PROGRAMME AND ABSTRACTS

3rd International Conference on
Econometrics and Statistics (EcoSta 2019)

http://cmstatistics.org/EcoSta2019

National Chung Hsing University, Taiwan
25 – 27 June 2019

ISBN: 978-9963-2227-6-6

Co-chairs:
Wen-Han Hwang, Min-ge Xie, Geoffrey McLachlan, Sangyeol Lee.

EcoSta Editors:
Ana Colubi, Erricos J. Kontoghiorghes, Manfred Deistler.

Scientific Programme Committee:
Toshihiro Abe, Makoto Aoshima, Yuko Araki, Mauricio Castro, Baojiang Chen, Feng Chen, Jeng-Min Chiou, Daniel Henderson, Kei Hirose, MingHung Kao, Yongdai Kim, Victor Hugo Lachos, Andrew Lawson, Thomas Lee, Yoonkyung Lee, Wai-Keung Li, Yi Li, Heng Lian, Catherine Liu, Zudi Lu, Shujie Ma, Shu-Kay Ng, David Nott, Marc Paolella, Debdeep Pati, Frederick Kin Hing Phoa, Artem Prokhorov, Peter Radchenko, Jeroen Rombouts, Yang Shen, Xinyuan Song, Minh-Ngoc Tran, Berwin Turlach, Jane-Ling Wang, Liqun Wang, Naisyin Wang, Sherry Wang, Hoi Ying Wong, Jingjing Wu, Han Xiao, Hiroshi Yadohisa, Feng Yao, Dalei Yu, Xibin Zhang, Xingqiu Zhao, Ping-Shou Zhong, Ding-Xuan Zhou and Ji Zhu

Local Organizing Committee:
Tsung-I Lin (NCHU), Chyong-Mei Chen (NYMU), Ray-Bing Chen (NCKU), Shih-Feng Huang (NUK), Chang-Yun Lin (NCHU), Li-Hsien Sun (NCU), Henghsiu Tsai (Academia Sinica),Wan-Lun Wang (FCU).

Dear Colleagues,

It is a great pleasure to welcome you to the 3rd International Conference on Econometrics and Statistics (EcoSta 2019). 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 the Department of Applied Mathematics and the Institute of Statistics of the National Chung Hsing University (NCHU), Taichung, Taiwan.

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

The EcoSta 2019 consists of about 155 sessions, four keynote talks, four invited sessions, and over 610 presentations. There are about 660 participants. These numbers confirm the support of the involved research communities to this important initiative. It is indeed promising that the EcoSta conference will become a successful medium for the dissemination of high quality research in Econometrics and Statistics, and facilitate 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 NCHU provides excellent facilities and a fantastic environment. The local host, volunteers, and sponsoring universities have substantially contributed through their effort to the successful organization of the conference. We thank them all for their support. Particularly we express our sincere appreciation to the host and main sponsor, Department of Applied Mathematics and the Institute of Statistics of the NCHU.

It is hoped that the quality of both the scientific programme and the NCHU will provide the participants with a productive, stimulating conference, and an enjoyable stay in Taiwan.

The Elsevier journals of Econometrics and Statistics (EcoSta) and Computational Statistics & Data Analysis (CSDA) are associated with CFEnetwork, CMStatistics, and the EcoSta 2019 conference. The participants are encouraged to submit their papers to special or regular peer-reviewed issues of EcoSta and CSDA, and to join the networks.

Finally, we are happy to announce that the 4th International Conference on Econometrics and Statistics (EcoSta2020) will take place at the Yonsai University, Seoul, South Korea, from Monday the 20th to Wednesday the 22nd of July 2020. Tutorials will take place on Thursday the 23rd of July 2020. You are invited to participate actively in these events.

Ana Colubi, Erricos J. Kontoghiorghes and Wen-Han Hwang
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 over 1650 members and the following specialized teams

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

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

CFEnetwork COMPUTATIONAL AND FINANCIAL ECONOMETRICS

http://www.CFEnetwork.org

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

Specialized teams

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

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

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

SCHEDULE

2019-06-24	2019-06-25	2019-06-26	2019-06-27
	Opening, 08:45 - 09:00	**F** EcoSta2019 08:35 - 10:15	**J - Keynote** EcoSta2019 09:00 - 09:50
	A - Keynote EcoSta2019 09:00 - 09:50	**Coffee Break** 10:15 - 10:45	**Coffee Break** 09:50 - 10:20
	Coffee Break 09:50 - 10:20	**G** EcoSta2019 10:45 - 12:25	**K** EcoSta2019 10:20 - 12:25
	B EcoSta2019 10:20 - 12:25	**Lunch Break** 12:25 - 14:00	**Lunch Break** 12:25 - 14:00
	Lunch Break 12:25 - 14:00	**H** EcoSta2019 14:00 - 15:40	**L** EcoSta2019 14:00 - 15:40
	C EcoSta2019 14:00 - 15:40	**Coffee Break** 15:40 - 16:10	**Coffee Break** 15:40 - 16:10
Registration & Ice Breaker 16:00 - 19:00	**Coffee Break** 15:40 - 16:10	**I** EcoSta2019 16:10 - 17:50	**M** EcoSta2019 16:10 - 17:25
	D EcoSta2019 16:10 - 17:25		**N - Keynote** EcoSta2019 17:40 - 18:30
	E - Keynote EcoSta2019 17:40 - 18:30		**Closing**, 18:30 - 18:45
	Welcome Reception 18:30 - 20:00	**Conference Dinner** 19:00 - 22:00	

REGISTRATION AND SOCIAL EVENTS

- *Registration*. The registration will be open on Monday the 24th of June 2019, 16:00 - 19:00, Tuesday the 25th of June 2019, 08:00-18:00, Wednesday the 26th of June 2019, 08:15-17:00, and Thursday the 27th of June 2019, 08:40-18:00. It will take place at Room S113, 1F of the College of Science Building (No. 38) **(see maps on pages VIII and IX)**. The conference badges have a QR code with the registration information of the participants. For this reason, it is mandatory to always bring the conference badge.
- *The coffee breaks* will take place at the hall of the College of Science Building (No. 38, 1F) **(see maps on pages VIII and IX)**. Participants must have your conference badge in order to attend the coffee breaks.
- *Resting area*. Participants can use the Rooms S201, S202 and S203 of the College of Science Building (No. 38, 2F) **(see maps on page VIII and IX)** to rest, have their lunches, etc. These rooms are fully air-conditioned.
- *Welcome Reception, Tuesday the 25th of June 2019, 18:30 - 20:00*. The Welcome Reception is open to all registrants who have pre-registered and accompanying persons who have purchased a reception ticket. It will take place at the small auditorium of NCHU campus (N0. 7) **(see map on page VIII)**. Conference registrants must bring their conference badge in order to attend the reception. Information about the welcome reception booking is embedded in the QR code on the conference badge. Preregistration is required due to health and safety reasons, and the limited capacity of the venue. Entrance to the reception venue will be strictly allowed only to those who have prebooked.
- *Conference Dinner, Wednesday the 26th of June 2019, from 19:00 to 22:00*. The conference dinner is optional and registration is required. It will take place at the the Hotel National **(see map on page XIV)**. Four shuttle buses, departing beside the campus library at 17:50, 18:00, 18:10 and 18:20 (see map on page VIII) will drive participants to the lobby of Hotel National. Conference registrants must bring their conference badge in order to attend the conference dinner. Participants must bring their conference badge to attend the dinner, as the information about the conference dinner booking is encoded in the QR code at the badge.
- *Lunches*. Participants can buy lunch at restaurants and cafes at the Student Center of the campus (No. 6) **(see map on page VIII and table on page XV)** or in the nearby area **(see map on page XIV)**.

TUTORIAL

A tutorial on *Semi-parametric financial tail risk forecasting incorporating realized measures* will be given by Prof. Richard Gerlach, The University of Sydney Business School, Australia. The tutorial will take place on Friday the 28th of June 2019 from 08:30 to 13:00 at Room U414, Information Science Building (No. 36) **(see maps on pages VIII and XI)**. Pre-registration is required.

SPECIAL MEETINGS by invitation only

- The *EcoSta session organizers* meeting will take place on Monday the 24th of June 2019, 17:30-17:45 at Room S101, 1F of the College of Science Building (No. 38) **(see maps on pages VIII and IX)**. After the meeting there will be a small reception at the Hall of the same building.
- The *CSDA & Econometrics and Statistics (EcoSta) Editorial Board* meeting and dinner will take place on Thursday the 27th of June 2019, 19:15-21:30 at a restaurant that will be announced in due course. The meeting is by invitation only.

GENERAL INFORMATION

Addresses of venues (see maps on page VIII)

The Conference venue is the National Chung Hsing University (NCHU), Taichung 402, Taiwan.

Lecture rooms (see maps on pages VIII to XIII)

The opening and keynote talks will take place at Room S1A03, 1F of the College of Science building. The paper presentations will take place at the College of Science Building (No. 38), the Information Science Building (No. 36) and the Applied Science & Technology Building (No. 16) of the NCHU **(see maps on pages VIII to XIII)**. The poster sessions will take place at the Hall of the College of Science Building (No. 38, 1F) **(see maps on pages VIII and IX)**. We advise that you visit the venue in advance.

Presentation instructions

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

Posters

The posters should be displayed only during their assigned session. The authors will be responsible for placing the posters in the poster panel displays and removing them after the session. The maximum size of the poster is A0.

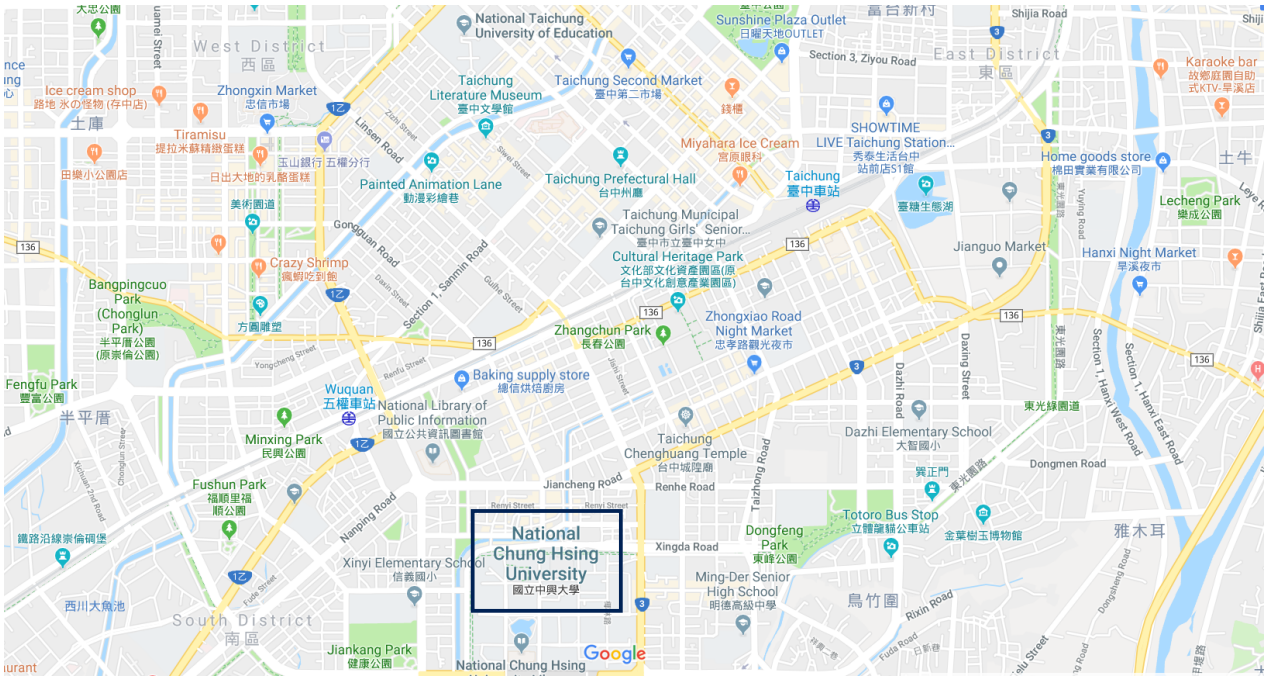
Internet connection

The WIFI name (SSID) is NCHU, the ID and password are both Sta2019. The domain must be kept blank (do not choose any domain).

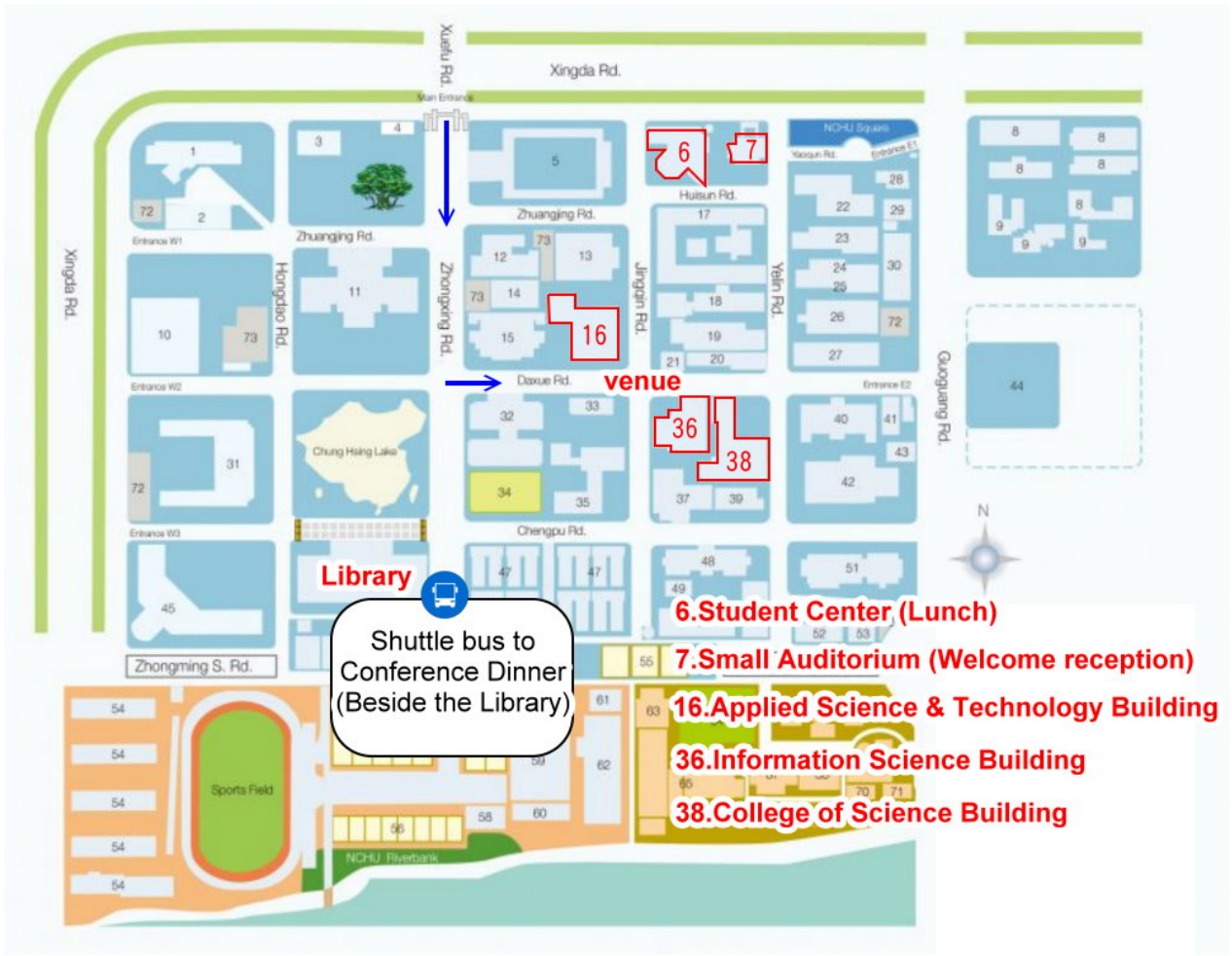
Exhibitors

Elsevier, Statistics Sinica.

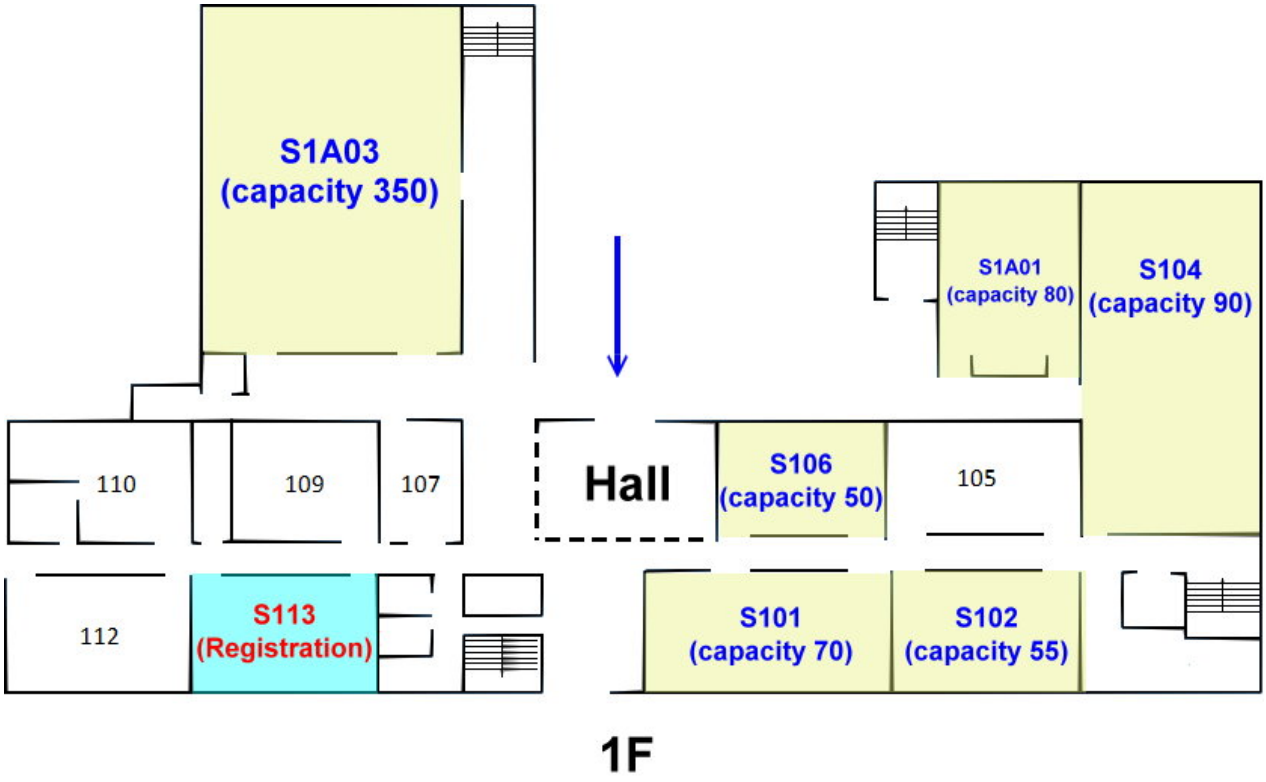
Google map of the venue and nearby area



Map of NCHU

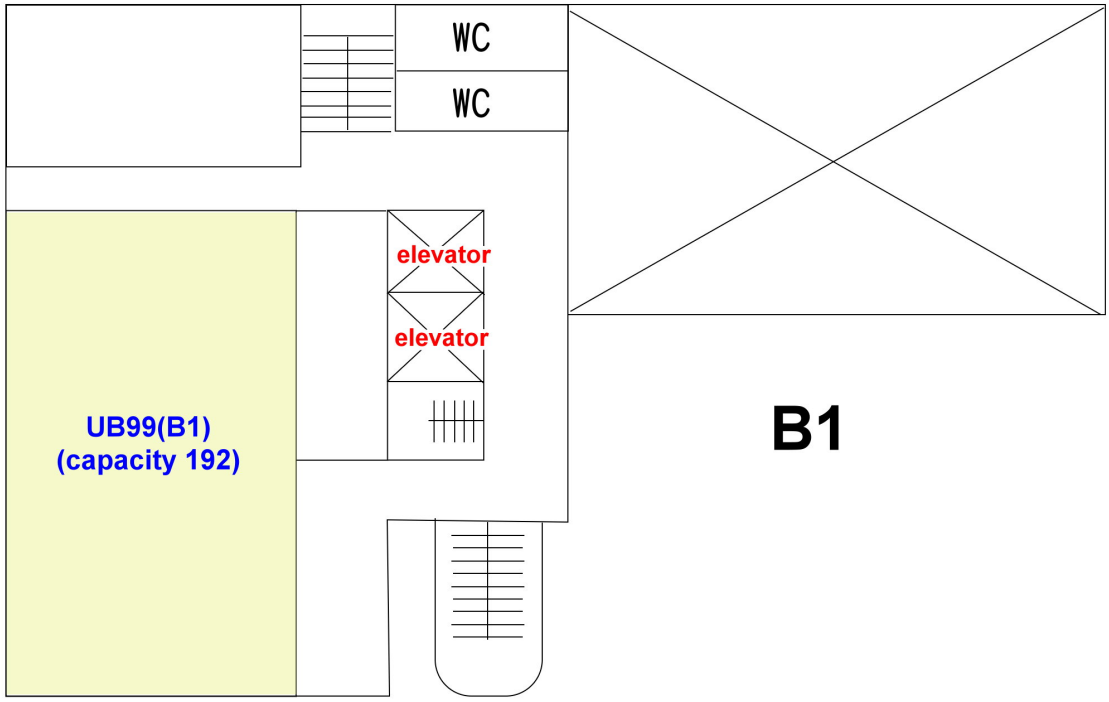


Venue: 1F of the College of Science Building

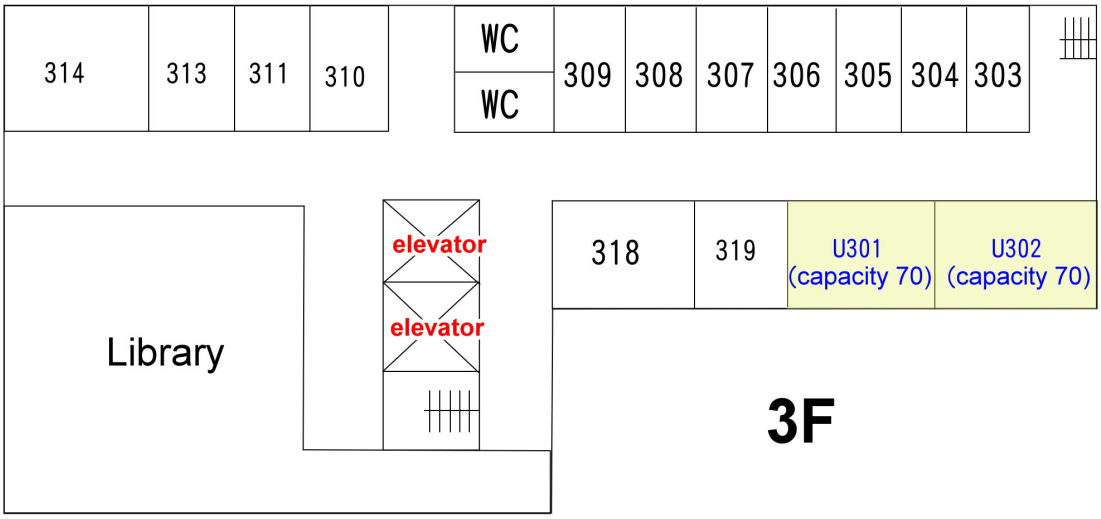


Venue: 2F of the College of Science Building

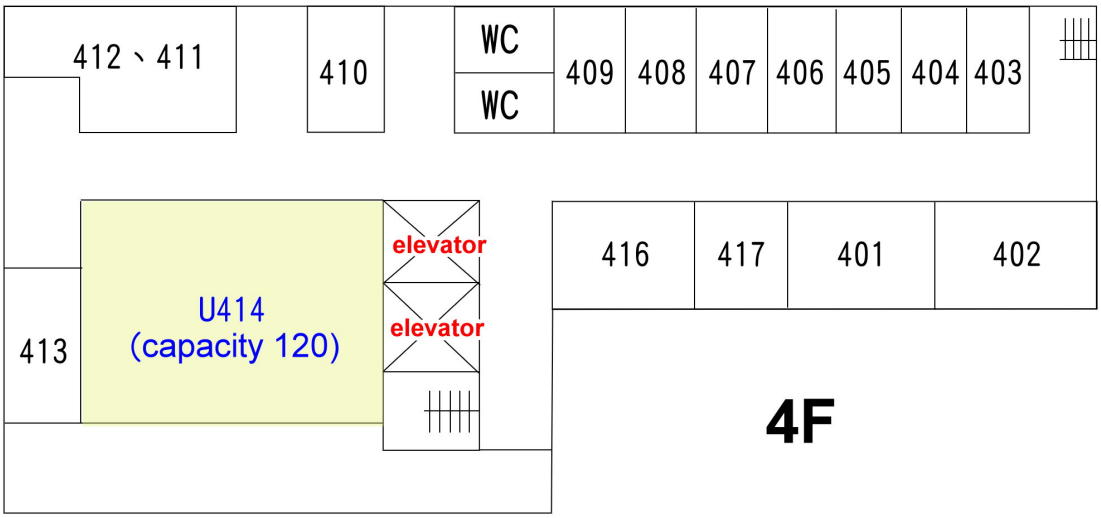
Venue: B1 of the Information Science Building



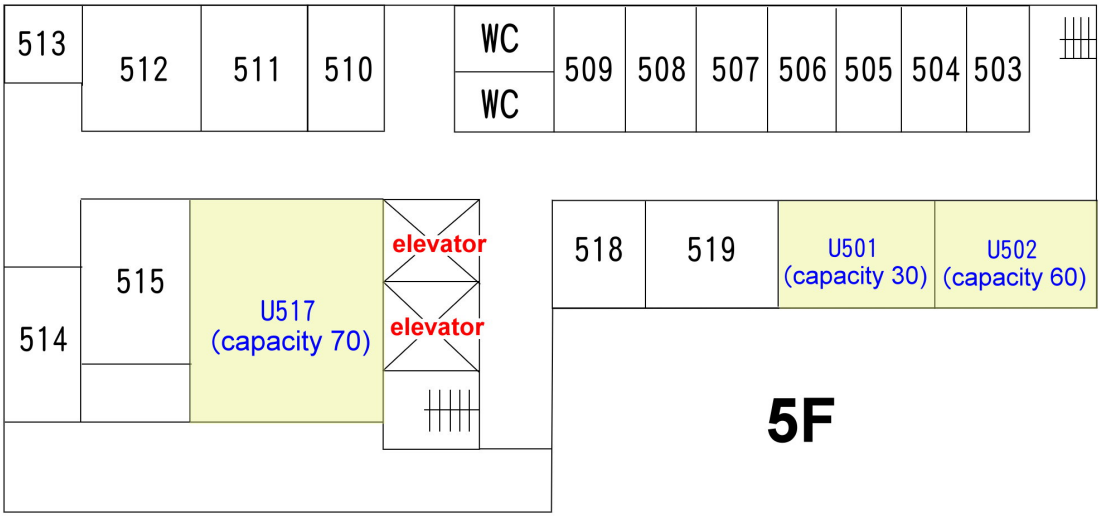
3F of the Information Science Building



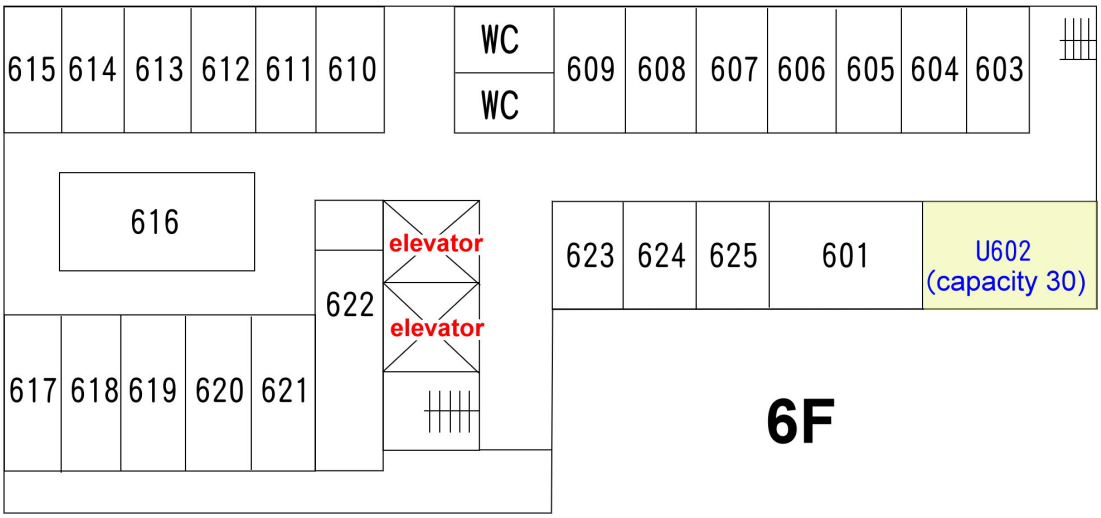
Venue: 4F of the Information Science Building



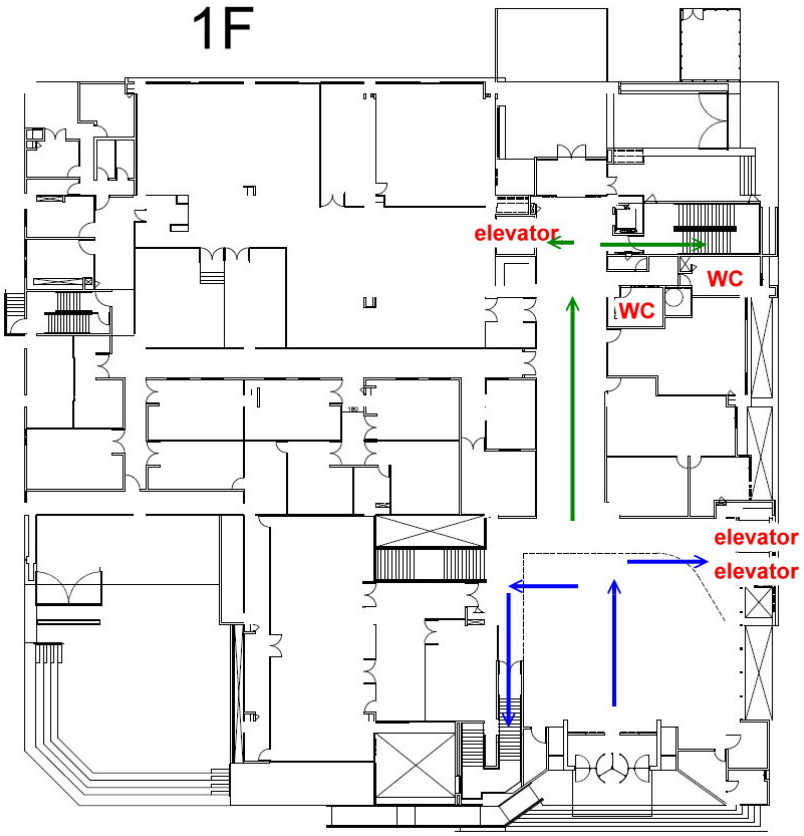
Venue: 5F of the Information Science Building



