoregressive moving average (ARMA) model, it not only inherits desirable properties such as parsimony and rich temporal dependence structures, but also avoids two well-known drawbacks of the former: (i) non-identifiability and (ii) computational intractability even for moderate-dimensional data. Moreover, its parameter estimation is scalable with respect to the complexity of temporal dependence, namely the number of decay patterns constituting the autoregressive structure; hence it is called the scalable ARMA (SARMA) model. In the high-dimensional setup, we further impose a low-Tucker-rank assumption on the coefficient tensor of the proposed model. The resulting model has the form of a regression with embedded dynamic factors and hence can be especially suited for financial and economic data. Non-asymptotic error bounds for the proposed estimator are derived, and a tractable alternating least squares algorithm is developed. Theoretical and computational properties of the proposed method are verified by simulation studies, and the advantages over existing methods are illustrated in real applications.

E0586: **Estimation based on martingale difference divergence with insufficient instrumental variables**
Presenter: **Kunyang Song**, The University of Hong Kong, Hong Kong

Finding valid instrumental variables (IVs) is important but hard in the linear regression model. However, the classical estimation method such as 2-stage least square estimator is not applicable when the linear regression model is under-identified without enough valid IVs. Based on the martingale difference divergence (MDD), a new estimator is proposed for the general univariate nonlinear regression model, and this estimator is applicable even when the regression model is under-identified. Under certain regular conditions, the consistency and asymptotic normality of this MDD-based estimator is established. As an extension, a similar MDD-based estimator is also proposed for the multivariate nonlinear regression model. Simulations are given to illustrate the importance of the proposed estimators.

E0629: **Bootstrapping robust goodness-of-fit tests for GARCH models**
Presenter: **Muyi Li**, Xiamen University, China

A random weighting (RW) bootstrap approach is considered to conduct the robust goodness-of-fit tests for the generalized autoregressive conditional heteroskedastic (GARCH) models, which is estimated by the least absolute deviation (LAD) method. The RW bootstrap method perturbs both the minimand of the objective function and autocovariances of the transformed residual sequence, such that the test is applicable for very heavy-tailed innovations with only finite fractional moments. The testing procedure is easy to implement and the test statistic is robust to the choice of the random weights. The first-order consistency of the RW procedure is proved and the finite sample performance of the proposed test is assessed by numerical experiments. Finally, a real data analysis illustrates the usefulness of the test.

E0645: **Testing and modelling for the structural change in covariance matrix time series with multiplicative form**
Presenter: **Wai-keung Li**, The Education University of Hong Kong, Hong Kong

A new generalized Hausman test is constructed for detecting the structural change in a multiplicative form of the covariance matrix time series model. This generalized Hausman test is asymptotically pivotal, and it has non-trivial power in detecting a broad class of alternatives. Moreover, we propose a new semiparametric covariance matrix time series model, which has a time-varying long-run component to take the structural change into account, and a BEKK-type short-run component to capture the temporal dependence. A two-step estimation procedure is proposed to estimate this semiparametric model, and the asymptotic properties of the related estimators are established. Finally, the importance of the generalized Hausman test and the semiparametric model is illustrated by simulations and an application to realized covariance matrix data

EO123 Room Virtual R5 RECENT ADVANCES IN APPLIED PROBABILITY AND STATISTICS Chair: Li-Hsien Sun

E0197: **Nonparametric estimation of the continuous treatment effect with measurement error**
Presenter: **Wei Huang**, University of Melbourne, Australia
Co-authors: Zheng Zhang

The average dose-response function (ADRF) for a continuously valued error contaminated treatment is identified by a weighted conditional expectation. We then estimate the weights nonparametrically by maximising a local generalised empirical likelihood subject to an expanding set of conditional moment equations incorporated into the deconvolution kernels. Thereafter, we construct a deconvolution kernel estimator of ADRF. We derive the asymptotic bias and variance of our ADRF estimator and provide its asymptotic linear expansion, which can help conduct statistical inference. To select our smoothing parameters, we adopt the simulation-extrapolation method and propose a new extrapolation procedure to stabilise the computation. Monte Carlo simulations and a real data study illustrate our method's practical performance.

E0409: **Extracting stock predictive information in fund managers decisions through machine learning with hypergraph**
Presenter: **Chu-Lan Kao**, National Yang Ming Chiao Tung University, Taiwan
Co-authors: Vincent Tseng, You-Sin Chen

A machine learning framework is proposed that incorporates mutual fund managers' portfolio decisions to predict stock price movements. It employs a weighted hypergraph to connect the information of the funds' portfolio weight changing to the corresponding stocks in the portfolio, which is further aggregated with the price information extracted through classical machine learning methods. The framework then predicts the stock price movements as a classification problem. We put this framework to the test through the Taiwanese stock and mutual fund data, discovering that the managers' decisions do provide stock predictive information in addition to the classical technical data. Moreover, we determined that the level of information provided by the managers' decisions is asymmetric under different market conditions, which further establishes the existing findings in the literature on the mutual fund behavior under different market conditions. Further discussions on incorporating other expert major trading behaviors in Taiwan and model comparisons are provided.