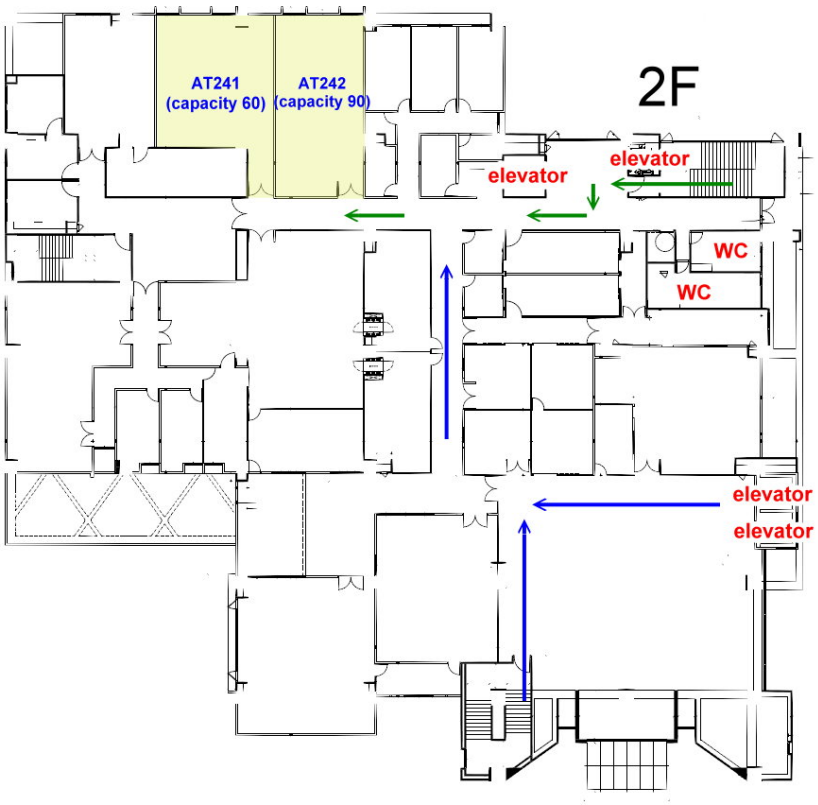
Venue: 6F of the Information Science Building



Venue: 1F of the Applied Science and Technology Building



Venue: 2F of the Applied Science and Technology Building

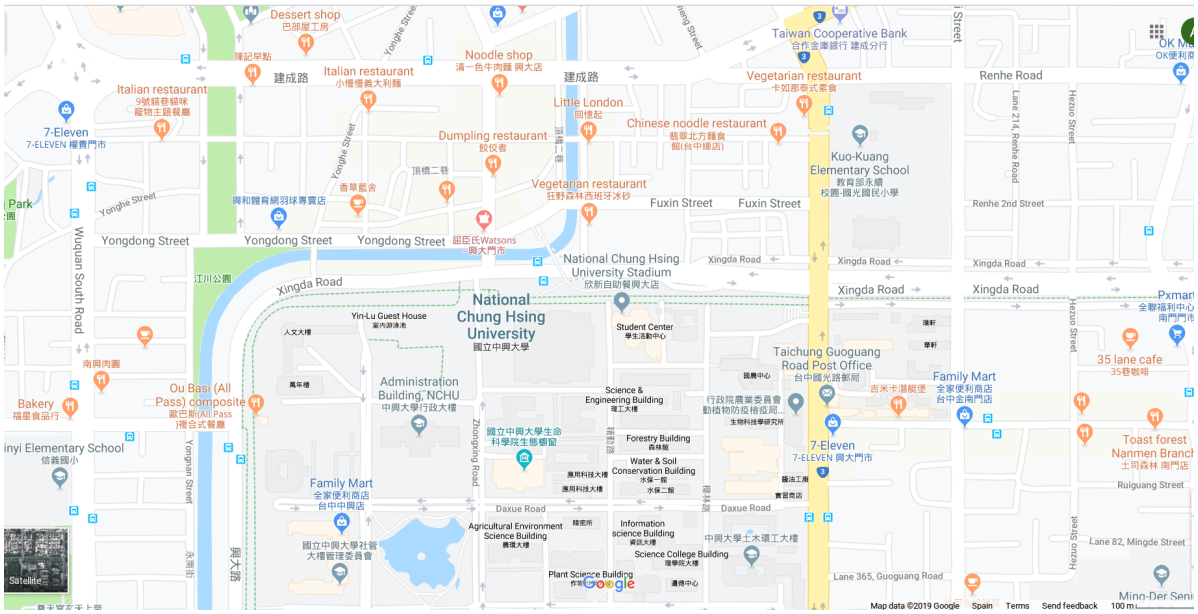


Venue: 3F of the Applied Science and Technology Building

From NCHU to the Conference Dinner Hotel through public transportation

Please note that four private shuttle buses, departing at 17:50, 18:00, 18:10 and 18:20 beside the campus library (see map at page VIII) will drive participants of the Conference Dinner from the NCHU to the Hotel National. The Hotel National is about 15-20 minutes by taxi.

Restaurants outside the campus



List of restaturants at the Student Center (No. 6)

Name (Chinese/English)		Building N. 6	Floor	Type/Cuisine	Spending/ps ($NTD)
小木屋	SHINE MOOD	Student Center	1F	Muffin light meals	>35
	Subway	Student Center	1F	American submarine fort	>45
路意莎	louisa	Student Center	1F	Café and sandwiches	>30
比時地	big steve's	Student Center	1F	American burger	>100
麥味登	My Warm Day	Student Center	1F	Brunch	>35
7-11	7-ELEVEn	Student Center	1F	Convenience Store	>50
林記重慶麻辣燙		Student Center	1F	Mala Tang. braised snacks	>70
利蝂和食		Student Center	1F	Pot Burn Noodles. Donburi	>50
鮮茶道	Presotea Ottawa	Student Center	1F	Hand drink	>25
泰鄉味料理		Student Center	1F	Thai cuisine	>75
潘記天津蔥抓餅		Student Center	1F	Flaky scallion pancake. Douhua. Job's tears milk	>30
欣新自助餐		Student Center	2F	Chinese-food Buffet	>50
五花馬	wu hua ma	Student Center	2F	Dumpling house	>55
素之香		Student Center	2F	Vegetarian cuisine	>50
越南王		Student Center	2F	Vietnamese cuisine	>70
168 町拉麵		Student Center	2F	Japanese Ramen	>75
義拾	EATS	Student Center	2F	Italian noodles	>75
大也牛排		Student Center	2F	steak served on a hot iron plate	>60
隨主滾	Health it	Student Center	2F	Water cooking	>75
台南炒飯		Student Center	2F	Fried rice	>65
DAVID 香港茶水灘		Student Center	2F	Hong Kong-style dim sum	>60
夜市小吃區		Student Center	1F	Night market snacks	>35

PUBLICATION OUTLETS

Econometrics and Statistics (EcoSta)

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

Econometrics and Statistics (EcoSta), published by Elsevier, is the official journal of the networks Computational and Financial Econometrics (CFEnetwork) and Computational and Methodological Statistics (CMStatistics). It publishes research papers in all aspects of econometrics and statistics and comprises two sections:

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

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

The journal consists, preponderantly, of original research. Occasionally, reviews and short papers from experts are published, which may be accompanied by discussions. Special issues and sections within important areas of research are occasionally published. The journal publishes as a supplement the Annals of Computational and Financial Econometrics.

Call For Papers Econometrics and Statistics (EcoSta)

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

Papers containing novel components in econometrics and statistics are encouraged to be submitted for publication in special peer-reviewed, or regular issues of the new Elsevier journal Econometrics and Statistics (EcoSta) and its supplement Annals of Computational and Financial Econometrics.

Call For Papers Computational Statistics & Data Analysis (CSDA)

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

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

Contents

Tuesday 25.06.2019 09:00 - 09:50 Room: S1A03 Chair: Chun-houh Chen Keynote talk 1

Statistical learning of big dependent data

Speaker: **Ruey Tsay, The University of Chicago Booth School of Business, United States**

In the modern big data environment, most real applications are concerned with data that are spatial and/or temporal dependent. For example, in environmental studies, the PM2.5 indexes are typically collected at various monitoring stations over a long period of time. Yet most of the methods for analyzing such data developed in machining learning and statistical literature were derived under the independent assumption. The dynamical dependence in the dependent variable of interest and in the cross-sectional and serial dependence among the predictors may lead to erroneous inference if such dependence is overlooked. We present some methods that can handle big dependent data and provide theoretical justifications for some commonly used methods in the presence of dependent data. Real examples are used throughout to illustrate the effects of dynamical dependence on the traditional methods and to show the gains obtained by the proposed methods.

Tuesday 25.06.2019 17:40 - 18:30 Room: S1A03 Chair: Erricos John Kontoghiorghes Keynote talk 2

Order determination for large-dimensional matrices

Speaker: **Lixing Zhu, Hong Kong Baptist University, Hong Kong**

In sufficient dimension reduction field, a long-standing problem is under-determination of the structural dimension of the central subspace when the criteria that are based on eigendecomposition of target matrices are used. First, due to the existence of some dominating eigenvalues compared to other nonzero eigenvalues, the true dimensionality is often underestimated. Second, the estimation accuracy of any existing method often relies on the uniqueness of minimum/maximum of the criterion. Yet, it is often not the case particularly for the models that converge to a limit with smaller dimensionality. To alleviate these difficulties, we propose a thresholding double ridge ratio criterion. Unlike all the existing eigendecomposition-based criteria, this criterion can define a consistent estimate even when there are several local minima. This generic strategy is readily applied to many fields. As the applications, we give the details about dimension reduction in regressions with fixed and divergent dimensions; about when the number of projected covariates can be consistently estimated, when cannot if a sequence of regression models converges to a limiting model with fewer projected covariates; about ultra-high dimensional approximate factor models.

Thursday 27.06.2019 09:00 - 09:50 Room: S1A03 Chair: Jeroen Rombouts Keynote talk 3

Semi-parametric financial tail risk forecasting

Speaker: **Richard Gerlach, University of Sydney, Australia** Chao Wang

The finding of a class of loss functions that are elicitable for the well-known financial tail risk measures Value at Risk and Expected Shortfall (ES), considered jointly, has allowed some recent advances in the field of semi-parametric tail risk modelling and forecasting, as well as in the formal assessment of those ES forecasts. The aim is to present some of these developments and build upon them through the incorporation of realized measures, the addition of measurement equations and through allowing separate dynamics for the ES equation. Evidence is shown that a Bayesian approach to estimation and forecasting can yield favourable results and an application to financial market returns illustrates that the recent developments, especially when realized measures are included in the models, can generate improvements in the accuracy of forecasts of financial tail risk.

Thursday 27.06.2019 17:40 - 18:30 Room: S1A03 Chair: Tsung-I Lin Keynote talk 4

Recent advances on mixtures of skew distributions for modelling heterogeneous and asymmetric data

Speaker: **Geoffrey McLachlan, University of Queensland, Australia** Sharon Lee

Finite mixtures of skew distributions provide a flexible tool for modelling heterogeneous data with asymmetric distributional features. In recent years, several skew variants of the multivariate normal and t-distributions have been proposed. However, attention has been focused mainly on distributions that are limited to modelling skewness concentrated in a single direction in the feature sample space. We consider a general class of skew distributions that can model various types of skewness and asymmetry in the data, including being able to accommodate multiple directions of skewness. We also consider mixtures of skew factor analyzers for applications to high-dimensional data. The usefulness and potential of the proposed models are demonstrated using real datasets.

Tuesday 25.06.2019 08:45 - 09:00 Room: S1A03 Chair: Wen-Han Hwang Opening

Opening speech

Speaker: **Fuh-Sheng Shieu, NCHU, Taiwan**

Prof. Fuh-Sheng Shieu, President of NCHU, Taiwan, welcomes all participants to EcoSta 2019. In his opening speech, Prof. Shieu will give a concise overview of NCHU, including its history, current status, recent accomplishments and future missions.

Tuesday 25.06.2019 10:20 - 12:25 Parallel Session B – EcoSta2019

EI005 Room UB99(B1) RECENT ADVANCES IN HIGH-DIMENSIONAL DATA ANALYSIS Chair: Wen-Han Hwang

E0157: **On searching for valid instruments in high-dimensional time-series models**
Presenter: **CY Sin**, National Tsing Hua University, Taiwan
With the prevalence of the so-called big data, structural models/equations are often estimated with high-dimensional instruments. That said, research papers in the literature either (1) assumes all instruments are valid and considers an efficient estimator; or (2) proposes some confidence sets of the structural parameters, and investigates their properties under various assumptions on the number of valid instruments. We adopt and modify the OGA-HDAIC approach and search for valid instruments out of some high-dimensional potential instruments. Unlike Lasso, this algorithm is arguably more suitable for time-series data. We close with (i) Some comparisons with the high-dimensional Durbin-Wu-Hausman (DWH) test; and (ii) Some Monte-Carlo simulations.

E0158: **Asymptotic efficiency of Cp-type criterion in high-dimensional multivariate linear regression models**
Presenter: **Shinpei Imori**, Hiroshima University, Japan
Variable selection is one of the crucial problems in statistical fields. Consistency and asymptotic efficiency are known as typical asymptotic properties of selection methods. In univariate linear regression models, Cp criterion is a feasible method from the perspective of prediction because it has the asymptotic efficiency, whereas it does not have the consistency. Recently, the consistency property of Cp-type criterion is shown in high-dimensional multivariate linear regression models, where the dimension of a response matrix is large. We study the asymptotic efficiency of Cp-type criterion under the high-dimensional framework.

E0156: **Model selection for high-dimensional sparse nonlinear models using Chebyshev greedy algorithms**
Presenter: **Ching-Kang Ing**, National Tsing Hua University, Taiwan
Model selection problems in high-dimensional sparse nonlinear models are considered. We first use the Chebyshev greedy algorithm (CGA) to perform variable screening and derive, under a fairly general sparsity condition, its rate of convergence in terms of the number of iterations and the approximation error. We then introduce a high-dimensional information criterion (HDIC) to determine the number of CGA iterations and show that CGA used in conjunction with HDIC achieves the optimal rate of convergence. Finally, the proposed method is applied to the analysis of high-dimensional logistic and Cox regressions.

EO111 Room S101 RECENT ADVANCES IN MACHINE LEARNING Chair: Andreas Christmann

E0224: **Dimension-free error bounds from random projections**
Presenter: **Ata Kaban**, University of Birmingham, United Kingdom
Learning from high dimensional data is challenging in general – however, often the data is not truly high dimensional in the sense that it may have some hidden low complexity geometry. We give new, user-friendly PAC-bounds that are able to take advantage of such benign geometry to reduce dimensional-dependence of error-guarantees in settings where such dependence is known to be essential in general. This is achieved by employing random projection as an analytic tool, and exploiting its structure-preserving compression ability. We introduce an auxiliary function class that operates on reduced dimensional inputs, and a new complexity term, as the distortion of the loss under random projections. The latter is a hypothesis-dependent data-complexity, whose analytic estimates turn out to recover various regularisation schemes in parametric models, and a notion of intrinsic dimension, as quantified by the Gaussian width of the input support in the case of the nearest neighbour rule. If there is benign geometry present, then the bounds become tighter, otherwise they recover the original dimension-dependent bounds.

E0225: **Solving weakly-convex-weakly-concave saddle-point problems as successive strongly monotone variational inequalities**
Presenter: **Tianbao Yang**, University of Iowa, United States
First-order algorithms are considered for solving a class of non-convex non-concave min-max saddle-point problems, whose objective function is weakly convex (resp. weakly concave) in terms of the variable of minimization (resp. maximization). We propose an algorithmic framework motivated by the inexact proximal point method, which solves the weakly monotone variational inequality corresponding to the original min-max problem by approximately solving a sequence of strongly monotone variational inequalities constructed by adding a strongly monotone mapping to the original gradient mapping. The proposed framework is flexible because various subroutines can be employed for solving the strongly monotone variational inequalities. The overall computational complexities of our methods are established when the employed subroutines are subgradient method, stochastic subgradient method, gradient descent method and Nesterov's accelerated method and variance reduction methods for a Lipschitz continuous operator. To the best of our knowledge, this is the first work that establishes the non-asymptotic convergence to a nearly stationary point of a non-convex non-concave min-max problem.

E0334: **Stochastic optimization for AUC maximization in machine learning**
Presenter: **Yiming Ying**, State University of New York at Albany, United States
Stochastic optimization algorithms such as stochastic gradient descent (SGD) update the model sequentially with cheap per-iteration costs, making them amenable for large-scale streaming data analysis. However, most of the existing studies focus on the classification accuracy which can not be directly applied to the important problems of maximizing the Area under the ROC curve (AUC) in imbalanced classification and bipartite ranking. We will present recent work on developing novel SGD-type algorithms for AUC maximization. The new algorithms can allow general loss functions and penalty terms which are achieved through the innovative interactions between machine learning and applied mathematics. Compared with the previous literature which requires high storage and per-iteration costs, our algorithms have both space and per-iteration costs of one datum while achieving optimal convergence rates.

E0489: **Parameter-free machine learning through coin betting**
Presenter: **Francesco Orabona**, Boston University, United States
Machine Learning (ML) has been described as the fuel of the next industrial revolution. Yet, despite their name, the majority of the ML algorithms still heavily rely on having humans in the loop in order to set their "parameters". For example, when using regularized empirical risk minimization, the choice of the weight of the regularizer is critical to obtain theoretical and practical optimal performance. Moreover, the minimization itself, usually done through stochastic gradient descent (SGD), requires to set "learning rates" in order to get good performance. Are these parameters strictly necessary? Is it possible to have "parameter-free" ML algorithms? It will be shown that many ML problems can be reduced to a game of betting on a non-stochastic coin. Betting on a non-stochastic coin is a well known problem whose optimal strategy turns out to be a simple generalization of the Kelly betting criterion. Moreover, the optimal coin betting algorithm is parameter-free, giving rise to parameter-free ML and stochastic optimization algorithms. For example, this approach gives: 1) optimal rates of convergence in RKHS in the capacity independent setting without any parameter to tune; 2) a differentially private SGD without learning rates; 3) a new way to obtain finite-time iterated-logarithm martingale concentrations in Banach spaces.

E0423: **Robustness of localized learning**
Presenter: **Andreas Christmann**, University of Bayreuth, Germany
The computation of kernel methods is usually not fast enough for big data sets. A localized learning method based on kernels is investigated. Recent results on universal consistency and on statistical robustness properties of such localized learning methods will be given.

EO213 Room S102 BIOSTATISTICS: THEORY AND METHODS Chair: Chien-Ju Lin

E0580: **Evaluating the validity and reliability of multi item scales after multiple imputation**
Presenter: **Oya Kalaycioglu**, Bolu Abant Izzet Baysal University, Bolu, Turkey
Various multiple imputation (MI) methods for handling missing items in multi-item scales were evaluated based on the real data collected from a questionnaire consists of 39 Likert type items and four subscales. For different MI strategies at item, subscale and scale levels, the mean of the sub-scale scores were compared with simulation studies using the bias and coverage as the performance parameters. Additionally, commonly used measures to ensure validity and reliability of the multi-item scales were assessed. MI of each item separately outperformed in terms of bias and coverage of the mean sub-scale scores, as well as the validity and reliability measures. When the number of incomplete items was too large overfitting problems occurred with this method, therefore two different techniques were proposed to reduce the number of predictors in the imputation model. First, the predictors in the imputation model were selected with forward selection approach and second, rather than using item scores, sub-scale scores were used as predictors when imputing an item. All methods were most sensitive when missing data were not at random.

E0233: **Efficient estimation of a semiparametric zero-inflated Bernoulli regression model**
Presenter: **Chin-Shang Li**, University at Buffalo, United States
When the observed proportion of zeros in a data set consisting of binary outcome data is larger than expected under a regular logistic regression model, it is frequently suggested using a zero-inflated Bernoulli (ZIB) regression model. A spline-based ZIB regression model is proposed to describe the potentially non-linear effect of a continuous covariate. A spline, which can be expressed as a linear combination of B-spline basis functions, is used to estimate the unknown smooth function. The spline estimator of the nonparametric component is shown to be uniformly consistent and achieve the optimal convergence rate under the smoothness condition. The regression parameter estimators are shown to be asymptotically normal and efficient. A spline-based semiparametric likelihood ratio test is established, and a direct and consistent variance estimation method based on least-squares method is proposed. Extensive simulations are conducted to evaluate the finite-sample performance of the proposed method. A real-life data set is used to illustrate the practical use of the proposed methodology.

E0364: **Second-order estimating equations for clustered current status data from family studies using biased sampling**
Presenter: **Yujie Zhong**, Shanghai University of Finance and Economics, China
Co-authors: Richard Cook
Studies about the genetic basis for disease are routinely conducted through family studies under response-dependent sampling in which affected individuals are sampled, along with relatives providing current status information on disease onset times. We develop conditional second-order estimating equations for studying the nature and extent of within-family dependence which recognizes the biased sampling scheme employed in family studies and the current status data provided by the relatives. Simulation studies are carried out to evaluate the finite sample performance of different estimating functions and to quantify the empirical relative efficiency of the various methods. Sensitivity to model misspecification is also explored. An application to a motivating psoriatic arthritis family study is given for illustration.

E0630: **Nonparametric clustering approach for longitudinal cognitive measurements, baseline imaging and genetic data**
Presenter: **Brian Tom**, MRC Biostatistics Unit, United Kingdom
Co-authors: Anais Rouanet
Dementia is one of the most challenging global health problems of the 21st century, affecting over 47 million people globally with numbers expected to rise substantially over the next thirty years. At present there is a high failure rate for treatments tested for Alzheimers dementia. This may be due to treatments being tested on those who already have irreparable brain damage. Identifying persons early in disease through use of biomarkers may increase the likelihood that treatments will be more effective in slowing or arresting further progression of the disease. Clustering approaches provide a powerful means of profiling at-risk populations over time. We have developed a Bayesian Dirichlet process mixture model linking non-parametrically a longitudinal outcome and baseline variables which allows clustering structure within a heterogeneous population to be uncovered. It flexibly models the longitudinal outcome through cluster-specific Gaussian process priors and allows the number of clusters to vary through the use of a Dirichlet process prior. The methodology is applied to the ADNI cohort to identify typical profiles of subjects at high risk of dementia using longitudinal cognitive measurements, baseline socio-demographics, imaging and genetic data. Four clusters of subjects, including two with steep cognitive decline profiles, were obtained.

E0318: **A variance component score test applied to RNA-Seq differential analysis**
Presenter: **Boris Hejblum**, Bordeaux University Inria/Inserm Vaccine Research Institute, France
Co-authors: Marine Gauthier, Rodolphe Thiebaut, Denis Agniel
Gene expression measurement has shifted from microarrays to next generation RNA-sequencing, producing ever richer high-throughput data for transcriptomics studies. As such studies grow in size, frequency, and importance, there is an urgent need for statistical methods that better control the type-I error. We model transformed RNA-seq counts as continuous variables using nonparametric regression to account for their inherent heteroscedasticity, in a principled, model-free, and efficient manner. We rely on a powerful variance component score test that can deal with both covariates adjustment and data heteroscedasticity to identify the genes whose expression is significantly associated with one or several factors of interest in complex experimental designs (including longitudinal data). Our test statistic has a simple form and limiting distribution, which can be computed quickly. A permutation version of the test is also derived for small sample sizes. We show that our test has very good statistical properties in simulations, with an increase in stability and power when compared to state-of-the-art methods limma/voom, edgeR, and DESeq2. In particular, we show that those three methods can all fail to control the type I error when the sample size becomes larger, while our method behaves as expected. We apply our proposed method to two public datasets: one with repeated measurements investigating a candidate vaccine against EBOLA, and one studying tuberculosis.

EO015 Room S104 RECENT ADVANCES IN STATISTICAL METHODS FOR HIGH-DIMENSIONAL DATA Chair: Jingjing Wu

E0398: **Genetic association between amyotrophic lateral sclerosis and cancer**
Presenter: **Hsiuying Wang**, National Chiao Tung University, Taiwan
Amyotrophic lateral sclerosis (ALS) is a fatal neurodegenerative disease. An ALS drug, Riluzole, has been shown to induce two different anticancer effects on hepatocellular carcinoma (HCC). In light of this finding, we explore the relationship between ALS and cancer, especially for HCC, from the molecular biological viewpoint. We establish biomarkers that can discriminate between ALS patients and healthy controls. A principal component analysis (PCA) based unsupervised feature extraction (FE) is used to find gene biomarkers of ALS based on microarraygene expression data. Based on this method, 101 probes were selected as biomarkers for ALS with 0.95 high accuracy to discriminate between ALS patients and controls. Most of the genes corresponding to these probes are shown to be related to various cancers. These findings might provide a new insight for developing new therapeutic options or drugs for both ALS and cancer.

E0401: **Model checking for the single-index model: A residual decomposition approach**
Presenter: **Yih-Huei Huang**, Tamkang University, Taiwan
The single-index model is a popular semiparametric model. It has the flexibility of nonparametric regression and avoids the curse of dimensionality by adopting a single-index, instead of the whole covariate, as the predictor in the regression model. It has been widely used and a few diagnostic tools or goodness of fit tests had been developed for model checking. Nevertheless, there are drawbacks for traditional approaches. Existent tests could be complicated to implement, require quite an amount of computation or not powerful for certain alternative hypothesis. Motivated by these drawbacks. We decomposed the residual and found the conditions when the test statistics is asymptotic distribution free. It is easy in implementation, require no bootstrapping and has good power performance according to a small simulation study.

E0552: **Clustering microbiome data using finite mixture of Dirichlet-multinomial regression models**
Presenter: **Zeny Feng**, University of Guelph, Canada
The human microbiome is a fundamental component of our physiology, and exploring the relationship between biological/environmental covariates and the resulting taxonomic composition of a given microbial community is an active area of research. The advancement of biology techniques, allow us to sequence the high throughput microbial metagenomic with an affordable cost, such that microbiome data is available and accessible for the study. Previously, a Dirichlet-multinomial regression framework has been suggested to model this relationship, but it did not account for any latent group structure which has been observed across microbiome samples which share similar biota compositions (such as enterotypes). A finite mixture of Dirichlet-multinomial regression models is proposed and illustrated in order to account for this group structure and to allow for a probabilistic investigation of the relationship between bacterial abundance and biological/environmental covariates within each inferred group. Furthermore, finite mixtures of regression models which incorporate the concomitant effect of the covariates on the resulting mixing proportions are also proposed and examined within the Dirichlet-multinomial framework.

E0590: **Machine learning classification of functional brain imaging for Parkinsons disease stage prediction**
Presenter: **Guan-Hua Huang**, National Chiao Tung University, Taiwan
The aim is to analyze a dataset containing functional brain imaging from 6 normal healthy controls and 196 patients with Parkinson's disease (PD), which can be divided into 5 stages according to the severity of illness. The goal is to predict patients' PD illness stages via their functional brain images. Used approaches include multivariate statistical methods, ensemble learning models, and deep convolutional neural network (CNN). For statistical and ensemble models, PCA is performed to extract features, and the best combination of parameters is found by grid search. For CNN modeling, we use the technique of image augmentation to increase data size and build the model by the architecture of VGG16. It is found that the deep learning VGG16 model outperforms other approaches, which can capture significant features from imaging and reach higher classification accuracy.

E0710: **Group and individual variable selection in semiparametric transformation models**
Presenter: **Jingjing Wu**, University of Calgary, Canada
Co-authors: Wenyan Zhong, Xuewen Lu
The bi-level variable selection is investigated in semiparametric transformation models for right-censored data. The class of transformation models under consideration includes the proportional hazards model and the proportional odds model as special cases and has the capability to accommodate external time-varying covariates. In the framework of regularized regression, we propose a computationally efficient estimation method that selects significant groups and variables simultaneously. Group bridge, adaptive group bridge and composite group bridge penalties which can integrate grouping structure of covariates were adopted for bi-level variable selection purpose. We illustrate the finite sample performance of the proposed methods via simulations and two real data examples.

EO125 Room S106 NEW ADVANCES IN STATISTICAL COMPUTING AND COMPLEX DATA ANALYSIS Chair: Weixin Yao

E0240: **Gaussian-based visualization of Gaussian and non-Gaussian model-based clustering**
Presenter: **Christophe Biernacki**, Inria, France
Co-authors: Matthieu Marbac-Lourdelle, Vincent Vandewalle
A generic method is introduced to visualize in a Gaussian-like way, and onto R^d, results of Gaussian or non-Gaussian model-based clustering. The key point is to explicitly force a spherical Gaussian mixture visualization to inherit from the within cluster overlap which is present in the initial clustering mixture. The result is a particularly user-friendly draw of the clusters, allowing any practitioner to have an overview of the potentially complex clustering result. An entropic measure allows us to inform of the quality of the drawn overlap, in comparison to the true one in the initial space. The proposed method is illustrated on four real data sets of different types (categorical, mixed, functional and network) and is implemented on the R package ClusVis.

E0360: **Simultaneous confidence bands for partially linear single-index models for longitudinal data**
Presenter: **Suojin Wang**, Texas A and M University, United States
Co-authors: Li Cai, Lei Jin
An asymptotically accurate simultaneous confidence band (SCB) is proposed for the nonparametric link function in the partially linear single-index models for longitudinal data under general conditions including both unbalanced and non normal cases. To formulate such a band, a two-step semiparametric estimation method combing local linear smoothing and generalized estimating equations is adopted to estimate both the link function and the parametric components. The estimator for the link function is shown to be oracally efficient, in the sense that it is asymptotically equivalent to that with all true values of the parameters being known oracally. Furthermore, using the oracle efficiency the asymptotic distribution of the maximal deviation of the two-step estimator is provided, and hence an SCB for the link function is constructed. Finite-sample simulation studies are carried out which support our asymptotic theory. The proposed SCB is applied to a CD4 data set to analyze and test the trend of the CD4 cell numbers.