E0762: **Nonparametric estimation of general mediation effects by calibration weighting**
Presenter: **Lukang Huang**, Nankai University, China
Co-authors: Zheng Zhang

To investigate causal mechanisms, causal mediation analysis decomposes the total treatment effect into the natural direct and indirect effects. The aim is to examine the estimation of the direct and indirect effects in a general treatment effect model, where the treatment could be binary, multi-valued, continuous, or mixture valued. We propose generalized weighting estimators with weights estimated by solving an expanding set of equations. Under some sufficient conditions, we show that the proposed estimators are consistent and asymptotically normal. Specifically, when the treatment is discrete, the proposed estimators attain the semi-parametric efficiency bounds. Meanwhile, when the treatment is continuous, the convergence rates of the proposed estimators are slower than root N; however, they are still more efficient than that constructed from the true weighting function. Moreover, a simulation study reveals that our estimators exhibit a satisfactory finite-sample performance, while an application shows its practical value.

E0882: **Estimation for multiple-threshold regression models**
Presenter: **ChihHao Chang**, National University of Kaohsiung, Taiwan

Threshold linear regression models with multiple threshold points are considered. We provide algorithms to efficiently estimate threshold points, where the number of thresholds and the type of thresholds (kink or jump) can be detected simultaneously. We conduct several simulation studies to demonstrate the finite-sample performance of the proposed algorithm versus existing algorithms. The simulation results reveal that the our method has less computational burden, comparable estimation accuracy and better flexibility for threshold type detection.

EO379 Room Virtual R6 HIGH DIMENSIONAL STATISTICAL ANALYSIS AND APPLICATIONS Chair: Su-Yun Huang

E0441: **Statistical learning for AI assisted clinics**
Presenter: **Henry Horng-Shing Lu**, National Yang Ming Chiao Tung University, Taiwan

The co-developments of AI-assisted clinics with Taipei Veterans General Hospital are discussed. The designs of computer-assisted diagnosis systems with deep learning techniques by multi-modalities of medical images are discussed for specific clinical applications. The related issues are investigated for the integration of statistical models, computational algorithms and domain knowledge. The current developments are summarized and the future potential studies are discussed.

E0176: **A generalized information criterion for high-dimensional PCA rank selection**
Presenter: **Hung Hung**, National Taiwan University, Taiwan

Principal component analysis (PCA) is a commonly used statistical tool for dimension reduction. An important issue in PCA is to determine the rank, which is the number of dominant eigenvalues of the covariance matrix. Among information-based criteria, the Akaike information criterion (AIC) and the Bayesian information criterion (BIC) are the two most common ones. Both use the number of free parameters for assessing model complexity, which requires the validity of the simple spiked covariance model. As a result, AIC and BIC may suffer from the problem of model misspecification when the tail eigenvalues do not follow the simple spiked model assumption. To alleviate this difficulty, we adopt the idea of the generalized information criterion (GIC) to propose a model complexity measure for PCA rank selection. The proposed model complexity takes into account the sizes of eigenvalues and, hence, is more robust to model misspecification. Asymptotic properties of our GIC are established under the high-dimensional setting, where $n \to \infty$ and $p/n \to c > 0$. Our asymptotic results show that GIC is better than AIC in excluding noise eigenvalues, and is more sensitive than BIC in detecting signal eigenvalues. Numerical studies and a real data example are presented.

E0168: **Perturbation theory for cross data matrix-based PCA**
Presenter: **Shao-Hsuan Wang**, National Central University, Taiwan
Co-authors: Su-Yun Huang

PCA has long been a useful tool for dimension reduction. Cross data matrix (CDM)-based PCA is another way to estimate PCA components, through splitting data into two subsets and calculating singular value decomposition for the cross product of the corresponding covariance matrices. CDM-based PCA has a broader region of consistency than ordinary PCA for leading eigenvalues and eigenvectors. We will introduce the finite sample approximation results as well as the asymptotic behavior for CDM-based PCA via matrix perturbation. Moreover, we introduce a comparison measure for CDM-based PCA vs. ordinary PCA. This measure only depends on the data dimension, noise correlations, and the noise-to-signal ratio (NSR). Using this measure, we develop an algorithm, which selects good partitions and integrates results from these good partitions to form a final estimate for CDM-based PCA. Numerical and real data examples are presented for illustration.

E0472: **Contrastive modeling for Cryo-EM 3D orientation estimations**
Presenter: **Szu-Chi Chung**, National Sun Yat-sen University, Taiwan

Learning methods are becoming popular to determine the 3D structure of a protein. The most remarkable achievement is arguably the recent release of AlphaFold 2, which significantly advances the understanding of protein by performing highly accurate structure prediction from chains of amino acid sequences. However, these methods are mainly used to predict a static structure that resembles the X-ray crystal, which may not be the native state of a protein. It is noted that Cryogenic electron microscopy (cryo-EM) can capture the protein in its native states, but providing a robust initial volume for cryo-EM reconstruction is still time-consuming. The data include heavy noise, huge dimensions, and many unlabeled samples (no clean target is available for training) with unknown orientations, making it challenging to reach a robust computation. We explore the possibility of using the recent advances in contrastive modeling to directly estimate the orientations for each cryo-EM image. We elaborate on our approach, which represents the orientation using unit quaternion instead of the traditional Euler angle and several loss functions we choose. Finally, we will discuss some ongoing research, including utilizing the framework to estimate and refine the contrast transfer function parameters and the end-to-end orientation estimation procedure.

Authors Index