E0550: **Permutation test based on clustered data from a rotating sample plan**
Presenter: **Jiahua Chen**, University of British Columbia, Canada
A classical problem in mathematical statistics is the hypothesis test. Given a data set, we wish to decide whether or not the distribution behind the data violates the model structure of interest. Such a simple task may demand complex solutions when a realistic yet comprehensive model is hard to find. In an applied project, we have observations on samples from several connected populations. Due to a rotating sampling plan,random effects are suggested in its longitudinal direction as well as in cross-sectional respect. Besides, strong parametric model assumptions should be discouraged. It is difficult to quantify or model these random effects. The asymptotic theory becomes hard to develop, and therefore a good approximation hard to find for the distribution of the test statistics. We develop a permutation scheme to the symmetric in the data structure. The resulting test, therefore, automatically has the right size. Combined with a semi-parametric density ratio model and the composite likelihood approach, the proposed tests are found to work well for the targeted applications.

E0616: **Variable selection and estimation in generalized linear models with measurement error**
Presenter: **Liqun Wang**, University of Manitoba, Canada
Co-authors: Lin Xue
The variable selection problem in linear and generalized linear models is studied when some of the predictors are measured with error. We demonstrate through numerical examples how measurement error (ME) affects the selection results and propose a regularized instrumental variable (RIV) method to correct for the ME effects. We show that the proposed estimator has the oracle property in a linear model and we derive its asymptotic distribution under general conditions. We also investigate the performance of this method in generalized linear models. Our simulation studies show that the RIV estimator outperforms the naive estimator in both linear and some generalized linear models. Finally, the proposed method is applied to a real dataset.

E0687: **Maximum Lq-likelihood estimation for the mixture of dynamic covariance models**
Presenter: **Lin Xu**, Zhejiang University of Finance & Economics, China
Co-authors: Weixin Yao
A robust estimation is proposed for the mixture of dynamic covariance models based on the maximum Lq-likelihood inference procedure. The model is shown to be identifiable, and can be estimated robustly by a modified EM-type algorithm. Via the constrained quadratic optimization, a within-subject tuning parameter selection criterion is constructed. Additionally, We derive the asymptotic property of the maximum Lq-likelihood estimates. A simulation study and two real data examples are conducted to evaluate the finite sample performance of the proposed methodology.

EO025 Room S1A01 RECENT ADVANCES IN FUNCTIONAL DATA ANALYSIS Chair: Jeng-Min Chiou

E0213: **Wasserstein covariance for vectors of random densities**
Presenter: **Hans-Georg Mueller**, University of California Davis, United States
Co-authors: Alexander Petersen
Samples of data that consist of probability densities or distributions are encountered in various applications. Once a metric for densities is specified, the Frechet mean or barycenter is typically used to determine the average density. The Wasserstein metric is popular due to its good performance in applications and interpretive value as an optimal transport metric. Motivated by applications in neuroimaging, we consider data that consist of a p-vector of univariate random densities for each sampling unit. We introduce Wasserstein covariance to quantify the dependency of the component densities and provide corresponding estimators for fixed and diverging p, where the latter corresponds to continuously-indexed densities. Consistency and asymptotic normality are established, while accounting for errors introduced in the unavoidable preparatory density estimation step. The utility of the Wasserstein covariance matrix is demonstrated in applications that include functional connectivity in the brain and the secular evolution of mortality.

E0579: **Regularized subspace clustering for functional data**
Presenter: **Yoshikazu Terada**, Osaka University; RIKEN, Japan
Co-authors: Michio Yamamoto
The intrinsic high dimensional nature of functional data often makes possible the very good performance in supervised classification for functional data. In the supervised classification problems, it is known that, using the projection into the finite-dimensional subspace, we can extract the intrinsic high dimensional nature from functional data. In the context of unsupervised classification, there are several clustering methods based on the projection into the subspace. However, the projected data do not necessarily reflect the hidden true cluster structure in these existing methods. We propose a new regularized subspace clustering algorithm for functional data based on a cluster-separation criterion in the finite-dimensional subspace. The proposed algorithm monotonically decreases the objective function value. Moreover, we study asymptotic properties of our clustering method. The proposed method works well not only for the simulated data, but also for the real datasets which are difficult to obtain a good classification performance.

E0675: **Domain selection for functional linear models: A dynamic RKHS approach**
Presenter: **Jane-Ling Wang**, University of California Davis, United States
Co-authors: Shu-Chin Lin
In conventional scalar-on-function linear regression model, the entire trajectory of the predictor process on the whole domain is used to model the response variable. However, the response may only be associated with the covariate process X on a subdomain. We consider the problem of estimating the domain of association when assuming that the regression coefficient function is nonzero on a subinterval. This problem was first considered few years ago, and the difficulty in estimating the domain has been pointed out. We resolve this through a two-steps procedure to estimate the unknown components, where in the first step we estimate the domain based on the reproducing kernel Hilbert space (RKHS) approach and in the second step the regression function is estimated. We motivate the two-step procedure and show that it provides consistent estimator for the domain under mild smoothness assumptions. A simulation study illustrates the effectiveness of the proposed approach.

E0803: **Prediction with robust mixtures of Gaussian local mapping**
Presenter: **Naisyin Wang**, University of Michigan, United States
The uses of mixtures provide a powerful tool to simplify the task of modeling complicated associations between variables. Data with similar associations are grouped together and simple estimation techniques can be applied on each cluster. This strategy also allows flexibility in distinguishing structures of interest. However, recovering the appropriate structure embedded in mixture is not a trivial task under certain circumstances. The presence of outliers might have severe impacts on forming the suitable clusters. Thus, robust consideration is particularly important under non-standard scenarios. We propose a robust mixture regression approach coupling with the use of trimmed likelihood to established structured functional-regression modeling of scalar responses and functional predictors. Our modeling strategies form a family of assumed models. Model-averaging, model-selection and their hybrids are then naturally incorporated under a unified framework into a modified Expectation-Maximization algorithm. The outcomes provide the estimated parameters in the most favorable assumed model that satisfies pre-determined criteria and then

is used for prediction. We provide theoretical justifications behind the proposed procedures and illustrate their numerical performances using synthetic and real-world data sets.

E0181: **New tests for equality of several covariance functions for functional data**
Presenter: **Jin-Ting Zhang**, National University of Singapore, Singapore
Two new tests for the equality of the covariance functions of several functional populations, namely a quasi GPF test and a quasi Fmax test, are obtained via Globalizing a Point-wise quasi F-test statistic with integration and taking its supremum over some time interval of interest, respectively. Unlike several existing tests, they are scale-invariant. The asymptotic random expressions of the two tests under the null hypothesis and a local alternative are derived. It is showed that under some mild conditions, the asymptotic null distribution of the quasi GPF test is a chi-squared-type mixture whose distribution can be well approximated by a simple scaled chi-squared distribution. We also propose a random permutation method for approximating the null distributions of the quasi GPF and Fmax tests. Simulation studies are presented to demonstrate the finite-sample performance of the new tests against five existing tests. An illustrative example is also presented.

EO155 Room AT241 RECENT ADVANCES IN COMPUTATION FOR STATISTICAL LEARNING Chair: Eric Chi

E0216: **A penalty method for variance component selection**
Presenter: **Hua Zhou**, UCLA, United States
Co-authors: Juhyun Kim, Jin Zhou
Variance components models, also known as mixed effects model, are a central theme in statistics. When there are a large number of variance components, one wants to select a subset of those that are associated with response. Existing methods are limited to finding random components at individual level or within one variance component. We propose a selection of variance components based on a penalized log-likelihood with adaptive penalty. This is solved by a majorization-minimization (MM) algorithm, which is simple, numerically stable, and globally convergent. The performance of the proposed methodology is evaluated empirically through simulation studies and real data analysis. In theory, we establish a non-asymptotic error bound for the iterates from the algorithm and characterize the region in which the MM iterates converge to a global optimum of the population likelihood. This result provides a theoretical guideline in terminating MM iterations.

E0284: **Computational techniques for modeling non-life insurance claims**
Presenter: **Yi Yang**, McGill University, Canada
Co-authors: Wei Qian, Hui Zou
Tweedies Compound Poisson model is a popular method to model data with probability mass at zero and non-negative, highly right-skewed distribution. Motivated by wide applications of the Tweedie model in various fields such as actuarial science, we investigate a grouped elastic net method and a boosted nonparametric method for the Tweedie model in the context of the generalized linear model. For the grouped elastic net method, in order to efficiently compute the estimation coefficients, we devise a two-layer algorithm that embeds the blockwise majorization descent method into an iteratively re-weighted least square strategy. Together with the strong rule, the proposed algorithm is implemented in an easy-to-use R package HDtweedie, and it is shown to compute the whole solution path very efficiently. On the other hand, the linear form of the logarithmic mean in the Tweedie GLM sometimes can be too rigid for many applications. As a better alternative, we propose a boosted nonparametric Tweedie model for pure premiums and use a profile likelihood approach to estimate the index and dispersion parameters. To our knowledge, there is no existing nonparametric Tweedie method available before. Our method is capable of fitting a flexible nonlinear Tweedie model and capturing complex interactions among predictors. We have also implemented this method in a user-friendly R package that includes a nice visualization tool for interpreting the fitted model.

E0331: **Hierarchical community detection by recursive bi-partitioning**
Presenter: **Tianxi Li**, University of Virginia, United States
The problem of community detection in networks is usually formulated as finding a single partition of the network into some "correct" number of communities. We argue that it is more interpretable and in some regimes more accurate to construct a hierarchical tree of communities instead. This can be done with a simple top-down recursive bi-partitioning algorithm, starting with a single community and separating the nodes into two communities by spectral clustering repeatedly, until a stopping rule suggests there are no further communities. This class of algorithms is model-free, computationally efficient, and requires no tuning other than selecting a stopping rule. We show that there are regimes where it outperforms K-way spectral clustering, and propose a natural model for analyzing the algorithm's theoretical performance, the binary tree stochastic block model. Under this model, we prove that the algorithm correctly recovers the entire community tree under relatively mild assumptions. We also apply the algorithm to a dataset of statistics papers to construct a hierarchical tree of statistical research communities.

E0485: **A novel ADMM algorithm for graph-fused lasso**
Presenter: **Teng Zhang**, University of Central Florida, United States
A new algorithm is proposed for solving the graph-fused lasso (GFL), a method for parameter estimation that operates under the assumption that the signal tends to be locally constant over a predefined graph structure. The proposed method applies alternating direction method of multiplier (ADMM), which is based on the decomposition of the objective function into two components. While ADMM has been widely used in this problem, existing works decompose the objective function into the loss function component and the total variation penalty component. In comparison, the objective function is proposed to be decomposed into two parts, where one part is the loss function with some total variation penalty, and the other part is the remaining total variation penalty. Compared with existing works, this method has a smaller computational cost per iteration and fewer iterations to convergence in many settings. Experiments on artificial and real data sets confirm the competitive performance of our method.

E0346: **Network augmented classification**
Presenter: **Ji Zhu**, University of Michigan, United States
In classical classification, a data point is classified given its individual covariates. Often, additional network information describing the connectivity relationships between points are also available, which in principle can be used to improve classification performance. We develop a general statistical framework for network augmented classification. Under this framework, we derive the optimal Bayes classifiers for two general families of distributions incorporating both covariates and networks, one being generative and the other being discriminative. Further, we establish consistency results for plug-in classifiers with respect to the optimal classifiers under the generative and discriminative families,respectively. We also apply the general approaches to two specific models and propose two effective classification methods for practical use. The proposed methods have been evaluated using both simulation studies and real-world data examples, and the results are promising.

EO055 Room AT242 STATISTICAL METHODOLOGIES AND APPLICATIONS IN BUSINESS AND INDUSTRY Chair: Wendy Lou

E0208: **Process variation monitoring using a loss-function control chart**
Presenter: **Su-Fen Yang**, National Chengchi UNiversity, Taiwan, Taiwan

The quality and loss of products are crucial factors separating competitive companies in global market. Firms widely employ a loss function to measure the loss caused by poor quality. From the view point of Taguchis philosophy, monitoring the deviation from the process target value is important. In reality, there are many situations where the distribution of the quality variable may not be normal. The aim is to develop a new loss-function control chart for monitoring process variation under non-normal distributions. The properties and out-of-control detection performance of the new loss-function control chart are investigated. Furthermore, an adaptive control scheme for the proposed loss-function control chart is considered. Numerical results show the out-of-control detection performance of the proposed loss-function control chart. Keywords: Loss function, control chart, non-normal distribution.

E0246: **The empirical estimator of the boundary in an inverse firstexit problem**
Presenter: **Klaus Poetzelberger**, WU Vienna, Austria

First-passage problems for the Brownian motion (W_t) or general diffusion processes, have important applications. Given a boundary $b(t)$, the distribution of the first-exit time τ has to be computed, in most cases numerically. Inverse boundary crossing probabilities assume that the distribution of τ is given and the boundary b has to be found. The analysis is based on the fact that the boundary and the density of τ satisfy a Volterra integral equation. We propose and analyze estimators of b, when a sample $\tau_1,\ldots,\tau_n$ of first exit times is given. The first class of estimators are solutions of stochastic versions of the Volterra equation. The second class of estimators are approximate likelihood methods, using the idea of approximating the boundary $b(t)$ by a piecewise boundary $b_m(t)$. Define $W^m = (W_{t_1},\ldots,W_m)$. The density of τ for b_m conditional on $W^m = w^m$ is available in closed form. The Bayesian estimator chooses a prior on b and then uses Gibbs sampling to iterate the generation of $b \mid (W^m,\tau_1,\ldots,\tau_n)$ and $W^m \mid (b,\tau_1,\ldots,\tau_n)$. Typical inverse problems are sequential testing in statistics or the estimation of a ruin boundary, for instance in credit risk modelling. A company defaults if a process (V_t), called the value of the firm, crosses a boundary $b(t)$. (V_t) cannot be observed. It is correlated with (S_t), which includes published relevant information on (V_t).

E0611: **A Markov chain approach for control charts in a double-sampling scheme**
Presenter: **Hsing-Ming Chang**, National Cheng Kung University, Taiwan

A tool commonly used in many industries to monitor the quality of something as the result of a process is by devising a control chart for a statistic observed in the process over time. Keep in mind that, in practice or application, the so-called "control limits" may not be chosen merely due to the behaviour of a statistic described by its distribution but the specs required or expected by a customer or an experienced engineer. In view of this, a Markov chain approach is proposed as a general method to study the properties of control charts with zone rules for a double-sampling scheme.

E0638: **A hierarchical process for composite indices**
Presenter: **Wendy Lou**, University of Toronto, Canada

Quantifying a multifaceted concept for analytical purposes often involves multiple data sources with attributes requiring input from domain experts whose opinions likely vary across domains (e.g environmentalist, health care provider, urban planner etc.). Motivated by an ongoing Canadian urban environmental health research project, a composite urban green index is developed based on a hierarchical approach with fuzzy weights that incorporate expert opinions. The methodology will be described first, followed by simulations and numerical comparisons to illustrate the practical applications.

E0839: **First passage time distribution of jump-diffusion processes and nonlinear boundaries**
Presenter: **Zhiyong Jin**, University of Manitoba, Canada
Co-authors: Liqun Wang

First passage time (FPT) model of jump-diffusion processes is a very useful tool in finance and insurance, neuroscience and other scientific disciplines. The calculation of the FPT distribution for diffusion processes is a long-standing and notoriously difficult problem. While it is well-known that explicit formulas exist only for some special processes and boundaries, the problem is even more challenging for processes with jumps. We derive new formulas for piecewise linear boundary crossing probabilities and first passage time densities of Brownian motion with random jumps where the jump process can be any integer-valued counting process and jump sizes can be correlated and non-identically distributed. These formulas can be used to approximate the first passage time distributions for general nonlinear boundaries. The method can be extended to more general diffusion processes such as geometric Brownian motion and Ornstein-Uhlenbeck processes with jumps. The numerical computation is done through Monte Carlo integration which is straightforward and easy to implement. Some numerical examples are presented.

EO151 Room U301 FINANCIAL MATHEMATICS AND STATISTICS Chair: Hoi Ying Wong

E0378: **Open-loop equilibrium strategy for mean-variance portfolio problem under stochastic volatility**
Presenter: **Tingjin Yan**, The Chinese University of Hong Kong, China
Co-authors: Hoi Ying Wong

The open-loop control framework for time-consistent mean-variance (TCMV) portfolio problems is formulated in incomplete markets with stochastic volatility (SV). We offer the existence and uniqueness results of the TCMV equilibrium controls for general SV models and derive explicit closed-form equilibrium controls for several popular models, including the Heston, Hull-White and 3/2 SV models. The uniqueness of the equilibrium controls are related to the mean-reverting speed of the volatility and the investment horizon.

E0379: **Stochastic volatility asymptotics for optimal subsistence consumption and Investment with bankruptcy**
Presenter: **Kexin Chen**, The Chinese University of Hong Kong, Hong Kong
Co-authors: Hoi Ying Wong, Mei Choi Chiu, Yong Hyun Shin

Subsistence consumption and investment problems with bankruptcy are classic constrained stochastic optimal control problem in financial economic, where the consumption rate should be greater than a positive number and the investor faces a bankruptcy payment. We derive novel asymptotic solution to this problem under the fast mean-reverting stochastic volatility model. We rigorously prove that the zeroth-order approximation for the optimal pair of consumption and investment strategies leads to the first-order accuracy of the objective function. In addition, this zeroth-order sub-optimal consumption-investment pair is asymptotically optimal in a class of admissible trading strategy pairs.

E0468: **Consumption ratcheting and loss aversion**
Presenter: **Junkee Jeon**, Kyung Hee University, Korea, South
Co-authors: Kyoung Jin Choi, Hyeng Keun Koo

The aim is to investigate the optimal consumption, portfolio selection, and risk attitude of an economic agent who faces partial irreversibility of consumption decisions, formalizing a previously proposed theory. The irreversibility generates consumption ratcheting and dynamic loss aversion. We derive optimal policies and a measure of risk aversion implied by the optimal portfolio in closed form. The optimal consumption policy involves

an inaction interval for the consumption-wealth ratio; when the ratio is inside the interval it is optimal not to adjust consumption, and when the ratio is outside the interval, it is optimal to adjust consumption immediately to restore the ratio to the nearest boundary of the interval. In particular, we disentangle the effects of loss aversion from those of risk aversion, and show that loss aversion determines the frequency of consumption adjustments and the shape of the risky share inside the interval, which provides various novel implications for consumption decisions and risk attitudes. We also provide an extension of our model to that with durable or multiple goods.

E0477: **Optimal execution with regime-switching market resilience**
Presenter: **Chi Chung Siu**, The Hang Seng University of Hong Kong, Hong Kong
Co-authors: Ivan Guo, Song-Ping Zhu, Robert Elliott
The optimal placement of market orders in a limit order book (LOB) market is studied when the market resilience rate, which is the rate at which market replenishes itself after each trade, is stochastic. More specifically, we establish a tractable extension to the optimal execution model by modelling the dynamics of the resilience rate to be driven by a Markov chain. When the LOB replenishes itself stochastically through time, the optimal execution strategy becomes state-dependent, and is driven linearly by the current remaining position and the current temporary price impact, with their linear dependence based on the expectation of the dynamics of future resilience rate. A trader would optimally place more aggressive (respectively, conservative) market orders when the limit order book switches from a low to a high resilience state, (respectively, from a high to a low resilience state). Our cost saving analysis indicates that the incremental execution costs can be substantial when the agent ignores the stochastic dynamics of the market resilience rate by adopting the state-independent strategies.

E0606: **Mean-variance equilibrium asset-liability management strategy with cointegrated assets**
Presenter: **Mei Choi Chiu**, The Education University of Hong Kong, Hong Kong
Asset-liability management (ALM) problems are investigated in a continuous-time economy. When the financial market consists of cointegrated risky assets, institutional investors, on one hand, attempt to make profit from the cointegration feature and, on the other hand, needs to maintain a stable surplus level, that is, the company's wealth less its liability. Challenges occur when the liability is random and cannot be fully financed or hedged through the financial market. For mean-variance (MV) investors, an additional concern is the rational time-consistent (TC) issue which ensures the decision made in the future would not be restricted by the current surplus level. By putting all these ingredients together, the closed-form feedback equilibrium control the TC-MV ALM problem with cointegrated risky assets is derived. The solution procedure is built upon the HJB framework addressing time inconsistency.

EO241 Room U414 ADVANCES IN FINANCIAL TIME SERIES ANALYSIS Chair: Toshiaki Watanabe

E0276: **Bayesian network analysis of systemic risk in financial markets**
Presenter: **Thomas Chan**, Hong Kong University of Science and Technology, Hong Kong
Co-authors: Mike So
Analyzing systemic risk in financial markets has been an active research area in financial econometrics, risk management, and big data analytics. An approach based on network analysis is proposed to study the interrelationship between financial companies. We develop statistical models to understand how the financial network, and thus systemic risk, changes over time. We adopt Bayesian inference methods to estimate the financial network, do network prediction and use listed companies in Hong Kong to illustrate our idea.

E0430: **Realized jump beta: Evidence from high-frequency data on Tokyo stock exchange**
Presenter: **Masato Ubukata**, Meiji Gakuin University, Japan
The focus is on jump betas of some sector portfolios in the Japanese stock market and to assess the dynamics in realized jump betas estimated from the high-frequency data. We test the null hypothesis of constant jump betas over years and months. The results show that annual constant jump betas are always rejected at 1% significance level, while monthly constant jump betas are not rejected for about a half of the monthly samples. Given the worldwide evidence of fractional integration in realized variance and covariances, the estimation results under the assumption of a pure fractional noise process of the monthly jump betas indicate a small average degree of integration, namely ARFIMA(0, 0.2, 0). However, a further analysis implies that the monthly realized jump betas are arguably modeled as stationary I(0) processes.

E0434: **Investigating the interaction between returns and order flows: Endogeneity, intraday variations, and macro announcements**
Presenter: **Makoto Takahashi**, Hosei University, Japan
The aim is to examine the interaction between returns and order flow imbalances (differences between buy and sell orders), constructed from the best bid and offer files of S&P 500 E-mini futures contract, using a structural vector autoregressive (SVAR) model. The well-known intraday variation in market activity is considered by applying the SVAR model for each short interval each day, whereas the endogeneity due to time aggregation is handled by estimating the structural parameters via the identification through heteroskedasticity. The estimation results show that significant endogeneity exists and that the estimated parameters and associated quantities, such as the return variance driven by order flow imbalances, vary over time, reflecting intense or mild order submission activities. Further, order flow imbalances are shown to be more informative several minutes away from macroeconomic news announcements and that inactive order submission periods exist when they occur.

E0465: **Estimation of smoothly time varying coefficient partial adjustment model**
Presenter: **Kosuke Oya**, Osaka University, Japan
The partial adjustment model can capture how financial asset prices adjust to information. Speeds of adjustment towards intrinsic values of asset prices are the measures of the degrees of over and under-reactions for new information. The conventional partial adjustment model has a constant coefficient which reflects the adjustment speed. Although we can see the degrees of adjustments speed through the estimated coefficients, it is hard to see how adjustment speeds change during the estimation period. To cope with the difficulty, we introduce the partial adjustment model with smoothly time varying coefficient. The time varying scheme is established through the same way as the smooth transition autoregressive model developed in the previous studies. For empirical application, the price discovery and tatonnement process during the preopening period in the Tokyo stock exchange is focused. Using the time varying adjustment coefficient, we can find whether market discovers the equilibrium price of asset until the market opens.

E0507: **News implied volatility and aggregate economic activity: Japanese evidence**
Presenter: **Mototsugu Shintani**, University of Tokyo, Japan
Co-authors: Keiichi Goshima, Hiroshi Ishijima
Because options on Japanese government bonds (JGB) futures are relatively new in the market, JGB-VIX Index cannot be computed before 2007. For the training period when JGB-VIX index is available as output data, we conduct the supervised learning using the daily newspaper articles as input. Using the estimated relationship between JGB-VIX and the news, we construct a new uncertainty measure based on contents of the newspapers from 1981 to 2017. Our uncertainty measure, which we call JGB-NVIX index, suggests that the volatility of JGB market increases with events related to stock market crashs, wars and government policy announcements. In the short run, our JGB-NVIX index is found to be useful in predicting the industrial production in Japan. Furthermore, using an identification strategy based on a VAR model with JGB-NVIX index,

we confirm that uncertain shocks have a negative impact on the real economic activities in Japan.

EO071 Room U517 RECENT DEVELOPMENTS IN TIME SERIES ANALYSIS Chair: Chiu-Hsing Weng

E0214: **Prediction intervals for time series and their applications to portfolio selection**
Presenter: **Hsiang-Ling Hsu**, Institute of Statistics, National University of Kaohsiung, Taiwan
Co-authors: Shih-Feng Huang
Prediction intervals for time series are considered, and the results are applied to portfolio selection. The dynamics of the high and low underlying returns are depicted by time series models, which lead to a prediction interval of future returns. We propose an innovative criterion for portfolio selection based on the prediction interval. A new concept of coherent risk measures for the interval of returns is introduced. An empirical study is conducted with the stocks of the Dow Jones Industrial Average Index. A self-financing trading strategy is established by daily reallocating the holding positions via the proposed portfolio selection criterion. The numerical results indicate that the proposed prediction interval has promising coverage, efficiency and accuracy for prediction. The proposed portfolio selection criterion constructed from the prediction intervals is capable of suggesting an optimal portfolio according to the economic conditions.

E0232: **Estimation of regression coefficients when fixed effects and random effects are correlated**
Presenter: **Chun-Shu Chen**, National Changhua University of Education, Taiwan
Co-authors: Yung-Huei Chiou
In spatial regression models, when covariates and unobservable random effects are correlated, past researches have shown that ignoring this relationship would have significant influences on the estimation of regression coefficients but how to modify them remains an active research topic. To solve this problem, an idea of restricted spatial regression is used to ensure that the unobservable spatial random process is orthogonal to covariates. Then, an adjusted generalized least squares estimation method is proposed to estimate regression coefficients, resulting in estimators that perform better than conventional methods. Statistical inferences of the proposed methodology are justified both theoretically and numerically.

E0264: **Kalman filter for innovations state space models**
Presenter: **Chiu-Hsing Weng**, National Chengchi University, Taiwan
Co-authors: Chan-Yuan Hsu
Exponential smoothing methods have been widely used for time series forecasting. To calculate likelihood and prediction intervals, many have specified various state space models that underlie exponential smoothing methods. Among these state space models, the innovations formulation, called innovations state space models, is a popular one. With a state space representation, one can easily incorporate regressors in exponential smoothing methods. We derive Kalman filter type algorithms for innovations state space models with and without regressors, where the regressors can be fixed or random. Some examples are used for illustration.

E0348: **The asymptotic excess risk of possibly non-stationary time-series**
Presenter: **Shu-Hui Yu**, Institute of Statistics, Taiwan
Model selection criteria are often assessed by the so-called asymptotic risk. Asymptotic risk is defined either with the mean-squared error of estimated parameters; or with the mean-squared error of prediction. The literature focuses on i.i.d. or stationary time-series data though. Using the latter definition of asymptotic risk, the conventional AIC-type and BIC-type information criteria, which are arguably most suitable for univariate time series in which the lags are ordered, are assessed. Throughout we consider a univariate AR process in which the AR order and the order of integratedness are finite but unknown. We prove that the BIC-type information criterion, which penalty goes to infinity, attains zero asymptotic excess risk. In contrast, the AIC-type information criterion, which penalty goes to a finite number strictly greater than 1, renders a strictly positive asymptotic excess risk. Further, the asymptotic excess risk increases with the admissible number of lags, a result that gives a warning about certain high-dimensional analyses when the true data generating process is of low-dimension. In sum, we extend the existing results in threefold: (i) a general I(d) process; (ii) same realization prediction; and (iii) an information criterion more general than AIC. Some simulation study shows these asymptotic results are valid for fairly small sample sizes.

E0491: **Large portfolio management with clustering techniques**
Presenter: **Huei-Wen Teng**, National Chiao Tung University, Taiwan
Large portfolio management faces many numerical problems and statistical difficulties. For instance, it is non-trivial to estimate a large covariance matrix that remains semi-positive, it is time demanding in the optimizing process when the dimension is high, and the optimized portfolio may not be stable or with high turnover rate. Instead of proposing an alternative approach to estimate the large covariance matrix, the aim is to propose a clustering method to overcome the above problems. With hierarchical clustering techniques, we partition the assets into several groups, so that assets behave similarly within groups but vary among groups. In each group, the asset closest to its centroid is selected as the candidate asset. The optimization procedure is then implemented for the selected small portfolio. With empirical analysis, we will show that the proposed method is comparable with that optimized directly from the large portfolio.

EC344 Room AT335 CONTRIBUTIONS IN COMPUTATIONAL AND BAYESIAN METHODS Chair: Tzy-Chy Lin

E0791: **Inferring medication adherence using health outcomes with Bayesian state-space models**
Presenter: **Kristen Hunter**, Harvard University, United States
Co-authors: Mark Glickman, Luis Campos
Patients' non-adherence to their prescribed medication is a serious obstacle to successful medication therapy and a widespread problem in clinical care. Providers and patients are likely more empowered to make more informed decisions if they have accurate information about medication adherence. Current methods to summarize medication adherence are generally not practical or accurate enough to be useful in clinical settings. We develop an approach to infer medication adherence rates from commonly-collected clinical data, including: (1) health outcomes measured over time that are likely to be directly impacted by differential adherence, and (2) baseline health characteristics and sociodemographic data. Our approach uses efficient Bayesian computational methods for the goal of inferring recent adherence behavior, and uses information not typically utilized in adherence models. The method we adopt can be understood in two steps. First, we fit a Bayesian State-Space Model (SSM) to health outcomes as a function of time-varying adherence. Second, we infer a particular patient's medication adherence given their observed health outcomes and baseline health and sociodemographic information using a Sequential Monte Carlo (SMC) algorithm, which accomplishes efficient sampling in high dimensional spaces. Summaries of adherence, including interval estimates, can be determined directly from the SMC posterior draws.

E0828: **Sequential methods for learning under cognitive diagnosis modeling**
Presenter: **Sangbeak Ye**, University of Missouri Kansas City, United States
Cognitive diagnostic modeling with binary latent attributes classifies each subject into a specific skills profile in a multidimensional binary domain. In the application of e-learning or intelligent tutoring system, the goal is to provide pedagogical resources until the binary latent attribute of the subject consistently corresponds with observational responses that indicate a complete mastery in such a domain under the framework of CDM.

The process of transitioning from any state to a complete mastery profile of multiple attributes is viewed as a sequential change-point problem. If each item is assumed to carry different magnitude of stimulus to transition one or more attribute from non-mastery to mastery, irreversibly, the selection criteria of each item may affect the duration until a complete mastery. A variation of item selection methods that adaptively induce the change points and improve the detection accuracy of a complete mastery to gain efficiency was developed. The item selection methods showcase adopting different statistical approaches including Bayesian principles and survival analysis modeling. A simulation study is conducted to compare the performance of the item selection methods.

E0746: **Bayesian joint models for longitudinal binary and survival data using general random effects covariance matrix**
Presenter: **Keunbaik Lee**, Sungkyunkwan University, Korea, South

Joint models are proposed to analyze longitudinal binary data with survival times data. Unlike the previous researches, random effects covariance matrix in our proposed joint models is assumed to be serially correlated and heterogeneous using modified Cholesky decomposition. The resulting parameters are estimated via linear/loglinear models, and the estimated random effects covariance matrix is positive-definite. The proposed methods are illustrated using real data.

E0801: **Estimating the competitive storage model with stochastic trend: A particle MCMC approach**
Presenter: **Kjartan Kloster Osmundsen**, University of Stavanger, Norway
Co-authors: Tore Selland Kleppe, Atle Oglend, Roman Liesenfeld

The structural parameters of the competitive storage model with stochastic trend, completely bounded storage and i.i.d. supply shocks are estimated using particle Markov chain Monte Carlo, relying only on price data. Applied to several real data sets of monthly commodity prices, the estimated storage model exceeds the log-likelihood values obtained by commonly used time series models.

E0800: **Compartmentalisation of variational approximate inference for inverse problems models**
Presenter: **Luca Maestrini**, University of Technology Sydney - School of Mathematical and Physical Sciences, Australia
Co-authors: Robert G Aykroyd, Matt P Wand

Inverse problems are essentially statistical regression problems where a response depending on a number of causal factors or parameters is measured and the goal is to estimate the parameter values. However, they may be highly multivariate and have predictors which are highly correlated. Hence even linear inverse problems cannot be solved by classical regression methods, nor can they be solved using standard dimension reduction or regularised regression techniques. A remedy is to use Markov random field models, which can be slow to fit via Markov chain Monte Carlo methods. Variational message passing updating algorithms for factor graph fragments arising in inverse problem Bayesian models are identified, catalogued and derived. The resultant factor graph fragments facilitate streamlined implementation of fast approximate algorithms for inverse problems and form the basis for software development. Contemporary inverse problems models give rise to new factor graph fragment types for different penalization strategies. Nevertheless, the variational message passing approach on factor graph fragments is such that algorithm updates and streamlining steps only need to be derived once for a particular fragment, which can be integrated in an arbitrarily complex model. The first applications are one- and two-dimensional deconvolution problems motivated by archaeology data.

EC347 Room AT337 CONTRIBUTIONS IN APPLIED STATISTICS Chair: Jong Soo Lee

E0793: **Joint graduation of male and female mortality rates using penalised splines**
Presenter: **Kai Hon Tang**, University of Southampton, United Kingdom
Co-authors: Erengul Dodd, Jon Forster

Splines with an exponentially increasing penalty are applied to period mortality data, creating a flexible yet robust method for graduation of mortality rates. Previous parametric graduation methods usually involve complicated mathematical functions due to the peculiar mortality pattern, the estimated parameters are often highly correlated hence introducing difficulties during the fitting process. By utilising P-splines, the model enjoys computational efficiency and flexibility. Observed data at high ages are often sparse and unreliable, an additional difference penalty between male and female spline coefficients is included to further strengthen the robustness and borrow information from each other. Constrained splines are used to prevent the cross-over of male and female mortality patterns when they are modelled jointly.

E0807: **Data mining swimming behavior of adult zebrafish**
Presenter: **Natalie Karavarsamis**, La Trobe University, Australia

Data mining methods are applied to extract important characteristics from studies conducted on swimming behavior of one month old zebrafish that may link genetic causes in childhood epilepsy. Observed were a control group and treated group exposed to hyperthermia at five days. Data arise from footage of individual fish at 30 days where each fish was recorded in a single testing arena filmed for 7hrs hours. Time is discretized to intervals of 0.06 second. Zebrafish swimming behaviour involves both circular and linear variables described by direction and distance of movement. Derivative variables include velocity, acceleration and relative meander. Identifying important differences in swimming behaviour may be indicative of spontaneous seizures used for creating associations between febrile seizures and temporal lobe epilepsy in fish. Distinct features from this big data scenario are mined with generalized linear mixed effects models (GLMMs), time series methods, graphical analysis, and functional data analysis methods. In addition, higher activity in fish indicates higher stress and anxiety levels. Thus, periods of active and inactive swimming is investigated. Identifying key characteristics and patterns in swimming behaviour is vital for fitting more complex models.

E0290: **Modeling and Regionalization of China's PM2.5 Using Spatial-Functional Mixture Model**
Presenter: **Decai Liang**, Peking University, China
Co-authors: Hui Huang

Severe air pollution affects billions of people around the world, particularly in developing countries such as China. Effective emission control policies rely primarily on proper assessment of air pollutants and accurate spatial clustering outcomes. Unfortunately, emission patterns are difficult to observe as they are highly confounded by many meteorological and geographical factors. The standard clustering techniques generally fail to exploit the spatiotemporal features of data. We propose a novel approach for modeling and clustering daily PM2.5 concentrations all over China. Observed concentrations from monitoring stations are modeled as spatially dependent functional data. We assume the latent emission processes originate from a functional mixture model with each component as a spatiotemporal process. Cluster memberships of stations follow a Markov random field model and geographical factors are also considered. The superior performance of our approach compared to others is demonstrated using extensive simulation studies. Our method is effective in dividing China into several regions based on PM2.5 concentrations, suggesting separate local emission control policies are needed.

E0760: **Comparison of statistical methods for analyzing missing survival times in studies of census and death registry data**
Presenter: **Wei-Ting Hwang**, University of Pennsylvania, United States
Co-authors: Arielle Marks-Anglin, Frances Barg

Observational studies that evaluate the effect of risk factors on mortality are sometimes limited to population registries for outcome information. Such studies could have missing death dates, particularly when the population under study is identified by census, where the distributions of age

and other risk factors that impact mortality vary significantly. To date, there are no well-developed statistical methods to address this situation. Motivated by this gap in methodology, we conduct simulation studies to compare the performance of four analytical approaches (complete case analysis, censoring at the age of the census, inverse probability weighting, and multiple imputation of the residual lifetime), in estimating the exposure effects and median survival times in the presence of missing survival times. We also explore the effects of different missing data mechanisms (e.g., MCAR, MAR, MNAR), exposure-survival associations, censoring and missing percentages on their performance. The methods are applied to a cohort of the residents in Ambler, PA established using the 1930 US census, from which only 2,440 out of the total 4,514 individuals had death records retrievable from publicly available data sources and death certificate. Using this cohort, we examine the effects of occupational and paraoccupational asbestos exposure on survival and disparities in mortality by race and gender.

E0723: **Estimation parameter for two-stage randomized response technique in logistic regression model**
Presenter: **Kim Hung Pho**, Feng Chia University, Taiwan

When a survey study is related to sensitive issues, such as political orientation, sexual orientation, and income, interviewee may not be willing to respond truthfully, which leads to bias results. In order to protect the interviewee privacies and improve their willingness to provide the correct answer, the randomized response (RR) technique has been proposed, in which the respondents randomly selects questions by means of devices in order to ensure that they are protected. This design has been extended to propose a two-stage RR design. Not only can this method be used to estimate the proportion of a sensitive group, but also to estimate the honest answer rate. We use the two-stage randomized response design to apply a logistic regression model to investigate and estimate the effects of covariates on the proportion of a sensitive feature and the honest response rate. As an application, the proposed methodology is applied to analyze the survey data of sexuality of freshmen at Feng Chia University in 2016.

EC342 Room U302 CONTRIBUTIONS IN FINANCIAL ECONOMETRICS Chair: Jiri Witzany

E0740: **Latent volatility Granger causality and spillovers in renewable energy and crude oil ETFs**
Presenter: **Yu-Ann Wang**, National Chung Hsing University, Taiwan
Co-authors: Chia-Lin Chang, Michael McAleer

The purpose is to examine latent volatility Granger causality for four renewable energy Exchange Traded Funds (ETFs) and crude oil ETF (USO), namely solar (TAN), wind (FAN), water (PIO), and nuclear (NLR). Data on renewable energy and crude oil ETFs are from 18 June 2008 to 20 March 2017. From the underlying stochastic process of a vector random coefficient autoregressive (VRCAR) process for the shocks of returns, we derive Latent Volatility Granger causality from the Diagonal BEKK multivariate conditional volatility model. We follow a previous definition of the co-volatility spillovers of shocks, which calculate the delayed effect of a returns shock in one asset on the subsequent volatility or co-volatility in another asset, and extend the effects of the co-volatility spillovers of shocks to the effects of the co-volatility spillovers of squared shocks. The empirical results show there are significant positive latent volatility Granger causality relationships between solar (TAN), wind (FAN), nuclear (NLR), and crude oil (USO) ETFs, specifically significant volatility spillovers of shocks from solar ETF on the subsequent wind ETF co-volatility with solar ETF, and wind ETF on the subsequent solar ETF co-volatility with wind ETF. Interestingly, there are significant volatility spillovers of squared shocks for the renewable energy ETFs, but not with crude oil ETFs.

E0745: **Filtering momentum life cycles, price acceleration signals and trend reversals for stocks, credit derivatives and bonds**
Presenter: **Periklis Brakatsoulas**, Charles University Prague, Czech Republic

The aim is to develop a diversified portfolio approach to price distortion signals using daily position data on stocks, credit derivatives and bonds. An algorithm allocates assets periodically and new investment tactics take over upon price momentum signals and across different ranking groups. We focus on momentum life cycles, trend reversals and price acceleration signals. The main effort here concentrates on the density, time span and maturity of momentum phenomena to identify consistent patterns over time and measure the predictive power of buy-sell signals generated by these anomalies. To tackle this, we propose a two-stage modelling process. First, we generate forecasts on core macroeconomic drivers. Secondly, satellite models generate market risk forecasts using the core driver projections generated at the first stage as input. Moreover, using a combination of the ARFIMA and FIGARCH models, we examine the dependence of consecutive observations across time and portfolio assets since long memory behavior in volatilities of one market appears to trigger persistent volatility patterns across other markets. We believe that this is the first work that employs evidence of volatility transmissions among derivatives, equities and bonds to identify momentum life cycle patterns.

E0789: **Dynamic quantile model for bond pricing**
Presenter: **Frantisek Cech**, UTIA AV CR vvi, Czech Republic
Co-authors: Jozef Barunik

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

E0795: **Stylised facts for high frequency cryptocurrency data**
Presenter: **Yuanyuan Zhang**, University of Manchester, United Kingdom

The term "stylised facts" has been extensively researched through the analysis of many different financial datasets. More recently, cryptocurrencies have been investigated as a new type of financial asset, and provide an interesting example, with a current market value of over 500 billion dollars. We analyse the stylised facts in terms of the Hurst exponent, using both the DFA and R/S methods, of the four most popular cryptocurrencies ranked according to their market capitalisation. The analysis is conducted on high frequency returns data with varying lags. In addition to using the Hurst exponent, our analysis also considers features of dependence between the different cryptocurrencies.

E0813: **Modelling foreign exchange time series with SVJD models with stochastic drift**
Presenter: **Milan Ficura**, University of Economics in Prague, Czech Republic
Co-authors: Jiri Witzany

Several models with stochastic volatility, jumps, and stochastic drift are proposed and applied to the foreign exchange time series on the daily frequency. Alternative specifications of the drift term are tested, including Markov-Switching, continuous diffusion, pure-jump and jump-diffusion processes. The estimation is performed with Sequential Gibbs Particle Filter algorithm and the proposed models are compared based on their Bayes factors. In addition to the statistical evaluation, investment strategies utilizing the stochastic drift are proposed, and their risk-adjusted profitability is computed. Implications with respect to a possible violation of the weak form Efficient Market Hypothesis are discussed.

EG008 Room U502 CONTRIBUTIONS IN APPLIED ECONOMETRICS I **Chair: Massimiliano Caporin**

E0753: **To what extent globalization affects exchange rate pass-through: The role of global value chains**
Presenter: **Jan Hagemejer**, Narodowy Bank Polski, Poland
Co-authors: Aleksandra Halka, Jacek Kotlowski

The aim is to investigate to what extent the increasing participation in the global value chains influences the exchange rate pass-through to inflation. We use a broad panel data set to check whether the growing trade in intermediate goods leads to decrease of ERPT. We use a novel one-step approach and apply the panel smooth transition regression model to address the problems of co-linearity and also to account for potential non-linearities in the impact of GVC participation on ERPT. We test for non-linearity first and we find ERPT being non-linear in respect to intensity of GVC participation. We also find the logistic function as a proper transition function in PSTR model. In contrary to existing research, we investigate direct ERPT to producer prices instead of import prices. We employ several control variables to account for potential impact of other factors, which constitute an alternative explanation for decline in exchange rate pass-through. We use also different measures of GVC participation to increase the robustness of our results. We find that the growing GVC backward participation of the suppliers of imported intermediate input results in reducing the ERPT to producer prices. We also evidence that this effect is non-linear. The ERPT for countries strongly involved in the global value chains production is significantly smaller than for economies not participating in GVC.

E0336: **A financial Kuznets curve in the transitional economy: The tale of China**
Presenter: **Chi-Yang Chu**, National Taipei University, Taiwan
Co-authors: Mingming Jiang

Using manually collected grouped income data for urban households, different measures of income inequality are constructed for 30 Chinese provinces in the past three decades and the nonlinear impacts of financial depth on income distribution is quantified by using a partially linear semi-parametric panel data model. The estimates suggest a robust and inverted-U shaped relation between financial depth and income inequality, i.e., a financial Kuznets curve arises during the transition of Chinese economy. We find the marginal impacts of financial depth are affected by a couple of factors, including human capital, housing price, economic ownership, economic development, and initial inequality.

E0273: **Testing for distributional features in varying coefficient panel data models**
Presenter: **Alexandra Soberon**, Universidad de Cantabria, Spain
Co-authors: Winfried Stute, Juan Manuel Rodriguez-Poo

Several tests for skewness and kurtosis for the error terms are provided in a one-way fixed effects varying coefficient panel data model. In order to obtain these tests, estimators of higher order moments of both error components are obtained as solutions of estimating equations. Also, to obtain the nonparametric residuals, a local constant estimator based on a pairwise differencing transformation is proposed. The asymptotic properties of these estimators and tests are established. The proposed estimators and test statistics are augmented by simulation studies and they are also illustrated in an empirical analysis about the technical efficiency of the European Union companies.

E0697: **Volatility modelling using range-based measures and weighted exogenous threshold CARR model**
Presenter: **Kok-haur Ng**, University of Malaya, Malaysia

Three volatility measures including the squared returns and range-based Parkinson and Garman Klass were applied to estimate the financial volatility. These measures are then fitted to conditional autoregressive range (CARR) models and its weighted exogenous threshold extensions using generalised Beta type two distribution. The daily All Ordinaries index is studied by fitting the three volatility measures to the two types of CARR models and compare their model performances. Results show that the Garman Klass measure fitted to weighted exogenous threshold CARR model gives the best in-sample model fit based on Akaike information criterion. Different levels of value-at-risk are also provided.

E0779: **Diversification power of Turkish real estate market securities: The role of data frequency and dividend policies**
Presenter: **Metin Ilbasmis**, University of Aberdeen Business School & Wenzhou-Kean University, United Kingdom
Co-authors: Marc Gronwald, Yuan Zhao

Dynamic correlations are investigated between stock and REIT markets in both Turkey and the U.S., using an Asymmetric DCC - GJR - GARCH model to estimate the dynamic correlation at daily, weekly, and monthly frequencies. The contribution is threefold. First, we show, in all three data frequencies, a downward trend in conditional correlation in the Turkish market, which is contrary to the literature, while the upward trend in the correlation of the two U.S. markets is consistent with the literature. Second, we observe that the trend in the correlation changes the direction with the 2008 Global Financial Crisis. The negative trend in Turkish market becomes positive and the positive trend in the U.S. market becomes negative after the crisis, which could indicate a structural break in the REIT market with the crisis. Third, the dividend policy of REITs plays an important role on the dynamics of the correlation. Dividend payments by Turkish REITs are negatively associated with the correlation while no such relationship is detected in the U.S. Furthermore, we argue that effects of dividend payments by REITs on REIT correlation with the stock index is associated with the different regulatory environment of REITs in Turkey.

Tuesday 25.06.2019 14:00 - 15:40 Parallel Session C – EcoSta2019

EI011 Room UB99(B1) RECENT ADVANCES IN ECONOMETRIC TIME SERIES Chair: Sangyeol Lee

E0153: **Bayesian analysis for heterogeneity**
Presenter: **Jaeyoung Kim**, Seoul National University, Korea, South
A convenient new approach is studied for analyzing empirical problems in the presence of heterogeneity. The approach is a Bayesian semi-parametric method with an empirical prior capturing heterogeneity. The considered heterogeneity is a general type of heterogeneity that covers a variety of cases studied in economic theory and econometric practice. It is also applicable for developing inference for model misspecification. To develop our framework to get empirical prior, we adopt a minimum distance method with an information distance to solve an ill-posed deconvolution problem. Inferential issues of detection and estimation of heterogeneity are also studied based on the framework. The methods are evaluated by a pair of MonteCarlo experiments whose results confirm usefulness of our methods for practical applications. We also applied our methods for two interesting empirical examples.

E0154: **Intraday range-based stochastic volatility models with application to the Japanese stock index**
Presenter: **Toshiaki Watanabe**, Hitotsubashi University, Japan
Co-authors: Jouchi Nakajima
Realized stochastic volatility (RSV) models, where the true volatility is modelled jointly with a realized measure (RM) of volatility taking account of the bias in the RM, are extended for the analysis of high-frequency intraday volatility. The proposed model consists of the persistent autoregressive stochastic volatility process, seasonal components of the intraday volatility patterns, and correlated jumps in prices and volatilities. The range of the logarithmic prices within each intraday time interval is used as a RM in the proposed model. A Bayesian method for the analysis of this model is developed using Markov chain Monte Carlo (MCMC) with the exact multi-move sampler for the SV process. Using this method, the proposed model is applied to the 5-minute returns of Nikkei 225 index. It is also examined whether the intraday range-based RSV model improves the predictive ability of volatility compared with the intraday SV model without the range information and commonly-used models for daily realized volatility.

E0155: **Estimation of error correction model with measurement errors**
Presenter: **Sung Ahn**, Washington State University, United States
Effects of measurement errors on the analysis of error correction models (ECMs) of vector processes observed with measurement errors were studied previously. It was found that statistically undesirable effects on the analysis attributable to endogeneity in the ECM induced by measurement errors, even in their simplest form. Therefore, we propose a method using instrumental variables and derive the asymptotic distributions of the reduced rank estimator that eliminate the undesirable effect of endogeneity. We also propose a method of using a moving-average term to deal with endogeneity. These methods yield estimators that are consistent and asymptotically unbiased. We investigate the effects of the measurement errors on the proposed methods through a Monte Carlo simulation study.

EO047 Room S101 MACHINE LEARNING THEORY Chair: Ding-Xuan Zhou

E0366: **A strategy for identifying informative variables: Prediction of developmental outcomes of preterm neonates**
Presenter: **Sergiy Pereverzyev**, Johann Radon Institute for Computational and Applied Mathematics (RICAM), Austria
Co-authors: Sergiy Pereverzyev-Jr, Vasyl Semenov
The problem of detecting the most informative coordinates of inputs allowing an accurate reconstruction of the corresponding outputs is discussed. We are motivated by predicting neurodevelopmental outcomes of preterm neonates from the ratios of the amplitudes of the peaks of metabolite spectra. One of the difficulties of the above prediction problem is that the available clinical data contain only a few input-output pairs associated with neurodevelopmental impairments. The construction of a predictor from training data can be seen as the recovery of a multivariable function from its evaluations at given points. An attempt to approximate the ideal prediction function by using an insufficient number of variables generates an error above the best guaranteed one. This remark hints that informative variables can be detected by finding the minimum value of the prediction errors observed for the predictors from the considered input combinations. The predictors employed in our study are obtained by kernel ridge regression (KRR) with various input variables as regressors. In KRR we use universal Gaussian kernels and the kernels constructed from the data according to a recently proposed approach. In this way, we try to cover a variety of predictors exhibiting universality or specificity.

E0377: **A geometrical framework for covariance matrices and covariance operators in machine learning and applications**
Presenter: **Minh Ha Quang**, RIKEN, Japan
Symmetric positive definite (SPD) matrices, in particular covariance matrices, play important roles in many areas of mathematics and statistics, with numerous applications various different fields, including machine learning, brain imaging, and computer vision. The set of SPD matrices is not a subspace of Euclidean space and consequently algorithms utilizing the Euclidean metric tend to be suboptimal in practice. A lot of recent research has therefore focused on exploiting the intrinsic geometrical structures of SPD matrices, in particular the view of this set as a Riemannian manifold. We will present a survey of some of the recent developments in the generalization of the geometrical structures of finite-dimensional covariance matrices to those of infinite-dimensional covariance operators. Computationally, we focus on covariance operators in Reproducing Kernel Hilbert Spaces (RKHS). This direction exploits the power of kernel methods from machine learning in a geometrical framework, both mathematically and algorithmically. The theoretical formulation will be illustrated with applications in computer vision, which demonstrate both the power of kernel covariance operators as well as of the algorithms based on their intrinsic geometry.

E0394: **Approximate kernel PCA: Computational vs. statistical tradeoff**
Presenter: **Bharath Sriperumbudur**, Pennsylvania State University, United States
Co-authors: Nicholas Sterge
Kernel principal component analysis (KPCA) is a popular non-linear dimensionality reduction technique, which generalizes classical linear PCA by finding functions in a reproducing kernel Hilbert space (RKHS) such that the function evaluation at a random variable X has maximum variance. Despite its popularity, kernel PCA suffers from poor scalability in big data scenarios as it involves solving an $n \times n$ eigensystem leading to the computational complexity of $O(n^3)$ with n being the number of samples. To address this issue we consider random approximations to kernel PCA which requires solving an mxm eigenvalue problem and therefore has a computational complexity of $O(m^3)$, implying that the approximate method is computationally efficient if $m < n$ with m being the number of random features. The goal is to investigate the trade-off between computational and statistical behaviors of approximate KPCA, i.e., whether the computational gain is achieved at the cost of statistical efficiency. We show that the approximate KPCA is both computationally and statistically efficient compared to KPCA in terms of the error associated with reconstructing a kernel function based on its projection onto the corresponding eigenspaces.

E0365: **Theory of deep convolutional neural networks for deep learning**
Presenter: **Ding-Xuan Zhou**, City University of Hong Kong, Hong Kong

Deep learning has been widely applied and brought breakthroughs in speech recognition,computer vision, and many other domains. The involved deep neural network architectures and computational issues have been well studied in machine learning. But there lacks a theoretical foundation for understanding the approximation or generalization ability of deep learning methods with network architectures such as deep convolutional neural networks (CNNs) with convolutional structures. The convolutional architecture gives essential differences between the deep CNNs and fully-connected deep neural networks, and the classical approximation theory of fully-connected networks developed around 30 years ago does not apply. An approximation theory of deep CNNs associated with the rectified linear unit is described. In particular, we show the universality of such a deep CNN, meaning that it can be used to approximate any continuous function to an arbitrary accuracy when the depth of the neural network is large enough. Our quantitative estimate, given tightly in terms of the number of free parameters to be computed, verifies the efficiency of deep CNNs in dealing with large dimensional data.

EO139 Room S102 RECENT ADVANCES IN ADAPTIVE CLINICAL TRIAL DESIGN Chair: Yisheng Li

E0384: **Bayesian shrinkage models and designs for Phase 1b cohort expansion trials**
Presenter: **Yuan Ji**, The University of Chicago, United States
Co-authors: Jiaying Lyu

A class of Bayesian hierarchical models for phase 1b trials aiming to expand multiple doses on multiple indications will be discussed. Shrinkage priors are used to differentiate the degree of information sharing across doses and indications. Due to the hierarchical structure, proper shrinkage and multiplicity control is achieved in the inference, which results in increased power and smaller sample size. An example will be given to illustrate the performance of the new approach with comparison to existing methods. The new design has been applied to real-life oncology trials and is expected to help improve the efficiency of overall drug development.

E0388: **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
Co-authors: Xiao Su, Peter Mueller, Kim-Anh Do

While a number of phase I dose-finding designs in oncology exist, the commonly used ones are either algorithmic or empirical model-based. Other statistical designs that incorporate pharmacokinetic (PK) data mainly focus on summary PK information (such as the area under the concentration-time curve [AUC] or maximum serum concentration [Cmax]). We aim to: 1) propose an extended framework for modeling the dose-toxicity relationship, by incorporating dynamic PK and pharmacodynamic (PD) information via PK/PD modeling; and 2) apply this modeling framework in the design of phase I trials. The proposed modeling framework naturally incorporates the information on the impact of dose, schedule and method of administration (e.g., drug formulation and route of administration) on toxicity. We conduct extensive simulation studies to evaluate the performance of the proposed design and compare it with existing 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.

E0440: **Biomarker-based Bayesian randomized phase II clinical trial design to identify a sensitive patient subpopulation**
Presenter: **Satoshi Morita**, Kyoto University Graduate School of Medicine, Japan

The benefits and challenges of incorporating biomarkers into the development of anti-cancer agents have been increasingly discussed. Prospective exploration of sensitive subpopulations of patients may enable us to efficiently develop definitively effective treatments, resulting in accelerated drug development and a reduction in development costs. We consider the development of a new molecular-targeted treatment in cancer patients. We propose a Bayesian randomized phase II clinical trial design incorporating a biomarker that is measured on a graded scale. We compare two Bayesian methods, one based on subgroup analysis and the other on a regression model, to analyze a time-to-event endpoint such as progression-free survival (PFS) time. Extensive simulation studies evaluate these methods' operating characteristics under a wide range of clinical scenarios. Although both methods' performance depends on the distribution of treatment effect and the population proportions across patient subgroups, the regression-based method shows more favorable operating characteristics.

EO113 Room S104 RECENT ADVANCES ON HIGH-DIMENSIONAL STATISTICAL INFERENCE Chair: Xiaohui Chen

E0263: **A new framework for distance and kernel-based metrics in high dimensions**
Presenter: **Xianyang Zhang**, Texas A&M University, United States
Co-authors: Shubhadeep Chakraborty

New metrics are presented to quantify and test for (i) the equality of distributions and (ii) the independence between two high-dimensional random vectors. First, we show that the energy distance based on the usual Euclidean distance cannot completely characterize the homogeneity of two high dimensional distributions in the sense that it only detects the equality of means and a specific covariance structure in the high dimensional setup. To overcome such a limitation, we propose a new class of metrics which inherit some nice properties of the energy distance in the low-dimensional setting and is capable of detecting the pairwise homogeneity of the low dimensional marginal distributions in the high dimensional setup. The key to our methodology is a new way of defining the distance between sample points in the high-dimensional Euclidean spaces. Second, we propose new distance-based metrics to quantify the dependence between two high dimensional random vectors. In the growing dimensional case, we show that both the population and sample versions of the new metrics behave as an aggregation of the group-wise population and sample distance covariances. Thus, it can quantify group-wise non-linear and non-monotone dependence between two high-dimensional random vectors.

E0370: **A neighborhood-assisted Hotelling's T2 test for high-dimensional means**
Presenter: **Jun Li**, Kent State University, United States
Co-authors: Yumou Qiu

Many tests have been proposed to remedy the classical Hotelling's T^2 test in the "large p, small n" paradigm, but how to incorporate data dependence in the sum-of-squares type test to enhance the power has not been explored. We show that, under certain conditions, the population Hotelling's T^2 test with the known Σ^{-1} attains the best power among all the L_2-norm based tests with the data transformed by Σ^{η} for $\eta \in (-\infty, \infty)$. To extend the result to the case of unknown Σ^{-1},we propose a Neighborhood-Assisted Hotelling's T^2 (NEAT) statistic obtained by replacing the inverse of sample covariance matrix in the classical Hotelling's T^2 statistic with a regularized covariance estimator. Utilizing a regression model, we establish its asymptotic normality under mild conditions. Without any structural assumption on Σ, the proposed NEAT test is able to enhance the power by incorporating variable dependence through an adaptively chosen neighborhood, and thus more powerful than other tests without utilizing dependence. An optimal neighborhood size selection procedure is proposed to maximize the power of the NEAT test via maximizing the signal-to-noise ratio. As a special case, our results demonstrate that if Σ happens to satisfy a certain bandable structure, the neighborhood exploration procedure leads to an optimal test that matches the population Hotelling's T^2 test.

E0527: **Inference of break-points in high-dimensional time series**
Presenter: **Likai Chen**, Washington University in Saint Louis, United States
Co-authors: Weining Wang, Wei Biao Wu
A new procedure is considered for detecting structural breaks in mean for high-dimensional time series. We target breaks happening at unknown time points and locations. In particular, at a fixed time point, the method is concerned with either the biggest break in one location or aggregating simultaneous breaks over multiple locations. We allow for both big or small sized breaks, so that we can 1), stamp the dates and the locations of the breaks, 2), estimate the break sizes and 3), make inference on the break sizes as well as the break dates. The theoretical setup incorporates both temporal and cross-sectional dependence, and is suitable for heavy-tailed innovations. We derive the asymptotic distribution for the sizes of the breaks by extending the existing powerful theory on local linear kernel estimation and high dimensional Gaussian approximation to allow for trend stationary time series with jumps. A robust long-run covariance matrix estimation is proposed, which can be of independent interest. An application on detecting structural changes of the US unemployment rate is considered to illustrate the usefulness of the method.

E0632: **A robust and high-dimensional bootstrap change point test for location parameter**
Presenter: **Mengjia Yu**, University of Illinois at Urbana-Champaign, United States
Co-authors: Xiaohui Chen
In change point analysis, the widely used cumulative sum (CUSUM) statistics are sensitive to outliers. We propose a robust test for change point detection problem of location-shift in high dimensions when the dimension p can be much larger than the sample size n. To achieve the robustness purpose in a nonparametric setting, we consider signal cancellations in the general U-statistics framework with anti-symmetric kernels of order 2. To calibrate the distribution of our test statistic, a Gaussian multiplier bootstrap is proposed. Subject to mild conditions on kernels, we derive the uniform rates of convergence of the multiplier bootstrap to the sampling distribution of the test statistic. The proposed test is fully data-dependent without any tuning parameter, and numeric studies are provided.

EO299 Room S106 RECENT ADVANCES IN META-ANALYSIS FOR MEDICAL RESEARCH Chair: Satoshi Hattori

E0432: **Efficient implementation of Copas selection model for publication bias in meta-analysis using clinical trial registry**
Presenter: **Ao Huang**, Department of Biomedical Statistics, Graduate School of Medicine, Osaka University, China
Co-authors: Sho Komukai, Satoshi Hattori
Copas selection model is one of the useful sensitivity analysis methods for publication bias in the standard meta-analyses. Despite its usefulness to quantify potential impacts of publication bias, it has some undesirable features. In conducting the sensitivity analysis, one needs to make inference with sets of fixed sensitivity parameters. This may lead us difficulty in interpreting the results of the sensitivity analysis. A method to estimate all the unknown parameters has been proposed based on data with an EM-type algorithm. However, this method is constructed under a strong assumption on funnel-plot symmetry. We extend the inference procedure for the Copas selection model by utilizing information from the clinicalTrials.gov, and propose two strategies in estimating the parameters of interest: one is the two stage method which estimates the parameters in the selection model first using a probit model, then with the parameters fixed in the likelihood, we estimate the bias-adjusted treatment effects; the other is utilizing the full likelihood function with all the information to estimate the parameters simultaneously. Through applications to real datasets and simulation studies, we show that our methods enable us to conduct the sensitivity analysis more stably and have more interpretable insights on publication bias.

E0557: **Bayesian selection model for publication bias correction in meta-analysis of prognostic studies**
Presenter: **Satoshi Hattori**, Osaka University, Japan
Publication bias is a serious issue in conduction meta-analyses. For meta-analyses to evaluate intervention effects, the funnel plot and the trim-and-fill method are widely used due to their simplicity. However, they rely on some strong assumptions and may give us misleading insights. The Copas selection model is an alternative as a tool to quantify potential influence of publication bias on the aggregated estimates for the intervention effects. In meta-analyses of prognostic studies, the summary receiver operating characteristics (ROC) curve is widely used. A Copas-type selection model is proposed for the summary ROC curve and an inference procedure under a frequentist setting. Although it enables us to evaluate potential impacts of publication bias on the summary ROC estimation, it has a drawback that an important parameter responsible for publication bias cannot be estimated based on data and then must be treated as a sensitivity parameter. To overcome this difficulty, we propose a Bayesian inference for the Copas-type selection model and demonstrate its usefulness in practice through application to some real datasets.

E0573: **Evidence synthesis in rare diseases**
Presenter: **Tim Friede**, University Medical Center Goettingen, Germany
Meta-analyses in rare diseases (also called orphan diseases) and small populations generally face particular problems, including small numbers of studies, small study sizes and heterogeneity of results. However, the heterogeneity is difficult to estimate if only very few studies are included. Motivated by a systematic review in immunosuppression following liver transplantation in children the characteristics of commonly used frequentist and Bayesian procedures for random effects meta-analyses will be explored and recommendations for their practical applications will be made. This will also include the common case of meta-analyses including only two studies. We will extend the meta-analytic methods to generalized evidence synthesis or cross design synthesis considering the special case of combing a randomized controlled trial with a non-randomized study, i.e. data from a clinical registry. This part is motivated by a clinical research project in Creutzfeldt-Jakob disease.

E0719: **An improved Henmi-Copas confidence interval for random effects meta-analysis with small number of studies**
Presenter: **Masayuki Henmi**, The Institute of Statistical Mathematics, Japan
Co-authors: Satoshi Hattori, Tim Friede
The DerSimonian Laird condence interval for the average treatment effect in meta-analysis is widely used in practice when there is heterogeneity between studies. However, it is well known that its coverage probability can be substantially below the target level of 95 per cent. It can also be very sensitive to publication bias. For solving this problem, a confidence interval (HC interval) has been proposed by using the fixed effects estimate as the center of the interval in the random effects setting. Although the HC interval has better coverage probability than the DerSimonian-Laird confidence interval and is less sensitive to publication bias, its coverage probability is still below the target level especially when the number of studies included in meta-analysis is small. This is because the HC interval uses the DerSimonian-Laird estimate of the between-study variance in its construction and this estimate is less accurate as the number of studies gets smaller. We propose an improved version of the HC interval by replacing the DerSimonian-Laird estimate with a more accurate estimate. Its performance is examined by simulation studies and we apply it to a real-data example for illustration.

EO037 Room S1A01 INFERENCES AND PREDICTION FOR SPATIAL OR DYNAMIC DATA Chair: Naisyin Wang

E0239: **Dynamic landmark prediction for mixture data**
Presenter: **Tanya Garcia**, Texas AM University, United States
Co-authors: Layla Parast
In kin-cohort studies, clinicians are interested in providing their patients with the most current cumulative risk of death arising from a rare deleterious mutation. Estimating the cumulative risk is difficult when the genetic mutation status in patients is unknown and, instead, only estimated probabilities of a patient having the mutation are available. We estimate the cumulative risk using a novel nonparametric estimator that incorporates covariate information and dynamic landmark prediction. The contributions are three-fold. Our estimator better informs patients of their risk of death, as it yields improved prediction accuracy over existing estimators that ignore covariate information. The estimator is built within a dynamic landmark prediction framework whereby we can obtain personalized dynamic predictions over time. Compared to current standards, a simple transformation of our estimator provides more efficient estimates of marginal distribution functions in settings where patient-specific predictions are not the main goal. We show our estimator is unbiased and has substantial gains in predictive accuracy compared to approaches that ignore covariate information and landmarking. Our method is motivated by and illustrated using data from a Huntington disease study; results illustrate the development of survival prediction curves incorporating gender and familial genetic information, and the creation of personalized dynamic risk trajectories over time.

E0707: **Spatially dependent functional data: Covariance estimation, principal component analysis, and kriging**
Presenter: **Yehua Li**, University of California at Riverside, United States
Spatially dependent functional data collected under a geostatistics setting are considered. Locations are sampled from a spatial point process and a random function is observed at each location. The functional response is the sum of a spatially dependent functional effect and a spatially independent functional nugget effect. Observations on each function are made on discrete time points and contaminated with measurement errors. Under the assumption of spatial stationarity and isotropy, we propose a tensor product spline estimator for the spatio-temporal covariance function. If a coregionalization covariance structure is further assumed, we propose a new functional principal component analysis method that borrows information from neighboring functions. Under a unified framework for both sparse and dense functional data, where the number of observations per curve is allowed to be of any rate relative to the number of functions, we develop the asymptotic convergence rates for the proposed estimators. Advantages of the proposed approach over existing methods are demonstrated through simulation studies and a real data application to the home price-rent ratio data in the San Francisco Bay Area.

E0706: **Optimal prediction in generalized functional linear model with semiparametric single-index interactions**
Presenter: **Yanghui Liu**, East China Normal University, China
Co-authors: Yehua Li, Naisyin Wang, Raymond Carroll, Riquan Zhang
A generalized functional linear model with semiparametric single-index interactions is considered. The response variable was assumed previously to depend on multiple covariates as well as on a finite number of features in the functional predictor. We incorporate all features of the functional predictor into our prediction model. We consider a two-stage estimation procedure, in which a regularized functional predictor based on functional principal component analysis is used. The asymptotic properties of our estimators and rate of prediction are derived. It is shown that, under mild regularity conditions, the parametric estimators are $\sqrt{n}$-consistent, and are asymptotically normal when the functional predictor is under-smoothed. Furthermore, the overall convergence rate of squared prediction error is dominated by the nonparametric link function in the single-index component part, while the prediction rate for the functional linear part attains the optimal rate for traditional functional linear model in the minimax sense. Our theory also suggests that it works well to use a K-fold cross-validation procedure to identify a range of suitable m. We note that within this range, the prediction errors are insensitive toward the choice of m and satisfactory outcomes are achieved. The finite sample properties of our methods are further illustrated by a simulation study and a crop yield prediction application.

E0345: **Tensor factorization for temporal recommender systems**
Presenter: **Annie Qu**, University of Illinois at Urbana-Champaign, United States
Recommender systems have been widely adopted by electronic commerce and entertainment industries, as they bring insightful business intelligence for decision makers and convey useful information to customers efficiently. As an effective predictive model, recommender systems use individual experience or preference to make personalized prediction and recommendation. In general, individual experiences and preferences change over time, and capturing such changes is essential for developing accurate recommender systems. We introduce a temporal recommender system applicable to product forecasting, which achievers more-accurate sales forecasting for stores and manufactures, and improves decision-making on new product introduction. Specifically, we develop a time-varying tensor function, implement tensor factorization under the nonparametric framework to provide the time-dependent predictions, and utilize joint information from subgroups to solve the "cold-start" problem in the absence of information from new customers, new products or new contexts. We develop the asymptotic consistency of the predictions and nonparametric parameter estimators. In addition, the proposed method is illustrated for the IRI marketing data. Our numerical studies indicate that the proposed method outperforms existing competitors in the literature.

EO137 Room AT241 RECENT TOPICS IN DESIGN OF EXPERIMENT Chair: Wei Zheng

E0215: **ICUDO: Incomplete U-statistic based on division and orthogonal arrays**
Presenter: **Wei Zheng**, University of Tennessee, United States
Co-authors: xiangshun kong
U-statistics are an important class of statistics. Unfortunately, its computation easily becomes impractical as the data size n increases. Particularly, the number of combinations, say m, that a U-statistic of order d has to evaluate is in the order of n^d. Many efforts have been made to approximate a U-statistic by a small subset of combinations. Such an approximation is called incomplete U-statistic. To the best of our knowledge, all existing methods require m to grow at least faster than n, albeit much slower than n^d, in order for the corresponding incomplete U-statistic to be asymptotically efficient in the sense of mean squared error. We introduce a new type of incomplete U-statistic, which can be asymptotically efficient even when m grows slower than n. In some cases, m is only required to grow faster than $\sqrt{n}$. Both one-sample and multi-sample cases and both non-degenerate and degenerate cases are thoroughly discussed.

E0672: **Efficient design and analysis for a selective choice process**
Presenter: **Qing Liu**, University of Wisconsin-Madison, United States
Variable selection is a decision heuristic that describes a selective choice process. Choices are made based on only a subset of product attributes, while the presence of other ("inactive") attributes plays no active role in the decision. Within this context, the authors address two integrated topics that have received scant attention: the efficient design of choice experiments and the analysis of data arising from a selective choice process. A new dual-objective compound design criterion is proposed that incorporates prior information for the joint purpose of efficient estimation of the effects of the active attributes and detection of the effects of attributes stated as inactive but may turn out to be active. The approach leverages self-stated auxiliary data as prior information both for individual-level customized design construction and in a heterogeneous variable selection

model. The authors demonstrate the efficiency advantages of the approach relative to design benchmarks and highlight practical implications using both simulated data and actual data from a conjoint choice experiment where individual designs were customized on-the-fly using self-stated active/inactive attribute status.

E0670: **A discrepancy-based design for A/B testing experiments**
Presenter: **Yiou Li**, DePaul university, United States
A/B tests (or "A/B/n tests") refer to the experiments and the corresponding inference on the treatment effect(s) of a two-level or multi-level controllable experimental factor. The common practice is to use a randomized design and perform hypothesis tests on the estimates. However, such estimation and inference are not always accurate when covariate imbalance exists among the treatment groups. To overcome this issue, we propose a discrepancy-based criterion and show that the design minimizing this criterion significantly improves the accuracy of the treatment effect(s) estimates. The discrepancy-based criterion is model-free and thus makes the estimation of the treatment effect(s) robust to the model assumptions. More importantly, the proposed design is applicable to both continuous and categorical response measurements. We develop two efficient algorithms to construct the designs by optimizing the criterion for both offline and online A/B tests. Through simulation study and a real example, we show that the proposed design approach achieves good covariate balance and accurate estimation.

E0776: **A nonparametric test of 2-factor factorial designs with unequal replicates for multivariate data**
Presenter: **Marcus Jude San Pedro**, School Statistics, University of the Philippines Diliman, Philippines
Co-authors: Mara Sherlin Talento, Erniel Barrios
MANOVA requires equal replicates for treatment with factorial design for its robustness to normality assumption and orthogonality of the basis of vector space. However, when data set is composed of Poisson and Normal error terms and replicates are unequal, normality and sphericity assumption will be rejected. The aim is to propose a nonparametric approach in testing interaction effect of two-factor factorial design even for highly unbalanced replicates and multivariate data with Poisson random variable. Results have shown small probability of committing type 1 error and increasing power for increasing effect of second treatment, increasing coefficient of interaction effect and increasing mean response of levels of second treatment effect, holding other factors constant.

EO067 Room AT242 CURRENT DEVELOPMENTS IN INDUSTRIAL AND ECOLOGICAL STATISTICS Chair: Tsung-Jen Shen

E0355: **Using a generalized area-based truncated model to resolve Fishers paradox when extrapolating biodiversity**
Presenter: **Youhua Chen**, Chengdu Institute of Biology, Chinese Academy of Sciences, China
Why are so many tropical tree species hyper-rare? Do they really have only one or two individuals on Earth? This question, the so-called Fishers paradox, was put forward by S.P. Hubbell when applying Fishers logseries to estimate tropical tree diversity. Herein, we developed an area-based truncated Fishers logseries model to partially, if not completely, resolve Fishers paradox by assuming that the occurrence of too-rare species is impermissible globally while being possible at local scales due to limited sampling efforts. An empirical test showed that alternative truncated models were indistinguishable at the local forest-plot scale, but they could be told apart at the regional scale. By comparison, a protracted speciation neutral model had similar behaviors. However, the exceptional merit of the truncated model is that by using a small truncation threshold, the prediction of regional species richness was similar to the value predicted by the original Fishers logseries, while completely excluding the possibility of the occurrence of too-rare species. Given the issue of the inability to distinguish among alternative models at the local scale, the truncation threshold might be pre-set by referring to real-world population sizes of trees. Alternatively, the threshold can be estimated if sufficient local biodiversity data are provided.

E0393: **On the design of experiments with ordered treatments**
Presenter: **Ori Davidov**, University of Haifa, Israel
There are many situations where one expects an ordering among $K \geq 2$ experimental groups or treatments. Although there is a large body of literature dealing with the analysis under order restrictions, surprisingly, very little work has been done in the context of the design of experiments. A principled approach to the design of experiments with ordered treatments is provided. In particular we propose two classes of designs which are optimal for testing different types of hypotheses. The theoretical findings are supplemented with thorough numerical experimentation and a concrete data example. It is shown that there is a substantial gain in power, or alternatively a reduction in the required sample size, when an experiment is both designed and analyzed using methods which account for order restrictions.

E0649: **Data-driven multistratum designs with the generalized Bayesian D-D criterion for highly uncertain models**
Presenter: **Chang-Yun Lin**, National Chung Hsing University, Taiwan
Multistratum designs have gained much attention recently. Most criteria, such as the D criterion, select multistratum designs based on a given model that is assumed to be true by the experimenters. However, when the true model is highly uncertain, the model used for selecting the optimal design can be seriously misspecified. If this is the case, then the selected multistratum design will be not efficient for fitting the true model. To deal with the problem of high uncertain models, we propose the generalized Bayesian D-D (GBDD) criterion, which selects multistratum designs based on the experimental data. Under the framework of multistratum structures, we develop theorems and formula that are used for conducting Bayesian analysis and extracting information about the true model from the data to reduce model uncertainty. The GBDD criterion is easy and flexible in use. We provide several examples to demonstrate how to construct the GBDD-optimal split-plot, strip-plot, and staggered-level designs. By comparing with the D-optimal designs and one-stage generalized Bayesian D-optimal designs, we show that the GBDD-optimal designs have higher efficiency on fitting the true models. The extensions of the GBDD criterion for more complicated cases, such as more than two stages of experiments and more than one class of potential terms, are also developed.

E0694: **Interaction-based variable selection approach for supersaturated design analysis**
Presenter: **Ray-Bing Chen**, National Cheng Kung University, Taiwan
Co-authors: Huei-Lun Siao, Inchi Hu, Shaw-Hwa Lo, Mong-Na Lo Huang
Supersaturated design is a well-organized design aiming at obtaining as much information as possible although while containing fewer factors involved. As it is not possible to estimate all effects in the experiment due to its size limitation, the main purpose of these types of designs is to discover influential factors under the factor sparsity assumption. We propose to identify the factors based on an influential measurement previously proposed. One major advantage of the new screening procedure is to be able to identify those factors with interaction effects influencing the experimental results. After influential factors and interaction effects are identified with fewer factors, we will apply the componentwise Gibbs sampler methodology to improve the accuracy of obtaining the exact set of significant factor effects. We will examine the effectiveness of the proposed two-stage analysis approach using simulations as well as four well-studied real examples in the literature. Finally, we compare our newly proposed method with others to examine the performances of screening factor effects.

EO013 Room AT335 BAYESIAN APPROXIMATE INFERENCE ALGORITHMS Chair: David Nott

E0496: **High-dimensional copula variational approximation through transformation**
Presenter: **David Nott**, National University of Singapore, Singapore
Co-authors: Michael Smith, Ruben Loaiza Maya

Variational approximation methods are attractive for computing posterior inferences for highly parametrized models and large datasets. They approximate a target distribution - either the posterior or an augmented posterior - using a simpler distribution that is selected to balance accuracy with computational feasibility. We approximate an element-wise parametric transformation of the target distribution as multivariate Gaussian or skew-normal. Approximations of this kind are copula models for the original parameters, with an implicit Gaussian or skew-normal copula function and flexible parametric margins. A key observation is that their adoption can improve the accuracy of variational inference in high dimensions at limited computational cost. To illustrate, we consider the Yeo-Johnson and G&H transformations of the target distribution, along with sparse factor structures for the scale matrix of the Gaussian or skew-normal. We also show how to implement efficient reparametrization gradient methods for these implicit copula models. The efficacy of the approach is illustrated in a number of examples. In each case, we show that the proposed copula model distributions can be more accurate variational approximations than the equivalent Gaussian distributions, but at only a minor increase in computational cost.

E0543: **An easy-to-use empirical likelihood ABC method**
Presenter: **Sanjay Chaudhuri**, National University of Singapore, Singapore
Co-authors: Subhroshekhar Ghosh, David Nott, Kim Cuc Pham

Many scientifically well-motivated statistical models in natural, engineering and environmental sciences are specified through a generative process, but in some cases, it may not be possible to write down a likelihood for these models analytically. Approximate Bayesian computation (ABC) methods, which allow Bayesian inference in these situations, are typically computationally intensive. Recently, computationally attractive empirical likelihood based ABC methods have been suggested in the literature. These methods heavily rely on the availability of a set of suitable analytically tractable estimating equations. We propose an easy-to-use empirical likelihood ABC method, where the only inputs required are a choice of summary statistic, its observed value, and the ability to simulate summary statistics for any parameter value under the model. It is shown that the posterior obtained using the proposed method is consistent, and its performance is explored using various examples.

E0545: **Approximate Bayesian inference for Potts models**
Presenter: **Yanan Fan**, University of New South Wales, Australia

Markov random fields play an important role in image analysis. A well-known problem is that for data on a large lattice, computation of the normalising constants quickly becomes intractable. We propose two approaches to overcome this problem on a regular lattice. In both cases, some form of approximation is used, with computational efficiency being the main trade-off.

E0629: **Variational Bayes estimation of discrete-margined copula models with application to time series**
Presenter: **Michael Smith**, University of Melbourne, Australia
Co-authors: Ruben Loaiza Maya

A new variational Bayes estimator is proposed for high-dimensional copulas with discrete, or a combination of discrete and continuous, margins. The method is based on a variational approximation to a tractable augmented posterior, and is faster than previous likelihood-based approaches. We use it to estimate drawable vine copulas for univariate and multivariate Markov ordinal and mixed time series. These have dimension rT, where T is the number of observations and r is the number of series, and are difficult to estimate using previous methods. The vine pair-copulas are carefully selected to allow for heteroskedasticity, which is a feature of most ordinal time series data. When combined with flexible margins, the resulting time series models also allow for other common features of ordinal data, such as zero in inflation, multiple modes and under- or over-dispersion. Using six example series, we illustrate both the flexibility of the time series copula models, and the efficacy of the variational Bayes estimator for copulas of up to 792 dimensions and 60 parameters. This far exceeds the size and complexity of copula models for discrete data that can be estimated using previous methods.

EO261 Room U302 CIRCULAR STATISTICS AND ITS RELATED TOPICS Chair: Hiroaki Ogata

E0304: **Frechet-Hoeffding copula bounds for circular data**
Presenter: **Hiroaki Ogata**, Tokyo Metropolitan University, Japan

A simple extension of the Frechet-Hoeffding copula bounds for circular data is proposed. Copulas are a powerful tool for describing the dependency of random variables. In two dimensions, Frechet-Hoeffding upper (lower) bound indicates the perfect positive (negative) dependence between two random variables. However, for circular random variables, the usual concept of dependency is not accepted, because of their periodicity. We redefine the Frechet-Hoeffding bounds and consider modified Frechet and Mardia families of copulas for modelling the dependency of two circular random variables. Simulation studies are also given to show the behavior of the model.

E0386: **Recent cylindrical models and their applications**
Presenter: **Toshihiro Abe**, Nanzan University, Japan

Correlation/covariance is a fundamental concept, and multivariate analyses often begin by investigating their presence in given multivariate data. In particular, if an objective variable is not a scalar but a vector with correlations, to construct a statistical model, we need multivariate probability distributions that can flexibly capture the correlations between the objective variables. For example, the multivariate Gaussian distribution is used in the Gaussian process and the geostatistical process to express stochastic uncertainty in temporally or spatially correlated objective variables. By focusing on cylindrical distributions, we show examples of cylindrical data. After a brief review of probability distributions on the line and on the circle, we introduce WeiSSVM distribution. Using a statistical model of forest tree data in Finland, we demonstrate an application of the cylindrical distributions to quantify the factors that affect asymmetric crown expansion.

E0387: **Testing symmetry of a mode and anti-mode preserving distributions on the circle**
Presenter: **Takayuki Shiohama**, Tokyo University of Science, Japan
Co-authors: Toshihiro Abe, Yoichi Miyata

The problem for testing symmetry of a circular distribution function is considered. Under the assumption that the underlying distribution function has mode and anti-mode preserving properties, we propose two types of test statistics for the null hypothesis of symmetry for the circular distribution. These test statistics are based on a circular distribution function and density function. Asymptotic distributions of the test statistics are investigated. Numerical simulations are provided to investigate the performances of the size and power for the proposed test statistics. Several symmetric tests using real data analysis are also illustrated.

E0400: **Discretized circular distribution**
Presenter: **Tomoaki Imoto**, University of Shizuoka, Japan
Co-authors: Kunio Shimizu
In many diverse scientific fields, there often appear data considered as points on a unit circle, called circular data, such as wind directions at a monitoring site, vanishing angles of birds from a starting point, intensive care unit arrival times on the 24-hour clock. For modeling circular data, many continuous circular distributions (CCDs) have been constructed using several methods such as projection, conditioning and maximizing entropy and so on. In practice, for different reasons, observed values are discretized. We propose a method to construct a discrete circular distribution (DCD) from a CCD. The pmf is defined to take the normalized values of the pdf at some pre-fixed equidistant points on the circle. When the pdf is represented by a Fourier series, the constructed pmf is concisely expressed by the cosine moments of the CCD. Simulation studies show that DCDs outperform the corresponding CCDs in modeling grouped circular data, and that minimum chi-square estimation is better than maximum likelihood estimation when the number of groups on the circle is not large.

EO029 Room U414 ADVANCES IN PRODUCTIVITY AND EFFICIENCY MODELLING Chair: Artem Prokhorov

E0231: **Evaluating the CDF of the distribution of the stochastic frontier composed error**
Presenter: **Wen-Jen Tsay**, Academia Sinica, Taiwan
In the stochastic frontier model, the composed error is the sum (or difference) of a normal and a half normal random variable. Often the composed error is linked to other errors using a copula, and evaluation of the copula requires evaluation of the cdf of the composed error. There is no analytical expression for this cdf, though there are several approximations. We propose a computationally efficient simulation based method of evaluation and use it to evaluate the accuracy of these approximations. We also derive the exact cdf of the composed error for the special case that the stochastic frontier relative variance parameter equals one, and we use this expression to investigate the accuracy of our evaluations and the existing approximations.

E0277: **Spatial autoregressive panel stochastic frontier models with fixed effects**
Presenter: **Kien C Tran**, University of Lethbridge, Canada
Co-authors: Levent Kutlu, Mike Tsionas
The estimation of spatial autoregressive (SAR) panel stochastic frontier model with fixed effects is considered. We apply the closed skew normal results to obtain the distribution of the first-difference transformation of the composed-error and the log-likelihood function. Indirect, direct and total marginal effects and inefficiency are calculated. Extension of the model to allow for endogenous regressors is also provided. Monte Carlo simulations indicate good finite-sample performance of the proposed approach. An empirical application of the proposed model and approach is presented.

E0404: **The panel stochastic frontier model with endogenous inputs and correlated random components**
Presenter: **Hung-pin Lai**, National Chung Cheng University, Taiwan
Co-authors: Subal Kumbhakar
The four-component stochastic frontier (4CSF) panel model, where the inputs are endogenous and correlated with the composite error, is considered. We assume that the correlation of inputs can be with one or more of the inefficiency components, or with all four random components in the production function. Furthermore, we allow correlation between the time-invariant and time-varying components. This correlation can arise in various ways. For example, the correlation can arise due to dependence between (i) the long- and short-run inefficiency components, (ii) firm-effects with short-run inefficiency, (iii) firm-effects and the noise term. We propose a three-step procedure to estimate the model parameters. In the first step, we use either within or difference transformation to eliminate the time invariant endogenous components. We use a previous approach to generate the instruments and obtain unbiased and consistent estimator the parameters in the frontier part, except the intercept. In the second step we use the maximum likelihood procedure to estimate the parameters associated with the distributions of the time-varying random components. In the third step, we estimate the intercept and the remaining parameters. We propose using copula approach to model the dependence between the time-varying and time-invariant components.

E0809: **A vine copula approach to estimation of a production system with technical and allocative inefficiency**
Presenter: **Jian Zhai**, University of Sydney, Australia
Co-authors: Artem Prokhorov
Models are considered where an equation for the stochastic frontier production function is complemented by the first order conditions for cost minimization. Such models allow for both technical and allocative inefficiencies. A recent paper proposed a new family of copulas that is appropriate for modelling dependence between that two types of inefficiency. The copula is called APS-d, where d is the number of inefficiency terms. We consider a vine copula construction that makes use of multiple APS-2 copulas and a Gaussian copula. We argue that this construction is natural for this type of productivity models and achieves a more comprehensive coverage of dependence than APS-d, where d can be substantially greater than 2. We also study the extra precision in the estimation of technical inefficiency scores, permitted by the use of dependence information among the error terms in such models of production.

EO225 Room U502 ADVANCES IN CAUSAL INFERENCE I Chair: Luke Keele

E0184: **Flexible models for time-dependent causal mediation**
Presenter: **Jason Roy**, Rutgers University, United States
Suppose interest is in estimating time-varying direct and indirect effects of exposure from observational data. We propose a Bayesian nonparametric approach for modeling the observed data. We then develop causal identifying assumptions and computational methods. We assess the performance of the method and compare with alternative approaches using simulation studies. Finally, we implement the method on longitudinal air pollution data.

E0186: **Causal estimation of scaled treatment effects with multiple outcomes**
Presenter: **Nandita Mitra**, University of Pennsylvania, United States
Co-authors: Edward Kennedy
Typical study designs aim to learn about the effects of an intervention on a single outcome; in many clinical studies, however, data on multiple outcomes are collected and it is of interest to explore effects on multiple outcomes simultaneously. Such designs can be particularly useful in patient-centered research, where different outcomes might be more or less important to different patients. We propose scaled effect measures (via potential outcomes) that translate effects on multiple outcomes to a common scale, using mean-variance and median-interquartile range based standardizations. We present efficient, nonparametric, doubly robust methods for estimating these scaled effects and for testing the null hypothesis that treatment affects all outcomes equally. We also discuss methods for exploring how treatment effects depend on covariates (i.e., effect modification). In addition to describing efficiency theory for our estimands and the asymptotic behavior of our estimators, we illustrate the methods using data from a community health worker study. Importantly, and in contrast to much of the literature concerning effects on multiple

outcomes, our methods are nonparametric and can be used not only in randomized trials to yield increased efficiency, but also in observational studies with high-dimensional covariates to reduce confounding bias.

E0236: **Instrumental variables estimation in cluster randomized trials with noncompliance**
Presenter: **Luke Keele**, University of Pennsylvania, United States
Many policy evaluations occur in settings with treatment randomized at the cluster level and there is treatment noncompliance at the unit level within each cluster. For example, villages might be assigned to treatment and control, but residents in each village may choose to not comply with their assigned treatment status. When noncompliance is present, investigators may choose to focus attention on either intention to treat effects or the treatment effect among the units that comply. When analysts focus on the effect among compliers, the instrumental variables framework can be used to evaluate identify and estimate causal effects. While a large literature exists on instrumental variables estimation methods, relatively little work has been focused on settings with clustered treatments. We review extant methods for instrumental variable estimation in clustered designs. We then show that these methods depend on assumptions that are often unrealistic in applied settings. In response, we develop an estimation method that relaxes these assumptions. Specifically, our method allows for possible treatment effect heterogeneity that is correlated with cluster size and uses a finite sample variance estimator.

E0621: **Evaluating instrumental variable assumptions using randomization tests**
Presenter: **Zach Branson**, Harvard University, United States
Instrumental variable (IV) analyses are a common approach for estimating causal effects in observational studies, where units are non-randomized to treatment. Even though treatment is non-randomized, estimating causal effects is tenable if there is an instrument that affects outcomes only through its correlation with covariates. Thus, a fundamental assumption of IV approaches is that the instrument is as-if randomly assigned to units. There are several falsification tests in the literature for this assumption, which compare balance on observed covariates by IV status to balance on observed covariates by treatment status, with the hope that the former is better than the latter. Adding to this literature, we propose a randomization test for this assumption. We use the balance that would have been produced under randomization as a standard by which to compare IV balance and treatment balance. A benefit of our test over other tests is that it can be used to perform a global balance assessment across covariates as well as assessments on individual covariates. Furthermore, these tests can incorporate blocking information if IV assignment depends on blocks. We demonstrate this approach using a recent application using bed availability in the ICU as an instrument for admission to the ICU.

EO065 Room U517 RECENT DEVELOPMENTS IN STATISTICS FOR COMPLEX DEPENDENT DATA Chair: Shih-Feng Huang

E0341: **Symbolic interval-valued data analysis for time series based on auto-interval-regressive models**
Presenter: **Liang-Ching Lin**, National Cheng Kung University, Taiwan
Interval-valued time series data are considered. To characterize interval time series data, we propose an auto-interval-regressive (AIR) model using the order statistics from normal distributions. Furthermore, to better capture heteroscedasticity in volatility, we designate an autoregressive conditional heteroscedasticity (AIR-ARCH) model. The likelihood functions of AIR and AIR-ARCH models are derived to obtain the maximum likelihood estimator. The corresponding predicted formulae are also given. Monte Carlo simulations are conducted to evaluate our methods of estimation, confirming their validity. Real data example is also carried out for the S&P 500 Index for illustration.

E0371: **Option hedging and parameter estimation of pricing models**
Presenter: **Lo-Bin Chang**, Ohio State University, United States
Typical parameter estimation methods such as the maximum likelihood estimation and implied parameter estimation rely upon the validity of model assumptions. Given the well-known fact that no model is perfect, the estimation criterion should be selected to adapt to the practical usage of the model. Focusing on option pricing and hedging, a novel criterion for estimating parameters is proposed, which is based on the magnitude of cumulative hedging error over the option lifetime. The theoretical property of this criterion for the Black-Scholes model is provided. Back-testing experiments of delta hedging with real stock data show that for both Black-Scholes and Merton's jump-diffusion models, the proposed hedging-optimization estimation results in better hedging performance over the maximum likelihood estimation and implied parameter estimation.

E0390: **Asymptotic theory of maximum likelihood estimators in mixed-effects models with a fixed number of clusters**
Presenter: **ChihHao Chang**, National University of Kaohsiung, Taiwan
A linear mixed-effects model with clustered structure is considered, where the parameters are estimated by maximum likelihood (ML) based on possibly unbalanced data. Inference of this model is typically done based on the asymptotic theory assuming that the number of clusters tends to infinity with the sample size. However, when the number of clusters is fixed, the traditional asymptotic theory developed under a divergent number of clusters is no longer valid. We establish the asymptotic properties of the ML estimators of the random-effects parameters under a general setting, including models with fixed numbers of clusters. The asymptotic theorems allow both the fixed effects and the random effects to be misspecified, and the dimensions of both effects to go to infinity with the sample size.

E0436: **Parameter estimation for misspecified diffusion with market microstructure noise**
Presenter: **Teppei Ogihara**, University of Tokyo, Japan
Statistical inference for stock prices modeled by diffusion processes with high-frequency observations is considered. In particular, we study parametric inference under the existence of market microstructure noise and nonsynchronous observations. We first consider maximum-likelihood-type estimation for parametric diffusion processes with noisy, nonsynchronous observations, assuming that the true model is contained in the parametric family. We show asymptotic mixed normality of the estimator with the convergence rate $n^{-1/4}$. We also see local asymptotic normality of the statistical model when coefficients of the stochastic differential equation are deterministic,and show asymptotic efficiency of the estimator. In practice for high-frequency financial data, it is not easy task to choose parametric family so that the true model is contained in the parametric family. The statistical model without this assumption is called 'misspecified model'. In this setting, the maximum-likelihood-type estimator cannot attain the optimal convergence rate $n^{-1/4}$ due to the asymptotic bias. We construct a new estimator which attains the optimal rate by using a bias correction and show the asymptotic mixed normality.

EG006 Room U301 CONTRIBUTIONS IN PORTFOLIO STRATEGIES Chair: Tomas Tichy

E0235: **Portfolio optimization based on forecasting models using vine copulas: An empirical assessment for the financial crisis**
Presenter: **Andreas Stephan**, Jonkoping University, Sweden
Co-authors: Maziar Sahamkhadam

Vine copulas are employed and examined in modeling the symmetric and asymmetric dependency structure and forecasting of financial returns. Asset allocation is performed during the 2007-2010 financial crisis and different portfolio strategies are tested including maximum reward-to-risk ratio, minimum variance and minimum conditional Value-at-Risk. Regular, drawable and canonical vine copulas are specified including Clayton, Frank, Joe and mixed copula. Both in-sample and out-of-sample analyses of portfolio performances are conducted. The out-of-sample portfolio back-testing shows that vine copulas reduce portfolio risk more than simple copulas. Considering portfolio out-of-sample CVaR, Frank and mixed vine copulas result in lower downside risk. The results of the VaR back-testing shows improvement in forecasting of the downside risk for all portfolio strategies obtained from using simple Clayton and mixed copula families, implying time-varying tail dependence of stock market returns. Copula families which capture no tail dependence (Frank) and upper tail dependence (Joe) lead to higher terminal values of portfolios over the financial crisis.

E0519: **Nonparametric performance hypothesis testing with the information ratio**
Presenter: **Jacque Bon-Isaac Aboy**, University of the Philippines, Philippines

A nonparametric bootstrap-based approach to testing performance of portfolios built through style investing and optimized through various diversification strategies is proposed. The test procedure is distribution-free and thus will not assume a special parametric distribution for the portfolio returns. The test was able to determine which styles are superior and which diversification strategy is optimal by testing which among them showcase a stochastically higher Information ratio. This test will help individual investors and the authorities in the financial market decide and choose which style portfolio to invest and which diversification strategy to use.

E0777: **Graphical least squares estimation and its application in portfolio optimisation**
Presenter: **Saeed Aldahmani**, United Arab Emirates University, United Arab Emirates

The Graphical Least Square Estimator (GLSE) has been proposed as an unbiased estimation method based on graphical models to be used for regression when the number of variables is much larger than the number of observations ($p > n$). Although this estimator can have useful applications, such as in portfolio optimisation in the world of finance, its main limitation is the heavy computational cost it entails when handling large datasets. To overcome this problem, a new algorithm is presented which heavily reduces the dimensionality of the search space during the process of finding the optimal graph. In other terms, the new algorithm facilitates the application of GLSE to the optimisation of the portfolio where all the assets in the portfolio need to be estimated without any bias, which as a result can lead to an optimal portfolio with a higher Sharpe ratio. In order to assess the proposed algorithm, three simulation scenarios and a data analysis from New York Stock Exchange are carried out. The results reveal that the algorithm helps to improve GLSEs performance and makes it superior to other existing portfolio optimisation methods such as the ridge.

E0808: **Portfolio strategies involving options and systemic risk alarm**
Presenter: **Tomas Tichy**, VSB-TU Ostrava, Czech Republic

Portfolio selection strategies involving stocks and options are further developed. We generally assume that the returns follow Markov processes that are approximated with proper Markov Chains. We consider investors with various risk attitude or utility function and apply complex rules utilizing systemic risk alarm identification. As an alternative, an alarm based on fuzzy logic and fuzzy natural logic is considered. Empirical results are based on examination of the US market.

Tuesday 25.06.2019 16:10 - 17:25 Parallel Session D – EcoSta2019

EO093 Room UB99(B1) LARGE-SCALE MULTIVARIATE MODELING OF FINANCIAL ASSET RETURNS Chair: Marc Paolella

E0288: **Dynamic currency hedging strategy with a common market factor non-Gaussian returns model**
Presenter: **Urban Ulrych**, University of Zurich and Swiss Finance Institute, Switzerland
Co-authors: Walter Farkas, Pawel Polak

A new foreign currency hedging strategy for international investors is motivated and studied. Model-free optimal foreign currency exposures for a risk averse investor are derived. Based on those, and assuming a very flexible non-Gaussian returns model for currency and portfolio returns, we build a dynamic currency hedging strategy. In the context of our model, each element of the vector return at time t is endowed with a common univariate shock, interpretable as a common market factor. It is shown that this mixing random variable plays the role of ambiguity (uncertainty about the return distribution), where its magnitude is expressed through the size of the market factor's conditional variance. Using the derived theoretical model and the proposed dynamic hedging strategy, an out of sample back test on the historical market data is performed. The results show that the approach yields a robust and highly risk reductive hedging strategy, obtainable with low transaction costs.

E0422: **On the use of random forest for two sample testing, with applications in empirical finance**
Presenter: **Simon Hediger**, University of Zurich, Switzerland
Co-authors: Jeffrey Naef

Several tests based on the Random Forest classifier are derived for the two-sample case. The tests are easy to use, require no tuning parameters, and are applicable for any p-dimensional distribution, notably for large p. The distribution of the test under the null of distributional equality is derived. Power analysis is conducted, both in theory, and with simulations. The R-package "hypoRF" contains the relevant codes. An application relevant to large-scale multivariate non-Gaussian GARCH models for financial asset returns is developed.

E0730: **Smooth FREE COMFORT and LASSO for GARCH groups**
Presenter: **Marc Paolella**, University of Zurich, Switzerland

GARCH-based multivariate models for asset returns can outperform their IID counterparts, but often only when not accounting for transaction costs, because of the high turnover induced by the ever-changing dispersion matrix. A setup is considered based on a smooth transition between IID and GARCH-based models, and presents a new method for the latter, termed Fast REduced Estimation (FREE), which cuts estimation time significantly with little or no performance degradation. A further enhancement is introduced that effectively clusters the univariate series into groups with the same GARCH structure. It is conducted using a new LASSO-based shrinkage paradigm, conditional on a filtered latent sequence of mixing random variables, and thus differs completely from related attempts at GARCH clustering.

EO211 Room S101 STATISTICAL MACHINE LEARNING METHODS FOR DATA SCIENCE Chair: Yen-Chi Chen

E0435: **Statistical inference with local optima**
Presenter: **Yen-Chi Chen**, University of Washington, United States

Many statistical analyses involve finding the maximum of an objective function. This is often done by applied a gradient-ascent type method such as the EM algorithm. When the objective function has multiple local maxima, there is no guarantee that maximum we obtain is truly the global maximum. Thus, many statistical inference such as confidence intervals and hypothesis test may not have the desired properties. We investigate the effect of local optima on statistical inference. We will discuss how the estimation theory and notions of coverage of a confidence interval has to be modified to account for its effect.

E0610: **Principal subspace analysis for high-dimensional compositional data**
Presenter: **Wei Lin**, Peking University, China
Co-authors: Jingru Zhang

Dimension reduction for high-dimensional compositional data plays an important role in many scientific studies, where the principal subspace of the basis covariance matrix is the parameter of interest. However, in practice, the basis variables are latent and rarely observed. The fact that the observed compositions lie in a simplex renders standard techniques inappropriate. To address this challenging problem, we relate the basis covariance to the centered log-ratio compositional covariance. We prove that the principal subspace of the basis covariance matrix is approximately identifiable as the dimensionality tends to infinity under some subspace sparsity assumptions, and derive nonasymptotic error bounds for the subspace estimation. Our theoretical analysis shows that the sparsity assumption not only helps to identify the principal subspace, but also benefits the estimation in high-dimensional settings. Moreover, we develop efficient proximal ADMM algorithms for solving the proposed nonconvex and nonsmooth optimization problems. Simulation results demonstrate that the performance of the proposed methods is nearly as good as the oracle method and significantly superior to those based on different transformations.

E0696: **Causal inference with confounders missing not at random**
Presenter: **Linbo Wang**, University of Toronto, Canada
Co-authors: Shu Yang, Peng Ding

It is important to draw causal inference from observational studies, which, however, becomes challenging if the confounders have missing values. Generally, causal effects are not identifiable if the confounders are missing not at random. We propose a novel framework to nonparametrically identify causal effects with confounders subject to an outcome-independent missingness, that is, the missing data mechanism is independent of the outcome, given the treatment and possibly missing confounders. We then propose a nonparametric two-stage least squares estimator and a parametric estimator for causal effects.

EO119 Room S102 ADVANCED METHODS IN SPATIO-TEMPORAL DATA ANALYSIS IN BIOSTATISTICS Chair: Yuko Araki

E0717: **Functional survival data analysis with direct and indirect effects for high dimensional data**
Presenter: **Yuko Araki**, Shizuoka University, Japan

Statistical methods are introduced for survival analysis which contains complex associations of several variables including very high dimensional intermediate variable. The proposed model deal with such high dimensional data as functional data with devices of basis expansions and sparse PCA for dimension reduction. The model is constructed based on structural equation modeling (SEM) for survival outcome. We extend SEM to be able to discuss a causal inference. The crucial issue is how to select the regularization parameters used in model building process. We introduced a model selection criterion to select this value. The proposed models were evaluated through simulations and real data example.

E0585: **Extension of Poisson point process based on quasi-linear modeling**
Presenter: **Osamu Komori**, Seikei University, Japan
The investigation to clarify the relationship between habitat distribution of some species and the environmental variables is important for its conservation and management purpose. To do this, the maximum entropy method (Maxent) or Poisson point process (PPP) is widely employed using the presence-only data. We propose an extension of PPP based on the quasi-linear modeling to improve the estimation accuracy. The effect of sampling bias is also considered in our model. Some simulation studies and real data analysis are conducted to show its practical utility.

E0324: **Detection and evaluation of multiple clusters in spatial epidemiology**
Presenter: **Kunihiko Takahashi**, Nagoya University Graduate School of Medicine, Japan
Co-authors: Hideyasu Shimadzu
A number of statistical tests have been proposed and are widely used in spatial epidemiology to investigate a regional or temporal tendency in the presence of certain diseases, whether the disease risk is relatively high to other surrounding regions or subsequent time periods. The scan statistic is one of the most powerful elements of the cluster detection test to detect and evaluate spatial and/or temporal disease clusters since it is based on the maximum likelihood ratio; examples include the Kulldorff's circular scan statistic along with the SaTScan software, and Tango and Takahashi's flexibly shaped scan statistic implemented in the FleXScan software. Although multiple clusters in the study space can be thus identified, current theoretical developments are mainly based on detecting a "single" cluster. The standard scan statistic procedure enables the detection of multiple clusters, recursively identifying additional "secondary" clusters. However, their p-values are calculated one at a time, as if each cluster is a primary one. Therefore, we proposed a new test procedure that can accurately evaluate multiple clusters as a whole, combining generalized linear models with an information criterion approach that selects an appropriate number of the clusters. This framework encompasses the conventional detection procedure as a special case. We present practical examples applying the proposed procedure and compare the results with ones by conventional procedures.

EO267 Room S104 HIGH DIMENSIONAL STATISTICS AND APPLICATIONS Chair: Lilun Du

E0342: **Random perturbation of low-rank matrices**
Presenter: **Ke Wang**, Hong Kong University of Science and Technology, Hong Kong
Co-authors: Zhigang Bao, Xiucai Ding
Computing the singular values and singular vectors of a large matrix is a basic task in high dimensional data analysis with many applications in computer science and statistics. In practice, however, data is often perturbed by noise. We consider the matrix model $Y = S + X$ where S is a low-rank deterministic matrix, representing the signal, and X is random noise. We give a precise description of the limiting distribution of the angles between the outlier singular vectors of Y with their counterparts, the leading singular vectors of S. It turns out that the limiting distribution depends on the structure of S and the distribution of X, and thus it is non-universal.

E0347: **Factor modeling for volatility**
Presenter: **Yi Ding**, The Hong Kong University of Science and Technology, Hong Kong
Co-authors: Robert Engle, Yingying Li, Xinghua Zheng
Under a high-frequency and high-dimensional setup, we establish a framework to estimate the factor structure in idiosyncratic and stock volatility. We show that the factor structure in idiosyncratic volatility can be consistently estimated by conducting principal component analysis on the idiosyncratic realized volatilities. Empirically, we confirm and identify the factor structure in idiosyncratic volatilities of S&P 500 Index constituents. Furthermore, motivated by strong empirical evidence, a single-factor volatility model is proposed. Empirical examination of the model reveals that the simple model well explains the co-movement feature of stock volatilities, and leads to substantial gain in forecasting.

E0493: **High dimensional analysis of chi-squared data**
Presenter: **Inchi Hu**, Hong Kong University of Science and Technology, Hong Kong
In a thought-provoking paper, the merit and limitation of an empirical Bayes method to correct selection bias based on Tweedie's formula have been investigated. The virtue of Tweedie's formula for the normal data lies in its representation of selection bias as a simple function of the derivative of marginal log likelihood. Since the marginal likelihood and its derivative can be estimated from the data directly without invoking prior information, bias correction can be carried out conveniently. We propose a Bayesian hierarchical model for chi-squared data such that the resulting Tweedie's formula has the same virtue. Because noncentral chi-squared distributions, the common alternative distributions for chi-squared tests, does not constitute an exponential family, our results cannot be obtained by extending existing results. Furthermore, the corresponding Tweedie's formula manifests new phenomena quite different from those of the normal data and suggests new ways to analyze chi-squared data.

EO251 Room S106 RECENT ADVANCES IN STATISTICAL MODELS AND THEIR APPLICATIONS Chair: Berwin Turlach

E0510: **Group based trajectory modelling with monotonicity constraints**
Presenter: **Kevin Murray**, University of Western Australia, Australia
Co-authors: Berwin Turlach, Michael Dymock
Existing methods for group based trajectory modelling have been developed over the past two decades with the addition of numerous extensions such as the ability to jointly analyse multiple trajectories as well as the handling of missing data. However, it is often known from underlying theory that a response trajectory will be monotonic and commonly used modelling techniques in group based trajectory modelling, such as polynomial regression, may fail to capture this behaviour. A new method for fitting group based trajectory models with monotonicity constraints is described in which an Expectation Maximisation algorithm is adopted in order to avoid the convergence issues commonly encountered by a Newton Raphson algorithm. Model selection and starting values are discussed and the methodology is described and implemented in R. Numerical experiments and an example on lung function data from the West Australian Busselton Health Study is used to illustrate the methodology.

E0453: **Bootstrap model selection for linear mixed models**
Presenter: **Garth Tarr**, University of Sydney, Australia
Co-authors: Alan Welsh, Samuel Mueller
Linear mixed effects models are widely used in applications because they provide flexible models for a variety of types of clustered data. Model selection, which often aims to choose a parsimonious model with other desirable properties from a possibly very large set of candidate statistical models, is a key part of many applications. We discuss the use of bootstrap model selection in linear mixed models. Bootstrap model selection was originally developed for simpler models with independent observations. It is an interesting approach because of the flexibility it allows in permitting the use of measures of fit different from those used to define the estimators used to fit the models. We discuss statistical properties as well as computational issues and present both theoretical and simulation results.

E0729: **Smooth backfitting of proportional hazards with multiplicative components**
Presenter: **Munir Hiabu**, University of Sydney, Australia

A proportional hazard model is proposed, where we assume an underlying conditional hazard with multiplicative components. No structural assumption is made on the components. The model is a generalization of the Cox proportional hazard model where components are assumed to be log-linear. We will fit the model via a smooth backfitting approach. Smooth backfitting has proven to have a number of theoretical and practical advantages in structured regression. It projects the data down onto the structured space of interest providing a direct link between data and estimator. Those ideas will be brought to the field of survival analysis. Asymptotic theory for the proposed estimator will be developed. In a comprehensive simulation study, we show that our smooth backfitting estimator successfully circumvents the curse of dimensionality and outperforms existing estimators. This is especially the case in difficult situations with higher dimensions and/or high correlation where other estimators tend to break down.

EO081 Room S1A01 CHALLENGES FOR FUNCTIONAL AND LARGE DATA Chair: Jane-Ling Wang

E0194: **Mean and covariance estimation for functional snippets**
Presenter: **Zhenhua Lin**, University of California, Davis, United States
Co-authors: Jane-Ling Wang

The focus is on the estimation of the mean function and the covariance function of functional snippets, which are short segments of functions possibly observed irregularly on an individual specific subinterval that is much smaller than the entire study interval. Estimation of the covariance function for functional snippets is challenging since information for the far off-diagonal regions of the covariance structure is completely missing. We address this difficulty by decomposing the covariance function into a variance function component and a correlation function component. The variance function can be effectively estimated by local linear smoothing, while the correlation part is modeled parametrically to handle the missing information in the far off-diagonal regions. Both theoretical analysis and numerical simulations suggest that this divide-and-conquer strategy is effective and efficient. In addition, we propose an efficient estimator for the variance of measurement errors and analyze its asymptotic properties. This estimator is required for the estimation of the variance function from noisy measurements.

E0295: **Detecting change points for a multivariate functional data sequence**
Presenter: **Yu-Ting Chen**, National Chenchi University, Taiwan
Co-authors: Jeng-Min Chiou

Functional data with a multivariate outcome is a common data type. We introduce a procedure for detecting change points for a multivariate functional data sequence based on the developed optimality criterion in defining the functional changes. The approach first searches for change point candidates using the dynamic segmentation that recursively adjusts the endpoints of the subsequences via the optimality criterion. Then, it verifies the statistical significance of these change point candidates by a resampling technique that requires very mild assumptions. We show the consistency property of the algorithm, illustrate the method by a traffic data application, and examine its practical performance through a simulation study.

E0163: **Selecting influential /predictive variables for a large data set**
Presenter: **Shaw-Hwa Lo**, Columbia University, United States

Prediction for very large data set is typically carried out in two stages, variable selection and pattern recognition. Ordinarily variable selection involves seeing how well individual explanatory variables are correlated with the dependent variable using a significance-based criterion. This practice neglects the possible interactions among the explanatory variables, so can choose less-predictive variables, because significance does not imply predictivity and important joint information may be omitted. When a subset of truly influential variables is identified, one may expect a noticeable increase of correct prediction rate, being true in both simple and complex data. However high dimensionality and complicated interactions have posed great difficulties for existing selection procedures. We consider an alternative selection approach that directly measures a variable set's ability to predict (termed "predictivity"), the I-score, without relying on the CV. We argue that the I-score not only reflects the true amount of interactions among variables, it can be related to a lower bound of the correct prediction rate and does not over fit. The values of the I-score measure the amount of "influence" of the variables set under consideration. We suggest searching for a new criterion to locate highly predictive variables using partition retention (PR) method with I-score. The PR was effective in reducing prediction error from 30% to 8% on a long-studied breast cancer data set.

EO023 Room AT241 RECENT ADVANCES IN DESIGN AND ANALYSIS OF EXPERIMENTS Chair: Frederick Kin Hing Phoa

E0299: **Opening up the black box: Gaussian process modeling using information from partial differential equation models**
Presenter: **Matthias Tan**, City University of Hong Kong, Hong Kong

Gaussian process (GP) emulators of computer models are typically constructed based purely on data from a computer experiment using a standard stationary GP prior with product Matern or Gaussian correlation function. This often ignores valuable engineering and mathematical knowledge about the behavior of the computer model. We will focus on the use of known behavior/properties of partial differential equation models solved numerically by computer codes to improve construction of GP emulators for this type of computer models.

E0274: **A construction of cost-efficient designs with guaranteed repeated measurements on interaction effects**
Presenter: **Frederick Kin Hing Phoa**, Academia Sinica, Taiwan
Co-authors: Yasmeen Akhtar

A useful class of cost-efficient designs is introduced for two-level multi-factor experiments. It provided guaranteed repeated measurements on all 2-tuples from any two factors and the number of repetitions was adjusted by the experimenters. Given the number of factors of interest, it utilized less resources than an orthogonal array while its repeated measurement provided a resistance towards outliers that a covering array failed to achieve. To bridge the wide spectrum between two extreme settings (orthogonal arrays and covering arrays) in terms of the number of repeated measures of tuples, we developed a systematic method to construct families of these designs, namely (supersaturated) repeated coverage design, with small run sizes under different number of factors and number of repetitions.

E0785: **Computer experiments with binary time series and applications to cell biology: Modeling, estimation and calibration**
Presenter: **Ying Hung**, Rutgers University, United States

Computer experiments have become ubiquitous in various applications from rocket injector designs to weather forecasts. Although extensive research has been devoted in the literature, computer experiments with binary time-series outputs have received scant attention. Motivated by the analysis of a class of cell adhesion experiments, we introduce a new emulator, as well as a new calibration framework for binary time-series outputs. More importantly, we provide their theoretical properties to ensure the estimation performance in an asymptotic setting. The application to the cell adhesion experiments illustrates that the proposed emulator and calibration framework not only provide an efficient alternative for the computer simulation, but also reveal important insight on the underlying adhesion mechanism, which cannot be directly observed through existing methods.

EO033 Room AT242 RECENT ADVANCES ON SURVIVAL ANALYSIS Chair: Xingqiu Zhao

E0362: **Nonparametric inference for right-censored data using smoothing splines**
Presenter: **Meiling Hao**, University of International Business and Economics, China
Co-authors: Yuanyuan Lin, Xingqiu Zhao

A penalized nonparametric maximum likelihood estimation of the log-hazard function for analyzing right-censored data is introduced. Smoothing splines are employed for a smooth estimation. Our main discovery is a functional Bahadur representation, which serves as a key tool for non-parametric inferences of an unknown function. The asymptotic properties of the resulting smoothing-spline estimator of the unknown log-hazard function are established under regularity conditions. Moreover, we provide a local confidence interval for this function, as well as local and global likelihood ratio tests. We also discuss the asymptotic efficiency of the estimator. The theoretical results are validated using extensive simulation studies. Lastly, we demonstrate the estimator by applying it to a real data set.

E0520: **Incorporating graphical structure of predictors in sparse quantile regression**
Presenter: **Wenlu Tang**, The Chinese University of Hong Kong, China
Co-authors: Zhanfeng Wang, Yuanyuan Lin, Xiaohui Liu

Quantile regression in high dimensional settings is useful in analyzing high dimensional heterogeneous data. Different from existing methods in quantile regression which treat all the predictors equally with the same priori, we take advantage of the graphical structure among predictors to improve the performance of parameter estimation, model selection and prediction in sparse quantile regression. It is shown under mild conditions that the proposed method enjoys the model selection consistency and the oracle properties. An alternating direction method of multipliers (ADMM) algorithm with a linearization technique is proposed to implement the proposed method numerically, and its convergence is justified. Simulation studies are conducted, showing that the proposed method is superior to existing methods in terms of estimation accuracy and predictive power. The proposed method is also applied to a microarray dataset.

E0523: **Exact one- and two-sample likelihood ratio tests based on type-I and joint type-II censored exponential data**
Presenter: **Xiaojun Zhu**, Xi'an Jiaotong-Liverpool University, China

The likelihood ratio test is one of the commonly used procedures for hypothesis testing. Several results on likelihood ratio test have been discussed for testing the scale parameter of an exponential distribution under complete and censored data; however, all of them are based on approximations of the involved null distributions. We first derive the exact distribution of the likelihood ratio statistic for testing the scale parameter of an exponential distribution based on a time-constrained life-testing experiment. We also obtain the asymptotic distribution, for the use in large sample size. We then discuss the derivation of its power function. Next, we consider the likelihood ratio test of $\theta_2 = \gamma_0 * \theta_1$ when data are obtained from two exponential distributions based on a time-constrained life-testing experiment. We derive both exact and asymptotic distributions of the likelihood ratio statistic and then use them to determine the reject region as well as the power function. We then extend the result to the case of Joint Type-II Censored Exponential Data and use the exact power function to design an optimal experiment. Finally, two illustrative examples are presented for demonstrating all the inferential results.

EO259 Room AT335 EAC-ISBA SESSION: BAYESIAN ANALYSIS WITH LARGE COMPLEX DATA Chair: Guanyu Hu

E0659: **A new Bayesian joint model for longitudinal count data with many zeros, intermittent missingness, and dropout**
Presenter: **Jing Wu**, University of Rhode Island, United States
Co-authors: Ming-Hui Chen, Elizabeth Schifano, joseph ibrahim, jeffrey fisher

In longitudinal clinical trials, it is common that subjects may permanently withdraw from the study (dropout), or return to the study after missing one or more visits (intermittent missingness). It is also routinely encountered in HIV prevention clinical trials that there is a large proportion of zeros in count response data. A sequential multinomial model is adopted for dropout and subsequently a conditional model is constructed for intermittent missingness. The new model captures the complex structure of missingness and incorporates dropout and intermittent missingness simultaneously. The model also allows us to easily compute the predictive probabilities of different missing data patterns. A zero inflated Poisson mixed-effects regression model is assumed for the longitudinal count response data. We also propose an approach to assess the ove all treatment effects under the zero-inflated Poisson model. We further show that the joint posterior distribution is improper if uniform priors are specified for the regression coefficients under the proposed model. Variations of the g-prior, Jeffreys prior, and maximally dispersed normal prior are thus established as remedies for the improper posterior distribution. An efficient Gibbs sampling algorithm is developed using a hierarchical centering technique. A modified logarithm of the pseudomarginal likelihood (LPML) is used to compare the models under different missing data mechanisms.

E0267: **Inflated density ratio and its variation and generalization for computing marginal likelihoods**
Presenter: **Yu-Bo Wang**, Clemson University, United States

In the Bayesian framework, the marginal likelihood plays an important role in variable selection and model comparison. The marginal likelihood is essentially the marginal distribution of the data after integrating out the parameters over the parameter space. However, this quantity is often analytically intractable due to the complexity of the model. We first examine the properties of the inflated density ratio (IDR) method, which is a Monte Carlo (MC) method for computing the marginal likelihood using a single MC or Markov chain Monte Carlo (MCMC) sample. We then develop a variation of the IDR estimator, called the dimension reduced inflated density ratio (Dr.IDR) estimator. We further propose a more general identity and then obtain a general dimension reduced (GDr) estimator. Simulation studies are conducted to examine empirical performance of the IDR estimator as well as the Dr.IDR and GDr estimators. We further demonstrate the usefulness of the GDr estimator for computing the normalizing constants in a case study on the inequality-constrained analysis of variance.

E0351: **Bayesian nonparametric nonhomogeneous Poisson process with applications**
Presenter: **Guanyu Hu**, University of Connecticut, United States

Intensity estimation is a common problem in statistical analysis of spatial point pattern data. A nonparametric Bayesian method is proposed for estimating the spatial point process intensity based on mixture of finite mixture (MFM) model. The MFM approach leads a consistent estimate on the intensity of spatial point patterns in different areas while considering heterogeneity. An efficient Markov chain Monte Carlo (MCMC) algorithm is proposed for our methods. Extensive simulation studies are carried out to examine empirical performance of the proposed methods. The usage of our proposed methods is further illustrated with the analysis of the Earthquake Hazards Program of United States Geological Survey (USGS) earthquake data.

EO229 Room AT337 RECENT DEVELOPMENTS IN MEASUREMENT ERROR AND MISSING DATA Chair: Jae Kwang Kim

E0210: **Causal inference with measurement error in outcomes**
Presenter: **Grace Yi**, University of Waterloo, Canada
Inverse probability weighting (IPW) estimation has been popularly used to consistently estimate the average treatment effect (ATE). Its validity, however, is challenged by the presence of error-prone variables. In applications, measurement error is ubiquitously present in data collection due to various reasons. Naively ignoring measurement error effects usually yields biased inference results. We will discuss the IPW estimation with mismeasured outcome variables. The impact of measurement error for both continuous and discrete outcome variables will be examined. We will describe estimation procedures with the outcome misclassification effects accommodated. Consistency and efficiency will be investigated. Numerical studies will be reported to assess the performance of the proposed methods.

E0383: **Locally efficient semiparametric estimators for a class of Poisson models with error-prone covariates**
Presenter: **Jianxuan Liu**, Syracuse University, United States
Co-authors: Yanyuan Ma
The presence of measurement error may cause bias in parameter estimation and can lead to incorrect conclusions in data analyses. Despite a large body of literature on general measurement error problems, relatively few works exist to handle Poisson models. We thoroughly study Poisson models with errors in covariates and propose consistent and locally efficient semiparametric estimators. The resultant estimators are shown to be root-n consistent, asymptotically normal and locally efficient. We assess the finite sample performance of the estimators through extensive simulation studies and illustrate the proposed methodologies by analyzing data from the Stroke Recovery in Underserved Populations Study.

E0663: **Semiparametric response model with nonignorable nonresponse**
Presenter: **Jae Kwang Kim**, Iowa State University, United States
Co-authors: Masatoshi Uehara
How to deal with nonignorable response is often a challenging problem encountered in statistical analysis with missing data. Parametric model assumption for the response mechanism is often made and there is no way to validate the model assumption with missing data. We consider a semiparametric response model that relaxes the parametric model assumption in the response mechanism. Two types of efficient estimators, profile maximum likelihood estimator and profile calibration estimator, are proposed and their asymptotic properties are investigated. Two extensive simulation studies are used to compare with some existing methods. We present an application of our method using Korean Labor and Income Panel Survey data.

EO207 Room U301 NONPARAMETRIC REGRESSION AND STATISTICAL INFERENCE Chair: Kuang-Yao Lee

E0459: **On post dimension reduction statistical inference**
Presenter: **Bing Li**, The Pennsylvania State University, United States
The methodologies of sufficient dimension reduction have undergone extensive developments in the past three decades. However, there has been a lack of systematic and rigorous development of post dimension reduction inference, which has seriously hindered its applications. The current common practice is to treat the estimated sufficient predictors as the true predictors and use them as the starting point of the downstream statistical inference. However, this naive inference approach would grossly overestimate the confidence level of an interval, or the power of a test, leading to the distorted results. We develop a general and comprehensive framework of post dimension reduction inference, which can accommodate any dimension reduction method and model building method, as long as their corresponding influence functions are available. Within this general framework, we derive the influence functions and present the explicit post reduction formulas for the combinations of numerous dimension reduction and model building methods. We then develop post reduction inference methods for both confidence interval and hypothesis testing. We investigate the finite sample performance of our procedures by simulations and a real data analysis.

E0471: **A three-step nonparametric regression approach for analyzing spatio-temporal data**
Presenter: **Peihua Qiu**, University of Florida, United States
Spatio-temporal data are common in practice. Research in proper analysis of spatio-temporal data has attracted much attention in different research communities, including statistics, epidemiology, geography, oceanography, environmental science, and more. Existing methods in the literature often employ parametric modelling with different sets of model assumptions. However, spatio-temporal data in practice usually have complicated structures, including complex spatial and temporal data variation, spatio-temporal data correlation, and data distribution. Because such data structures reflect the complicated impact of confounding variables, such as weather, demographic variables, life styles, and other cultural and environmental factors, they are usually too complicated to be described well by parametric models. We discuss a general modelling framework for analyzing spatio-temporal data based on nonparametric spatio-temporal regression. The suggested model and its estimation can well accommodate the complicated structure of real spatio-temporal data described above. Both theoretical arguments and numerical studies show that our proposed method could work well in practice.

E0596: **The CCP selector: Best subset selection for sparse regression from chance-constrained programming**
Presenter: **Xinwei Deng**, Virginia Tech, United States
Sparse regression and variable selection for large-scale data have been rapidly developed in the past decades. The focus is on considering the exact L_0 norm to pursue the sparse regression. We pave out a theoretical foundation to understand why many existing approaches may not work well for this problem, in particular on large scale datasets. Inspired by reformulating the problem as a chance-constrained program, we derive a novel mixed integer second order conic (MISOC) reformulation. Based the reformulation, we develop new scalable algorithms for sparse ridge regression with desirable theoretical properties. The proposed algorithms are proved to yield near-optimal solutions under mild conditions. The merits of the proposed methods are elaborated through a set of numerical examples in comparison with several existing ones.

EO308 Room U302 RECENT ADVANCES IN HIGH DIMENSIONAL TIME SERIES ANALYSIS Chair: Yue Niu

E0724: **Variance estimation for change-point model**
Presenter: **Yue Niu**, University of Arizona, United States
The variance of noise plays an important role in many change-point detection methods. For example, in binary segmentation and related methods, the variance is required to decide when to stop the procedure. In practice, people usually use conventional methods to estimate the variance based on residuals. However, these methods may be problematic when there are many change points. We will introduce a variance estimation method and show its advantage.

E0734: **Simultaneous prediction intervals for high-dimensional vector autoregressive model**
Presenter: **Mengyu Xu**, University of Central Florida, United States
Co-authors: Sayar Karmakar, Jayanta Kapat
The simultaneous prediction intervals for high-dimensional vector autoregressive model are studied. We consider a de-biased calibration for the lasso prediction and propose a Gaussian-multiplier bootstrap based method for one-step ahead prediction. The asymptotic coverage consistency of the prediction interval is obtained. We also develop simulation result to evaluate the finite sample performance of the procedure.

E0751: **Threshold factor models for high-dimensional time series**
Presenter: **Xialu Liu**, San Diego State University, United States
A threshold factor model is considered for high-dimensional time series in which the dynamics of the time series is assumed to switch between different regimes according to the value of a threshold variable. This is an extension of threshold modeling to a high-dimensional time series setting under a factor structure. Specifically, within each threshold regime, the time series is assumed to follow a factor model. The factor loading matrices are different in different regimes. The model can also be viewed as an extension of the traditional factor models for time series. It provides flexibility in dealing with situations that the underlying states may be changing over time, as often observed in economic time series and other applications. We develop the procedures for the estimation of the loading spaces, the number of factors and the threshold value, as well as the identification of the threshold variable. The theoretical properties are investigated. Simulated and real data examples are presented to illustrate the performance of the proposed method.

EO291 Room U414 NEW DEVELOPMENT IN TIME SERIES AND SPATIAL ECONOMETRICS Chair: Yongdeng Xu

E0624: **A varying coefficient panel data model with random individual effect and spatial errors**
Presenter: **Pipat Wongsa-art**, Cardiff University, United Kingdom
A varying coefficient panel data model with error components that are both spatially and time-wise correlated is considered. The model blends specifications typically considered in the spatial literature with those considered in the error components literature. We introduce the maximum likelihood estimators for estimating the spatial autoregressive parameter and the variance components of the disturbance process, which can be considered a generalized counterpart to the moments estimators suggested previously, within the non-varying coefficient context. We then use these estimators in the construction of the Nadaraya-Watson type estimation of the varying coefficient model. Nevertheless, how to conduct variable selection for the varying coefficient model in a computationally efficient manner is poorly understood. To solve the problem, we follow existing works, which combines the ideas of the local polynomial smoothing and the Least Absolute Shrinkage and Selection Operator, and presents new asymptotic results.

E0625: **A self-exciting threshold autoregressive count data model**
Presenter: **Namhyun Kim**, University of Exeter, United Kingdom
A new approach to modelling nonlinear time series of counts is proposed. In particular, the aims is to provide an instantaneous prediction of counts showing highly nonlinear phenomena such as limit cycles, jump resonance, harmonic distortion, modulation effects and chaos. Although the simple and intuitive benefits of using a Poisson process are well-known, the alternative, a Type II negative binomial (NB) process. The NB process provides a convenient way of introducing any type of serial dependence into counts by mixing a Poisson with a Gamma processes in which the dynamic evolution of counts is driven by the conditional mean. The conditional mean is specified with the well-known self-exciting threshold type of the process to describe the above nonlinearities in dynamic counts. The intrinsic presence of the latent stochastic conditional mean in the proposed likelihood motivates us to represent the proposed process with a legitimate linear state-space model by introducing a tuning parameter. Hence, the well-known linear filter is applied then the unknown parameters in the proposed process are estimated via quasi maximum likelihood estimation. The asymptotic properties of the proposed QMLEs are also studied and their finite sample performances are shown with Monte Carlo designs.

E0321: **DCC-HEAVY: A multivariate GARCH model with realized measures of variance and correlation**
Presenter: **Yongdeng Xu**, Cardiff University, United Kingdom
A new class of multivariate volatility models is proposed that utilising high-frequency data. We call this model the DCC-HEAVY model as key ingredients are the DCC model and the HEAVY model. We discuss the model's dynamics and highlight their differences from DCC-GARCH models. Specifically, the dynamics of conditional variances are driven by the lagged realized variances, while the dynamics of conditional correlations are driven by the lagged realized correlations in the DCC-HEAVY model. The new model removes well known asymptotic bias in DCC-GARCH model estimation and has more desirable asymptotic properties. We also derive a Quasi-maximum likelihood estimation and provide closed-form formulas for multi- step forecasts. Empirical results suggest that the DCC-HEAVY model outperforms the DCC-GARCH model in and out-of-sample.

EO177 Room U501 STATISTICAL METHODS FOR NETWORK DATA Chair: Yuan Zhang

E0207: **Logistic regression with network structure**
Presenter: **Guoyu Guan**, Northeast Normal University, China
As one of the most popular classification methods, logistic regression model has been extensively studied in the past literature. It basically assumes that individual's class label is influenced by a set of predictors. However, with the rapid advance of social network services, social network data are becoming increasingly available. As a result, how to take this additional network structure to improve classification accuracy becomes an important research problem. To this end, we propose a network based logistic regression model taking the network structure into consideration. Four interesting scenarios about link formation of the network structure are discussed under the NLR model. Furthermore, in order to figure out the impact of network structure on classification, asymptotic properties are derived for the prediction rule under different sparsities of network. Lastly, simulation studies are conducted to demonstrate the finite sample performance of the proposed method, and a real Sina Weibo dataset is analyzed for illustration purpose.

E0733: **Transform-based unsupervised point registration and unseeded low-rank graph matching**
Presenter: **Yuan Zhang**, Ohio State University, United States
Unsupervised estimation of the correspondence between two point sets has long been an attractive topic to CS and EE researchers. We focus on the vanilla form of the problem: matching two point sets that are identical over a linear transformation. We propose a novel method using Laplace transformation to directly match the underlying distributions of the two point sets. Our method provably achieves a decent error rate within polynomial time and does not require continuity conditions many previous methods rely on critically. Our method enables network comparison without strong model assumptions when node correspondence is unknown.

E0819: **Edge sampling using network local information**
Presenter: **Can Minh Le**, University of California, Davis, United States
Edge sampling is an important topic in network analysis. It provides a natural way to reduce network size while retaining desired features of the original network. Sampling methods that only use local information are common in practice as they do not require access to the entire network and can be parallelized easily. Despite promising empirical performance, most of these methods are derived from heuristic considerations and therefore still lack theoretical justification. To address this issue, we study a simple edge sampling scheme that uses network local information. We show that when local connectivity is sufficiently strong, the sampled network satisfies a strong spectral property. We quantify the strength of local connectivity by a global parameter and relate it to more common network statistics such as clustering coefficient and network curvature. Based on this result, we also derive a condition under which a hypergraph can be sampled and reduced to a weighted network.

EO035 Room U502 ADVANCES IN CAUSAL INFERENCE II Chair: Luke Keele

E0221: **Population-level cost-effectiveness analysis: The individual net benefit from a causal perspective**
Presenter: **Andrew Spieker**, Vanderbilt University Medical Center, United States
Co-authors: Jason Roy, Nandita Mitra
Health policy decisions regarding a particular treatment or intervention are generally made on the basis of aggregate information on both clinical effectiveness and cost. The net monetary benefit has been used as a measure to evaluate the comparative cost-effectiveness of two interventions. Briefly, this measure seeks to quantify the extent to which a treatment's level of efficacy justifies its associated cost, on average, at some willingness-to-pay threshold. Complexities of observational cost data including confounding, censoring, truncation by death, and time-dependent treatment can all impede causal estimation of cost-effectiveness. We present a marginal structural model framework for the net monetary benefit in order to overcome these challenges, thereby enabling causal interpretation and promoting refined cost-effectiveness comparison of clinically defined subgroups. We additionally present a novel visualization tool for cost-effectiveness that overcomes limitations of the current widely used cost-effectiveness acceptability curve.

E0308: **Measuring causal impacts on multifaceted outcomes with missingness: Learning welfare impacts of mobile credit**
Presenter: **Jacqueline Mauro**, UC Berkeley, United States
Co-authors: Blumenstock Joshua, Katherine Yen, Andrew Linxie
In the absence of a well-established financial sector, new tools for borrowing using mobile phones have arisen in East Africa. The welfare impacts of these high-interest, unsecured loans is not well known; ideally they provide a new avenue to credit to the underbanked, but they may also lead to debt spirals. In order to study the impacts on a multifaceted outcome like welfare, we extend nonparametric multivariate scaled ATE estimators to incorporate missing data, and develop a similar new LATE estimator. These estimators are doubly-robust and fully nonparametric, and allow us to robustly test for impacts of lending on multiple outcomes at once. We use data from a lender in the region to develop proxy measures of welfare which we analyze using these new tools. We compare our measures of welfare to self-reported measures of welfare by conducting interviews in Kenya.

E0827: **A Bayesian hierarchical model estimating CACE in meta-analysis of randomized clinical trials with noncompliance**
Presenter: **Haitao Chu**, University of Minnesota School of Public Health, United States
Co-authors: Jincheng Zhou, James Hodges, Fareed Suri
Noncompliance to assigned treatment is a common challenge in analysis and interpretation of randomized clinical trials. The complier average causal effect (CACE) approach provides a useful tool for addressing noncompliance, where CACE is defined as the average difference in potential outcomes for the response in a subpopulation of subjects who comply with their assigned treatments. We present a Bayesian hierarchical model to estimate the CACE in a meta-analysis of randomized clinical trials where compliance may be heterogeneous between studies. Between-study heterogeneity is taken into account with study-specific random effects. The results are illustrated by a re-analysis of a meta-analysis comparing epidural analgesia versus no or other analgesia in labor on the outcome of cesarean section, where noncompliance varied between studies. Finally, we present comprehensive simulations evaluating the performance of the proposed approach, and illustrate the importance of including appropriate random effects and the impact of over- and under-fitting.

EO095 Room U517 NONLINEARITY IN PANEL DATA ANALYSIS Chair: Feng Yao

E0259: **A fixed effect additive stochastic frontier model: A semiparametric estimation of inefficiency determinants**
Presenter: **Taining Wang**, West Virginia University, United States
Co-authors: Feng Yao
A semiparametric additive stochastic frontier model for panel data is proposed, where inputs and environment variables can enter the frontier individually and interactively through unknown smooth functions. The inefficiency has its mean function known up to certain parameters, and influenced by its determinants that may or may not appear on the frontier. We disentangle time invariant unobserved heterogeneities from inefficiency, which can be helpful to avoid overestimating the inefficiency level. Our model can be identified without the distribution assumption on the composite error, and consistently estimated without suffering from the curse of dimensionality and incidental parameter problems. Thus, our model can include a large number of interested variables as frontier or inefficiency determinants, a feature that can be potentially attractive to empirical studies. We illustrate the appealing finite-sample performance of the proposed estimator and two related hypotheses tests through the Monte Carlo study, and perform an application of world production frontier model with 116 countries during 2001-2013.

E0296: **On the influence of high leverage ratio on Chinese firms' performance: A semiparametric approach**
Presenter: **Jinjing Tian**, Dongbei University of Finance and Economics, China
Co-authors: Taining Wang, Feng Yao
Excessive leverage ratio of Chinese firms has raised concerns over its impact on productive efficiency. We employ a rm-level data set over 1998-2007 to investigate the role of debt in both the rm's production frontier and technical efficiency. The impact of debt on frontier is decomposed into a stand-alone neutral effect and indirect non-neutral effects, which alter the marginal product of production inputs. We estimate the effects through a semiparametric smooth coefficient stochastic frontier model. We allow a non-zero probability for the firms to be fully efficient, and model it as a function of debt and technical progress represented by time. We find that an increase in debt significantly reduces firms' frontier across different ownership, regions, and industries. Foreign and private firms are more efficient, with their full efficiency probability increased by debt and technical progress. In contrast,state-owned enterprises(SOEs) and collective firms are much less efficient and their probability of being fully efficient is not likely to increase with more debt. Furthermore, lower efficiency levels are concentrated in both the central and western regions, as well as in the mining and public utility industries.

E0541: **Efficient estimation in varying coefficient panel data model with different smoothing variables and fixed effects**
Presenter: **Feng Yao**, West Virginia University, United States

A varying coefficient panel data model is proposed to be estimated with different smoothing variables and fixed effects using a two step local linear regression approach. The pilot estimator removes fixed effect using kernel-based weight and estimates the varying coefficients by a marginal integration approach. We then use the pilot estimator to perform a one-step backfitting, which is shown to be efficient in the sense of being equivalent to a procedure knowing the other components of the varying coefficient. We obtain the asymptotic properties of both the pilot and efficient estimators. The Monte Carlo simulations show that our proposed estimator performs well. We illustrate their applicability by estimating a varying coefficient panel data production frontier, without assuming functional forms for distribution of efficiency and error terms.

EG010 Room U602 CONTRIBUTIONS IN ECONOMETRICS AND STATISTICS Chair: Alexandra Soberon

E0834: **The method of tail functions for confidence estimation**
Presenter: **Borek Puza**, ANU, Australia

The aim is to describe how tail functions can be used to construct confidence intervals with attractive properties. The theory is applied to several scenarios, including inference on the binomial proportion and normal mean. In each case, the standard interval is generalised using a class of tail functions and the optimal interval is then engineered, taking into account one or more criteria, such as the length of the interval, its maximum proportional length, or its expected length under a prior. The relationship between this approach, the Bayesian, and some others is also discussed.

E0836: **A comparative analysis of forecast models with online search volume: An application of price forecast on vegetables**
Presenter: **Yi-Wen Yeh**, National Taipei University, Taiwan

Along with the popularity of internet, people start to look for information by Internet. Search volumes gradually become the factor to affect vegetable prices. We take into account the information of search volume to enhance statistical modeling. Moreover, we compare the forecasting performances among time series model, the statistical modeling usually used in economy, and various analysis techniques with machine learning including neural networks and support vector regression. Furthermore, we attempt to build up a hybrid model which integrates time series with one of neural networks and support vector regression. An empirical analysis is conducted by using daily prices of cabbage. The results suggest an appropriate model for predicting cabbage prices.

E0838: **Statistical inference of non-Gaussian structural vector autoregressive (VAR) models**
Presenter: **Gigih Fitrianto**, Hiroshima University of Economics, Japan
Co-authors: Koichi Maekawa

Statistical inference in structural vector autoregressive (SVAR) models under non-Gaussian error is considered. Recent studies show that the non-Gaussian errors play an important role to identify a model and, they are useful to detect causal order of variables by using independent component analysis. This approach is gradually spreading in macro, micro, and financial econometrics. We conduct Monte Carlo experiments to see the performance of estimation methods and of testing long-run and short-run restrictions in a small SVAR model with t-distributed errors under acyclic and non-acyclic contemporaneous causal structure. The experiments show that performance of the test naturally depends on the degrees of freedom of t distribution and sample size. Furthermore, we apply this model to Japanese macroeconomic data to see whether Japans quantitative easing financial policy is effective or not. As a result, we found that SVAR model is potentially useful to detect causal order among macroeconomic variables.

Wednesday 26.06.2019 08:35 - 10:15 Parallel Session F – EcoSta2019

EO185 Room S101 RECENT ADVANCES IN PENALIZED LEARNING METHODS FOR COMPLEX DATA Chair: Jun Song

E0641: **Ensemble estimation and variable selection with semiparametric regression models**
Presenter: **Sunyoung Shin**, University of Texas at Dallas, United States
Co-authors: Yufeng Liu, Stephen Cole, Jason Fine
Scenarios are considered in which the likelihood function for a semiparametric regression model factors into separate components, with an efficient estimator of the regression parameter available for each component. An optimal weighted combination of the component estimators, named an ensemble estimator, may be employed as an overall estimate of the regression parameter, and may be fully efficient under uncorrelatedness conditions. This approach is useful when the full likelihood function is difficult to maximize but the components are easy to maximize. As a motivating example, we consider proportional hazards regression with prospective doubly-censored data, in which the likelihood factors into a current status data likelihood and a left-truncated right-censored data likelihood. Variable selection is important in such regression modelling but the applicability of existing techniques is unclear in the ensemble approach. We propose ensemble variable selection using the least squares approximation technique on the unpenalized ensemble estimator, followed by ensemble re-estimation under the selected model. The resulting estimator has the oracle property such that the set of nonzero parameters is successfully recovered and the semiparametric efficiency bound is achieved for this parameter set. Simulations show that the proposed method performs well relative to alternative approaches. Analysis of the multicenter AIDS cohort study illustrates the practical utility of the method.

E0705: **A note on ROC-optimizing support vector machines**
Presenter: **Seung Jun Shin**, Korea University, Korea, South
Unbalanced classification where one class dominates another is frequently-encountered in practice. Most classifiers that target to reduce misclassified examples may fail in such case. The ROCSVM directly optimize the AUC of ROC can be used as a natural alternative in the unbalanced classification, since AUC is a performance measure independent of threshold value that controls balance of two classes. We present some results about ROC SVM. First, we establish the piecewise linearity of the ROC-SVM solution and develop an efficient algorithm to recover entire trajectories of the solutions. Second, we develop the SCAD-penalized ROC-SVM to select informative variables in unbalanced classification.

E0708: **A computationally efficient algorithm for random effects selection in linear mixed models**
Presenter: **Mihye Ahn**, University of Nevada Reno, United States
The random effects selection has been received little attention in the literature. In linear mixed models, several methods for random effects selection have been proposed. However due to computationally intensive tasks, it is limited to apply the existing methods in practice. We propose two approximate methods of the moment-based method for random effects selection. The exact moment-based method has two challenging computation issues: nonlinear semidefinite programming and nonlinear programming with a linear inequality constraint. In particular, the most time-consuming step is the second computation to produce sparse solutions of the variance-covariance matrix of random effect factors. Since the objective function has up to fourth order terms and it makes the computation tedious, we suggest using a linear approximation to the penalized variance-covariance matrix. It reduces the objective function up to second order, and the quadratic programming can be easily implemented in some statistical software. By simulation studies, we show that the approximate methods also perform well and often outperform the exact method.

E0787: **Adaptive community detection via fused L1 penalty**
Presenter: **Yunjin Choi**, National University of Singapore, Singapore
Co-authors: Vincent Tan, zhaoqiang Liu
In recent years, community detection has been an active research area in various fields including machine learning and statistics. While a plethora of works has been published over the past few years, most of the existing methods depend on a predetermined number of communities. Given the situation, determining the proper number of communities is directly related to the performance of these methods. Currently, there does not exist a golden rule for choosing the ideal number, and people usually rely on their background knowledge of the domain to make their choices. To address this issue, we propose a community detection method that is equipped with data-adaptive methods of finding the number of the underlying communities. Central to our method is fused l-1 penalty applied on an induced graph from the given data. The proposed method shows promising results.

EO215 Room S102 NOVEL METHODS AND APPLICATIONS IN BIOSTATISTICS Chair: Haitao Chu

E0242: **Assessment of genetic impacts from twin study: A mixture distribution approach**
Presenter: **Zonghui Hu**, National Institutes of Health, United States
It is challenging to identify features that are genetically determined from those environmentally determined. We approach this by assessing the collective genetic impacts on a feature via the differential correlation in monozygotic twins versus dizygotic twins. Since the underlying order in a twin pair is mostly unclear, data are recorded in a random order, and conventional approaches for correlation coefficients are not valid. To handle the issue of missing order, we model twin data under the framework of mixture bivariate distribution. Taking full advantage of the properties of twin data, we construct and estimate under a combined likelihood function. Despite slow convergence associated with mixture distribution estimation, the combined likelihood induces improved convergency and allows effective statistical inference on the collective genetic impacts. The proposed method is applied to a twin study on immune traits.

E0253: **A Bayesian design for phase I cancer therapeutic vaccine trials**
Presenter: **Chenguang Wang**, Johns Hopkins University, United States
Co-authors: Gary Rosner, Richard Roden
Phase I clinical trials are the first step in drug development to test a new drug or drug combination on humans. Typical designs of Phase I trials use toxicity as the primary endpoint and aim to find the maximum tolerable dosage. However, these designs are poorly applicable for the development of cancer therapeutic vaccines because the expected safety concerns for these vaccines are not as much as cytotoxic agents. The primary objectives of a cancer therapeutic vaccine phase I trial thus often include determining whether the vaccine shows biologic activity and the minimum dose necessary to achieve a full immune or even clinical response. We propose a new Bayesian phase I trial design that allows simultaneous evaluation of safety and immunogenicity outcomes. We demonstrate the proposed clinical trial design by both a numeric study and a therapeutic human papillomavirus vaccine trial.

E0480: **Statistical modeling with imperfect data**
Presenter: **Li Tang**, St. Jude Children's Research Hospital, United States
Imperfect data is a long-standing statistical problem in practice. It often occurs due to method limitations or cost considerations. For example, in many real studies, either an exposure or a response variable or both may be misclassified, resulting in a common type of imperfect data. Another example of imperfect data involves study designs incorporating pooled samples. The third type of imperfect data one often encounters in reality is

missing information in variables. As such, potential threats to the validity of analytic results (e.g., estimates of effects) are widely known. Although much of the discussion has been made in literature, it is often restricted to oversimplified settings, which may not be satisfied in practice. Thus, clear illustrations of valid and accessible methods that deal with complicated settings are still in high demand. We propose novel frameworks that allow flexible modeling of common types of imperfect data and emphasize the utility when oversimplified assumptions are not met.

E0490: **Estimating organized large covariance matrix with l_0 penalty**
Presenter: **Shuo Chen**, University of Maryland, School of Medicine, United States
Interactions between features of high-dimensional biomedical data often exhibit complex and organized, yet latent, network topological structures. Estimating the organized large covariance matrix of these high-dimensional biomedical data while preserving and recognizing the latent network topology are challenging. A new l_0 shrinkage large covariance procedure is proposed that first detects latent network topological structures by implementing new penalized optimization and then regularizes the covariance matrix by leveraging the detected network topological information. The network topology guided regularization can reduce false positive and false negative rates simultaneously because it allows edges to borrow strengths from each other precisely. We provide applications to several large biomedical data examples including proteomics, genomics, and metabolomics data and demonstrate that organized latent network topological structures widely exist in high-dimensional biomedical data across platforms. In these applications, our methods show robust performance and scalability for identifying network structures. We also extend this method to detect network structures beyond the community structure.

EO203 Room S106 STATISTICAL METHODS IN COMPLEX DATA ANALYSIS Chair: ChienTong Lin

E0335: **On local Pearson correlation**
Presenter: **Li-Shan Huang**, National Tsing Hua University, Taiwan
The classical Pearson correlation measures the linear association between two variables. Its extension to measuring local association is widely used but little is known for its theoretical properties. We investigate the local Pearson correlation for both conditional and symmetric cases and establish properties using a kernel smoothing approach. Simulated examples confirm the asymptotic theory and a real data example is given for illustration.

E0497: **Nonlinear moderated mediation analysis with genetical genomics data**
Presenter: **Yuehua Cui**, Michigan State University, United States
Genetical genomics data provide promising opportunities for integrative analysis of gene expression and genotype data. Here we propose a nonlinear moderated mediation analysis taking into the causal mediation effect of gene expression on the relationship between genetic and phenotype, potentially moderated by environmental factors. Our goal is to select important genetic and gene expression variables that can predict a phenotypic response under a high-dimensional setup. As genes function in networks to fulfill their joint task, incorporating network or graph structures in a regression model can further improve gene selection performance. We propose a nonlinear graph-constrained penalized regression model to improve the selection performance via incorporating gene network structures. A two-step estimation procedure is adopted to obtain better variable selection and estimation. Simulation and real data analysis are conduced to show the utility of the method.

E0571: **Nonparametric confidence band for activity profiles based on wearable device data**
Presenter: **Hsin-wen Chang**, Academia Sinica, Taiwan
Co-authors: Ian McKeague
The motivation comes from applications to health care monitoring in which there is a need to analyze activity profiles constructed from wearable device data. We introduce a nonparametric likelihood ratio approach that makes efficient use of the activity profiles to provide a confidence band for their means. The procedure is calibrated using bootstrap resampling. A simulation study shows that the proposed procedure outperforms competing Wald-type functional data approaches. We illustrate the proposed methods using wearable device data from an NHANES study.

E0726: **Regularized variable selection for high dimensional survival data with unknown link function**
Presenter: **Haitao Zheng**, Southwest Jiaotong University, China
Aft model is one of the most commonly used models to handle survival data. However, the model assumption may not be correct in the real. We do not make any assumption on link function of the model. We use kernel estimation method to estimate the unknown link function. Then, we use a weighted least squares method with censoring constraints and sparse penalization to select high-dimensional covariates. In simulation studies, we compare the tpr, tnr, tdr and tndr of the proposed model with aft model and the neural network model from R package vsurf to illustrate the performance of each method.

EO233 Room S1A01 FUNCTIONAL AND HIGH DIMENSIONAL DATA WITH COMPLEX STRUCTURE Chair: Yehua Li

E0692: **Covariance function estimation for multidimensional functional data**
Presenter: **Raymond Ka Wai Wong**, Texas A&M University, United States
Co-authors: Xiaoke Zhang, Jiayi Wang
Multidimensional functional data are becoming more common in various domains such as climate studies, neuroimaging and chemometrics. We will present a nonparametric covariance function estimator for multidimensional functional data. It is based on an efficient spectral regularizations of an operator associated with a reproducing kernel Hilbert space. We will discuss the corresponding numerical and theoretical results.

E0815: **A functional mixed model for scalar on function regression with application to a functional MRI study**
Presenter: **Luo Xiao**, North Carolina State University, United States
Motivated by a functional magnetic resonance imaging (MRI) study, we propose a novel functional mixed model for scalar on function regression. The model extends the standard scalar on function regression for repeated outcomes by incorporating random subject-specific functional effects. Using functional principal component analysis, the new model can be reformulated as a mixed effects model and thus easily fit. A test is also proposed to assess the existence of the random subject-specific functional effects. We evaluate the performance of the model and test via a simulation study, as well as on data from the motivating fMRI study of thermal pain. The data application indicates significant subject-specific effects of the human brain hemodynamics related to pain and provides insights on how the effects might differ across subjects.

E0652: **An autocovariance-based framework for curve time series**
Presenter: **Cheng Chen**, London School of Economics, United Kingdom
Co-authors: Xinghao Qiao
It is commonly assumed in functional data analysis (FDA) that samples of each functional variable are independent realizations of an underlying stochastic process, and are observed over a grid of points contaminated by i.i.d. measurement errors. In practice, however, the temporal dependence across curve observations may exist and the parametric assumption on the error covariance structure could be unrealistic. We consider the model setting for serially dependent curve observations, when the contamination by errors is genuinely functional with a fully nonparametric covariance structure. The classical covariance-based methods in the FDA are not applicable here due to the contamination that can result in substantial estimation bias. We propose an autocovariance-based framework to address error-contaminated curve time series problems. Under the proposed

framework, we discuss several important problems in FDA, e.g. dimension reduction, functional linear regression, singular component analysis and high dimensional applications.

E0657: **Statistical analysis of longitudinal data on Riemannian manifolds**
Presenter: **Xiongtao Dai**, Iowa State University, United States
A manifold version of the principal analysis by conditional expectation (PACE) is proposed to represent sparsely observed longitudinal data that take values on a nonlinear Riemannian manifold. Typical examples of such manifold-valued data include longitudinal compositional data, as well as longitudinal shape trajectories located on a hypersphere. Compared to standard functional principal component analysis that is geared to Euclidean geometry, the proposed approach leads to improved trajectory recovery on nonlinear manifolds in simulations. As an illustration, we apply the proposed method on longitudinal emotional well-being data for unemployed workers. An R implementation of our method is available on GitHub.

EO269 Room AT242 RECENT ADVANCES IN INTERVAL CENSORED FAILURE TIME DATA Chair: Yang-Jin Kim

E0312: **Regression analysis of a semi-competing risks model when all transition times are interval-censored**
Presenter: **Jinheum Kim**, University of Suwon, Korea, South
In biomedical or clinical studies, semi-competing risks data in which one type of event may censor an other event, but not vice versa, are often encountered. We propose a multi-state model for analyzing these semi-competing risks data in the presence of interval censoring on both intermediate and terminal events. The proposed model can reflect diversities for which real data might frequently possess. We utilize the Cox proportional hazards model with a frailty effect to incorporate dependency between transitions of states. Weight allocations on sub-intervals of censored intervals are also used to construct the modified likelihood functions. Marginalization of the full likelihood is accomplished using adaptive importance sampling, and the optimal solution of the regression parameters is achieved through the iterative quasi-Newton algorithm. The proposed methodology is illustrated on several simulation studies and real data.

E0363: **Efficient estimation of varying coefficient linear transformation model for interval censored data**
Presenter: **Xingqiu Zhao**, The Hong Kong Polytechnic University Shenzhen Research Institute, China
A semiparametric varying-coefficient partially linear transformation model is proposed for the analysis of interval-censored failure time data and develop corresponding efficient estimation procedures. The asymptotic properties of the proposed estimators are established. Simulation studies are conducted to evaluate the proposed methods and compare them with the existing methods. Real data involving interval censoring are analyzed to illustrate the applications of the proposed methods.

E0547: **A general class of additive semiparametric models for recurrent event data**
Presenter: **Akim Adekpedjou**, Missouri University of Science and Technology, United States
Co-authors: Russell Stocker
Recurrent event data is a special case of multivariate lifetime data that is present in a large assortment of studies. Due to its pervasiveness, it is essential that appropriate models and inference procedures exist for its analysis. We propose a general class of additive semiparametric models for examining recurrent event data that uses an effective age process to take into account the impact of interventions applied to units after an event occurrence. The effect of covariates is additive instead of the common multiplicative assumption.We derive estimators of the regression parameter, baseline hazard function, and baseline survivor function. We also establish the asymptotic properties of the estimators using tools from empirical process theory. Simulation studies indicate that the asymptotic properties of the regression parameter closely approximate its finite sample properties. The analysis of a real data set consisting of lymphoma recurrence times provides a practical illustration of the class of models.

E0279: **Time dependent association for bivariate interval censored data**
Presenter: **Yang-Jin Kim**, Sookmyung Women University, Korea, South
The aim is to suggest a time dependent association measure for bivariate interval censored data. There are many statistical methods to consider the dependency between two failure time variables. Most approaches are global measures with shortage not reflecting a local dependency and based on the marginal survival function and joint survival function. However, the estimation of joint survival function under bivariate interval censored data seems to be difficult. Pseudo partial likelihood is extended to bivariate interval censored data. A two-stage procedure is proposed for the estimation. Simulation studies are conducted to assess the finite sample properties of the presented estimates. Real data from dairy cow udder infection time is analyzed for illustration.

EO358 Room AT335 EAC-ISBA SESSION: BAYESIAN THEORIES AND ALGORITHMS Chair: Xin Wang

E0349: **Weighted batch means estimators in Markov chain Monte Carlo**
Presenter: **James Flegal**, University of California - Riverside, United States
A family of weighted batch means variance estimators is proposed, which are computationally efficient and can be conveniently applied in practice. The focus is on Markov chain Monte Carlo simulations and estimation of the asymptotic covariance matrix in the Markov chain central limit theorem, where conditions ensuring strong consistency are provided. Finite sample performance is evaluated through auto-regressive, Bayesian spatial-temporal, and Bayesian logistic regression examples, where the new estimators show significant computational gains with a minor sacrifice in variance compared with existing methods.

E0526: **Hierarchical Bayesian model for unmixing remote sensing data**
Presenter: **Joon Jin Song**, Baylor University, United States
Co-authors: Jonathan Hobbs, Colin Lewis-Beck, Anirban Mondal, Xin Wang, Zhengyuan Zhu
Remote sensing satellites often have coarse spatial resolution and provide aggregate estimates of surface characteristics over wide footprints with heterogeneous ground cover. Disaggregating or un-mixing the observed signal is useful for identifying individual vegetation types and their unique signatures. We propose a hierarchical Bayesian model to deal with the un-mixing problem, which integrates information from multiple sources. Combining additional ground-based data sources with the satellite data allows us to identify the separate crop signals that make-up the original satellite measurements. The methodology is illustrated with data from the SMOS (Soil Moisture and Ocean Salinity) satellite mission.

E0454: **Geometric ergodicity of Polya-Gamma Gibbs sampler for Bayesian logistic regression with a flat prior**
Presenter: **Xin Wang**, Miami University, United States
The logistic regression model is the most popular model for analyzing binary data. In the absence of any prior information, an improper flat prior is often used for the regression coefficients in Bayesian logistic regression models. The resulting intractable posterior density can be explored by running the data augmentation (DA) algorithm. We establish that the Markov chain underlying the DA algorithm is geometrically ergodic. Proving this theoretical result is practically important as it ensures the existence of central limit theorems (CLTs) for sample averages under a finite second moment condition. The CLT in turn allows users of the DA algorithm to calculate standard errors for posterior estimates.

E0418: **A parametric approach to unmixing remote sensing crop growth signatures**
Presenter: **Zhengyuan Zhu**, Iowa State University, United States

Remote sensing data is often measured or reported over wide spatial footprints with heterogeneous ground cover. Different types of vegetation, however, have unique signatures that evolve throughout the growing season. Without additional information, signals corresponding to individual vegetation types are unidentifiable from satellite measurements. We propose a parametric mixture model to describe satellite data monitoring crop development in the US Corn Belt. The ground cover of each satellite footprint is primarily a mixture of corn and soybean. Using auxiliary data from multiple sources, we model the aggregate satellite signal, and identify the signatures of individual crop types, using nonlinear parametric curves. Estimation is performed using a Bayesian approach, and information from auxiliary data is incorporated into the prior distributions to identify distinct crop types. We demonstrate our parametric unmixing approach using data from the European Space Agency's Soil Moisture and Ocean Salinity satellite. Lastly, we compare our model estimates for the timing of key crop phenological stages to USDA ground-based estimates.

EO043 Room AT337 DIMENSION REDUCTION AND CLUSTERING Chair: Kei Hirose

E0381: **Model selection and estimation for latent variable models**
Presenter: **Emi Tanaka**, University of Sydney, Australia
Co-authors: Francis Hui

Latent variable models (LVMs), including the special case of factor analysis when the responses are conditionally normally distributed, are gaining traction in many scientific fields owing to both their statistical and computational advantages as a means of dimension reduction, and their attractive interpretation with the latent variables representing unmeasured predictors and factor loadings corresponding to the coefficients. Model selection for LVMs has an additional striking twist: as the name suggests, the latent variables are unobserved and have to be estimated from the data. Therefore, we need to select both the order and the structure of the factor loadings, where the former involves choosing the number of latent variables. We introduce a method for order selection in LVMs known as the Ordered FActor Lasso, which utilises penalised likelihood methods to encourage both element-wise and group-wise sparsity in the loadings. Specifically, we show how the OFAL penalty exploits both the grouped and hierarchical nature of the loadings, thus providing a natural approach to order selection, while also circumventing the issue of identifiability without the use of an arbitrary constraint and offering the potential for easier interpretability of the factor loadings. Additionally, we will discuss a computational algorithm for calculating the OFAL estimates based on a convenient reparameterisation of the penalty.

E0500: **A higher order local intrinsic dimension estimator by regression analysis**
Presenter: **Hideitsu Hino**, The Institute of Statistical Mathematics, Japan

Estimating intrinsic dimension of the observed dataset is an essential step prior to dimensionality reduction, manifold learning, and visualization. We propose a non-parametric method for estimating the intrinsic dimension of the observed data using the notion of local information dimension and generalized linear model with Poisson error structure. When fitting the power of distance from an inspection point, and the number of samples included inside a ball with radius equal to the distance, to a regression model, a goodness of fit is estimated. By using the maximum likelihood method, the intrinsic dimension around the inspection point is estimated. The method is shown to be comparative to conventional methods on both global and local intrinsic dimension estimation experiments.

E0549: **Visualization of heterogeneous class specific tendencies in categorical data**
Presenter: **Mariko Takagishi**, Osaka university, Japan
Co-authors: Michel van de Velden

In multiple correspondence analysis (MCA), both individuals and categories can be represented in a biplot that jointly depicts the relationships across categories or individuals. To enhance the interpretation of such a biplot, adding class information of individuals (e.g., gender, nationality) can be helpful. We consider interpreting class-specific tendencies in a biplot. By obtaining average points for each class, we can depict class-specific tendencies in the biplot. However, this approach only reveals tendencies of many individuals within a class. When a relatively small group in a class has a strong tendency towards a particular category not selected by the majority group in that class, this tendency would not be visible in the biplot. Such minority tendencies could still be interesting to consider, especially in order to characterize tendencies within classes. Therefore, we propose a new approach to find class-specific clusters, and depict them together with the category points. The resulting visualization allows us to identify different heterogeneous tendencies within classes in a single biplot, as well as the perceived relationships among classes.

E0576: **Cluster-based multiclass linear discriminant analysis**
Presenter: **Kei Hirose**, Kyushu University, Japan
Co-authors: Kanta Miura, Atori Koie

Multiclass linear discriminant analysis (LDA) is a well-known supervised learning based on a dimensionality reduction technique. In practice, there exists datasets which consist of a large number of classes. In many cases, some of the classes are easy to be classified, and some of them are difficult. In such a case, the LDA can lead to a large classification error, because the data in two similar classes are not appropriately projected onto a low-dimensional space. To handle this issue, we introduce a cluster-based LDA, in which the data in similar classes are categorized as one cluster and then the LDA are conducted to each cluster.

EO253 Room U301 ORDER RELATED STATISTICAL INFERENCE Chair: Johan Lim

E0311: **Continuity correction for RSS-structured cluster randomized designs with binary outcomes**
Presenter: **Soohyun Ahn**, Ajou University, Korea, South
Co-authors: Xinlei Wang, Johan Lim

Correction for continuity is commonly used to improve the inference procedures with binary data in which the interesting event rate is rare or sample size is small. A standard approach of bias reduction in logit estimation is to add a correction factor 0.5 to both event count and non-event count. The correction factor 0.5 is known as a factor rendering the estimation be unbiased up to the order of $O(K^{-1})$, where K samples are observed from simple random sampling (SRS). However, for more general designs beyond SRS, it no longer makes the bias in order of $O(K^{-1})$ be 0 in estimating the logit. We find the formula of the correction factor to make the bias be order of $O(K^{-2})$ for general designs. We then apply it to estimating the logit with the samples from cluster randomized design (CRD) with ranked set sampling (RSS), named as RSS-structured CRD (RSS-CRD). The RSS-structured CRD is a two-stage design which incorporates a cost-efficient ranked set sampling (RSS) into cluster randomized design (CRD) to have more efficient estimation on treatment effect. We propose two versions of the correction factors for the RSS-CRD. We numerically compare the proposed correction methods with the correction factor 0.5 in terms of the bias and mean squared error in estimating the treatment effect. Based on our results, recommendations and suggestions will be made to practitioners about when to use which correction factor.

E0350: **Post-stratified probability-proportional-to-size sampling from stratified populations**
Presenter: **Omer Ozturk**, The Ohio State University, United States

Statistical inference is developed based on post-stratified probability proportional-to-size (*pspps*) sample from a finite population. A *pspps* sample selects the sample units with selection probabilities proportional to their size and measures them for the characteristic of interest. For each measured

unit, the *pspps* sample further creates position information (rank) in a comparison set of size M. The sample is then post-stratified into ranking classes based on their position information in the comparison set. A *pspps* sample is expanded to stratified populations by selecting a *pspps* sample from each stratum population to form the stratified *pspps* sample. Using this stratified *pspps* sample we construct unbiased and Rao-Blackwell estimators for the mean of the stratified population. Different sample size allocation procedures for stratum sample sizes are investigated. The new sampling design is applied to apple production data to estimate the total apple production in Turkey.

E0474: **Proportion estimation in ranked set sampling in the presence of tie information**
Presenter: **Xinlei Wang**, Southern Methodist University, United States
Co-authors: Ehsan Zamanzade, Xinlei Wang
Ranked set sampling (RSS) is a statistical technique that uses auxiliary ranking information of unmeasured sample units in an attempt to select a more representative sample that provides better estimation of population parameters than simple random sampling (SRS). However, the use of RSS can be hampered by the fact that a complete ranking of units in each set must be specified when implementing RSS. Recently, to allow ties declared as needed, a modification of RSS has been proposed, which is to simply break ties at random so that a standard ranked set sample is obtained, and meanwhile record the tie structure for use in estimation. Under this RSS variation, several mean estimators were developed and their performance was compared via simulation, with a focus on continuous outcome variables. We extend that to binary outcomes and investigate three nonparametric and three likelihood-based proportion estimators (with/without utilizing tie information), among which four are directly extended from existing estimators and the other two are novel. Under different tie-generating mechanisms, we compare the performance of these estimators and draw conclusions based on both simulation and a data example of breast cancer prevalence. Suggestions are made about the choice of the proportion estimator in general.

E0602: **On constrained nonparametric maximum likelihood estimation of the survival functions for truncated survival data**
Presenter: **Hsun-Chih Kuo**, National Kaohsiung University of Science and Technology, Taiwan
It is known that the Nonparametric Maximum Likelihood Estimation (NPMLE) of the survival function with truncation will be underestimated if the truncation is ignored. We intend to study the constrained NPMLEs of survival functions with truncation under distributional ordering such as stochastic ordering. To start with, we will consider survival data with right censoring and left truncation under stochastic ordering. Monte Carlo simulations will be conducted to compare the performance of NPMLEs of survival functions with considering and without considering truncation in terms of either bias or variance.

EO147 Room U302 SEEMINGLY UNRELATED ECONOMETRIC PAPERS IN QUANTILE MODELS Chair: Jau-er Chen

E0317: **Unconditional quantile regression with endogenous regressors: A simulation study**
Presenter: **Yuya Katafuchi**, Research Institute for Humanity and Nature, Japan
Quantile regressions enable researchers to investigate the distributional effect of a target variable conditional on control variables, which complement the average effect identified by the ordinary least squares. Among them, the unconditional quantile regression (UQR) is becoming popular for empirical researchers who would like to measure the effect of a target variable on the unconditional quantile of the outcome distribution. In particular, to consistently estimate the causal effect of a variable of interest, the existing literature renders two approaches to deal with the endogeneity within the context of the UQR: the instrumental variable approach and the control function approach. There is, however, little research conducting a comparison study between these two approaches through simulation experiments. This issue is examined through Monte Carlo simulations. The cause of weak instruments with the UQR estimation is also investigated.

E0574: **Robust estimation of quantile treatment effects under unconfoundedness**
Presenter: **Yu-Chang Chen**, University of California, San Diego, United States
Quantiles are important inputs to several inequality measures, and inspecting quantile treatment effects (QTE) is a useful way to characterize effect heterogeneity. It has been previously established the identification of QTE under unconfoundedness and provided an inverse probability weighting (IPW) estimator. Although the IPW estimator is semiparametric efficient, several simulation studies suggest that IPW estimators are sensitive to model misspecification and vulnerable to lack of overlap. Consequently, the procedure fails to balance the treatment and control group, leading to substantial bias. To address this issue, we exploit the recent advance in balancing techniques to achieve a robust estimation of QTE. Specifically, we discuss the potential application of the entropy balancing methodology to the QTE estimation problem. Simulations show that the proposed method has substantially less bias compared to the standard IPW estimators. Although the theoretical property of the proposed estimator is still not fully known, recent literature has shown that entropy balancing method has a connection to penalized regression. Therefore, we suspect that the proposed method can have a link to penalized regressions.

E0581: **Estimation and inference for distribution functions and quantile functions in endogenous treatment effect models**
Presenter: **Tsung-Chih Lai**, Feng Chia University, Taiwan
Co-authors: Yu-Chin Hsu, Robert Lieli
Two-step nonparametric estimators are proposed for the distribution functions of potential outcomes among the group of compliers in an endogenous treatment effect model. Our estimator is monotonically increasing and bounded between zero and one. The monotonizing method we propose is an alternative to a previous one and it is easier to implement. We obtain the quantile function by inverting the estimated distribution function and show that both the distribution and the quantile estimators converge weakly to zero-mean Gaussian processes. For uniform inference, we propose a multiplier bootstrap procedure to approximate the limiting processes. Our methods generalize the pointwise results, and we also discuss the case for the treated compliers. Monte Carlo simulations and an application to the effect of fertility on family income distribution illustrate the usefulness of our results.

E0598: **Double/Debiased machine learning with gradient boosting for treatment effect**
Presenter: **Jui-Chung Yang**, National Tsing Hua University, Taiwan
Co-authors: Hui-Ching Chuang, Chung-Ming Kuan
A new double/debiased machine learning framework (DML) has been recently proposed for the estimation and inference of low-dimensional parameters in the presence of high-dimensional nuisance parameter. In DML, the low-dimensional parameters of interest are estimated using the Neyman orthogonal moment and the cross fitting technique, while the nuisance parameters are estimated by some machine learning algorithms with sufficient rates of convergence, namely, $o_p(n^{-1/4})$. We show that a) regression trees and random forests in general do not have the $o_p(n^{-1/4})$ rates and may result in serious bias and size distortion, and b) when unknown nuisance functions are additive, the gradient boosting with stumps provide consistent estimation and correct inference for the treatment effect. We apply our methods to recent debate of the treatment effect of the Big N auditors to the audit quality. Consistent to previous result, which use 3,000 designs of the propensity score matching, the method supports the existing of the Big N effect during 1988 to 2006 in U.S. DML-GB also identifies the non-linear associations of the firm characteristics and the audit quality.

EO133 Room U414 RECENT ADVANCES IN META-ANALYSIS AND DATA INTEGRATION Chair: Tiejun Tong

E0193: **Multiple-outcome network meta-analysis and personalized treatment ranking**
Presenter: **Yong Chen**, Univ. of Pennsylvania, United States

Network Meta-analysis (NMA), also known as multiple treatments meta- analysis (MTM), expands the scope of conventional pairwise meta-analysis by simultaneously synthesizing both direct comparisons of interventions within randomized clinical trials and indirect comparisons across trials. Network meta-analysis has been shown to have the advantage of providing broader, objective and inclusive view of available evidence for comparative effectiveness reviews. While network meta-analysis allows integrating information from clinical trials comparing different interventions, it also brings additional model assumptions and complexity. Furthermore, trials typically measure multiple outcomes, such as drug efficacy and safety. Simultaneously modeling all the outcomes allows borrowing information across outcomes since they are not only correlated but are also of interest for clinicians and patients. Therefore, considering the multivariate network meta-analysis is a necessity for aggregating all the existing evidence for comparative effectiveness reviews. We will introduce a simple but robust method for conducting multivariate NMA, as well as its implications to personalized treatment ranking.

E0188: **Estimating the sample mean and standard deviation from the sample size, median, mid-range, and/or mid-quartile range**
Presenter: **Tiejun Tong**, Hong Kong Baptist University, Hong Kong

Evidence-based medicine is attracting increasing attention to improve decision making in medical practice via integrating evidence from well designed and conducted clinical research. Meta-analysis is a statistical technique widely used in evidence-based medicine for analytically combining the findings from independent clinical trials to provide an overall estimation of a treatment effectiveness. The sample mean and standard deviation are two commonly used statistics in meta-analysis but some trials use the median, the minimum and maximum values, or sometimes the first and third quartiles to report the results. Thus, to pool results in a consistent format, researchers need to transform that information back to the sample mean and standard deviation. We will introduce our recent advances in the optimal estimation of the sample mean and standard deviation for meta-analysis from both theoretical and empirical perspectives. Specifically, we solve the problems by incorporating the sample size in a smoothly changing weight in the estimators to reach the optimal estimation. Our proposed estimators not only improve the existing ones significantly but also share the same virtue of the simplicity. The real data application indicates that our proposed estimators are capable to serve as rules of thumb and will be widely applied in evidence-based medicine.

E0199: **Divide and recombine approaches for fitting smoothing spline models with large datasets**
Presenter: **Yuedong Wang**, University of California - Santa Barbara, United States

Spline smoothing is a widely used nonparametric method that allows data to speak for themselves. Due to its complexity and flexibility, fitting smoothing spline models is usually computationally intensive which may become prohibitive with large datasets.To overcome memory and CPU limitations, we propose four divide and recombine (D&R) approaches for fitting cubic splines with large datasets. We consider two approaches to divide the data: random and sequential. For each approach of division, we consider two approaches to recombine. These D&R approaches are implemented in parallel without communication. Extensive simulations show that these D&R approaches are scalable and have comparable performance as the method that uses the whole data. The sequential D&R approaches are spatially adaptive which lead to better performance than the method that uses the whole data when the underlying function is spatially inhomogeneous.

E0219: **On the efficiency of network meta-analysis**
Presenter: **Lifeng Lin**, Florida State University, United States

Network meta-analysis (NMA) has become an increasingly used tool to compare multiple treatments simultaneously by synthesizing direct and indirect evidence in clinical research. However, the synthesized overall evidence is seldom compared with the direct evidence to validate the efficiency of an NMA. On the one hand, we propose three new measures (i.e., the effective number of studies, the effective sample size, and the effective precision) to preliminarily quantify overall evidence gained in NMAs at the pre-analysis stage. They permit evidence users to intuitively evaluate the benefit of performing NMAs, compared with pairwise meta-analyses based on only direct evidence. We use an illustrative example to demonstrate their derivations and interpretations. On the other hand, at the post-analysis stage, we use the recently proposed borrowing of strength (BoS) statistic to empirically evaluate the benefits by incorporating indirect evidence in 40 published NMAs. The BoS statistic quantifies the percentage reduction in the uncertainty of the effect estimate when adding indirect evidence to an NMA. We found that the incremental gain may reliably occur only when at least two head-to-head studies are available and treatments are well connected. Researchers should routinely report and compare the results from both network and pairwise meta-analyses.

EO287 Room U517 CONTEMPORARY INFERENCE ISSUES IN BIG DATA PROBLEMS Chair: Gourab Mukherjee

E0220: **Dynamic tracking and screening in massive datastreams**
Presenter: **Lilun Du**, HKUST, China

In the modern era, technological advances have led to the emergence of an increasing number of applications requiring analysis of large-scale datasreams, that are consisted of multiple indefinitely long and time evolving sequences. Consequently, it is often necessary to develop statistical methodologies that perform inferential tasks in an online manner, and can continuously revise the model to reflect the current status of the underlying process. In particular, we are interested in constructing a large scale dynamic tracking and screening (DTS) procedure capable of rapidly identifying irregular individual streams whose behavioral patterns deviate from that of the majority. By fully exploiting the sequential feature of datastreams, we first develop a robust estimation approach under a framework of varying coefficient model. The procedure naturally accommodates unequally-spaced design points and updates estimates as new data arrive without the need to store an ever increasing data history. A data driven choice of an optimal smoothing parameter is accordingly proposed. Then, we suggest a new model-specification test tailored to the streaming environment. The resulting DTS scheme is able to adapt time varying structures appropriately, track changes in the underlying models, and hence maintain high identification accuracy in detecting irregular individuals.

E0254: **Testing general linear hypotheses under a high-dimensional spiked model**
Presenter: **Debashis Paul**, University of California, Davis, United States

The problem of testing linear hypotheses associated with a high dimensional multivariate linear regression model is considered when the noise covariance has a spiked covariance structure. The classical test for this kind of hypotheses based on the likelihood ratio statistic suffers from substantial loss of power when the dimensionality of the observations is comparable to the sample size. To mitigate this problem, we propose a class of regularized test procedures that rely on a nonlinear shrinkage of the eigenvalues and eigen-projections of the sample noise covariance matrix, under the assumption that the population noise covariance matrix has a spiked covariance structure. We solve the problem of finding the optimal regularization parameter through a probabilistic formulation of the alternatives. We compare the performance of the proposed test with several tests proposed in the literature through numerical studies.

E0509: **On the construction of adaptive predictive densities for sparse count data**
Presenter: **Keisuke Yano**, The University of Tokyo, Japan
Co-authors: Ryoya Kaneko, Fumiyasu Komaki
Predictive densities under the Kullback-Leibler loss in high-dimensional sparse count data models are discussed. In particular, we consider Poisson sequence models under sparsity constraints. Sparsity in count data implies zero-inflation or quasi zero-inflation, that is, situations where there exists an excess of zeros or near-zero counts. We investigate the exact asymptotic minimax Kullback-Leibler risks in both sparse and quasi-sparse Poisson sequence models, providing a class of Bayes predictive densities that attain exact asymptotic minimaxity. We also discuss adaptation to an unknown sparsity. Our analysis also discuss the performance of the proposed Bayes predictive densities in settings where current observations are missing completely at random. We show the efficiency of the proposed Bayes predictive densities through both simulation studies and applications to real data.

E0542: **Model selection bias invalidates goodness of fit tests**
Presenter: **Joshua Loftus**, New York University, United States
Co-authors: Weichi Yao
Goodness of fit tests are studied in a variety of model selection settings and find that selection bias generally makes such tests conservative. Since selection methods choose the "best" model, a goodness of fit test will usually fail to reject, even when the incorrect model has been chosen. This is troubling, as it implies these tests in practice do not actually provide evidence in favor of the chosen model. We also explore methods of post selection inference to obtain conditionally unbiased goodness of fit tests and show how these outperform the unadjusted tests.

EC351 Room S104 CONTRIBUTIONS IN HIGH DIMENSIONAL AND COMPLEX DATA ANALYSIS Chair: Xialu Liu

E0265: **Interfirm relationship analysis for dynamic and dual-view company networks: A latent space modeling approach**
Presenter: **Ka Chung Ng**, The Hong Kong University of Science and Technology, Hong Kong
Co-authors: Mike So, Kar Yan Tam
Interfirm relationships are crucial to our understanding of firms' collective and interactive behavior. They have many business implications that help firms improve performance and governance. Toward this end, we propose a latent space approach to model temporal change in interfirm relationships, from a dynamic and dual-view company network, which is still under-researched in the literature. We assume that the probability of a link between firms depends only on a underlying latent space; firms that are close to each other in the latent space are more likely to develop a linkage. We expect to make four contributions: 1) three insightful business measures, business proximity, business reach and business attraction, are proposed to assist the analysis of interfirm relationship; 2) the latent space approach is extended to handle duel-view and dynamic networks in an integrated manner that is seldom addressed in the literature; 3) complex dynamic interfirm relationships are parsimoniously visualized and traced in a clear and interpretable way to help reduce cognitive load when making decision; 4) firms' relatedness and similarity are quantified, which can support managerial decision-making, portfolio management and private firm valuation.

E0752: **Bias corrected SVM with the Gaussian kernel in the HDLSS context**
Presenter: **Yugo Nakayama**, University of Tsukuba, Japan
Co-authors: Kazuyoshi Yata, Makoto Aoshima
A common feature of high-dimensional data is that the data dimension is high, however, the sample size is relatively low. We call such data HDLSS data. Asymptotic properties of the linear support vector machine (SVM) in the HDLSS context have been previously investigated. We study asymptotic properties of a non-linear SVM in HDLSS settings. We show that the non-linear SVM is heavily biased for imbalanced data. In order to overcome such inconvenience, we propose a bias-corrected SVM (BC-SVM) which is robust against extremely imbalanced data. In particular, we investigate asymptotic properties of the BC-SVM when having the Gaussian kernel. Since the performance of the BC-SVM is influenced by a scale parameter involved in the Gaussian kernel, we discuss a choice of the scale parameter in numerical simulations and actual data analyses.

E0806: **Shrinkage estimation of the mean of high-dimensional normal distribution**
Presenter: **Ryota Yuasa**, The University of Tokyo, Japan
Co-authors: Tatsuya Kubokawa
The problem of the estimation of the mean matrix of the multivariate normal distribution in high dimensional setting have been addressed. The Efron-Morris-type estimators with ridge-type inverse matrices are considered, and the proposal for estimation of optimal weights is estimator based on minimizing the risk function under the quadratic loss. Two approaches were considered, one is a method using random matrix theory, and the other is a method using Stein's identity. By considering these two methods, it can be seen that the estimators derived by the method using Stein's identity have optimality from the viewpoint of asymptotic minimization of the loss function when the prior distribution is assumed to be the true mean. It was also shown that under some assumptions, the proposed estimator has minimaxity. Numerical experiments are conducted to confirm the effect of the choice of ridge parameter on the estimation. The proposed estimators were compared with Efron-Morris estimator and James-Stein estimator. The proposed estimators provide better estimation accuracy than the Efron-Morris estimator, especially when both n and p are large, than the other estimators to be compared.

E0825: **Network analysis of Kaohsiung city mass transit**
Presenter: **Cheng-Han Chua**, National Sun Yat-sen University, Taiwan
The iPASS is a contactless smartcard operated by the iPASS Corporation in Taiwan. iPASS is accepted for public transport including Kaohsiung Metro, Taipei Metro, buses, bike and etc. The payment history of iPASS provides a rich source for investigating the major transport corridors of a city. We explore the iPass transport payment data and analyze the long term payment history of iPASS in Kaohsiung city. Time series models for daily traffic volume of metro, bus and bike are constructed. Intervention analysis of Winter Free mass transit policy is performed and investigated. Finally, we build the transport network model for the Kaohsiung city and use Rshiny to construct an interface for data exploration and network visualization.

EC343 Room U501 CONTRIBUTIONS IN COMPUTATIONAL ECONOMETRICS AND STATISTICS Chair: Eric Chi

E0756: **Classification strategies for time-constrained cost-sensitive decision tree induction with missing values**
Presenter: **Yong-Shiuan Lee**, National Chengchi University, Taiwan
Co-authors: Tsung-Chi Cheng

The induction of a cost-sensitive decision tree is one of extensively investigated issues in the study of classification. Among the studies, a newly developed algorithm generates a time-constrained minimal-cost tree, which is the first to build a cost-sensitive tree within a time limit. Their experiments show that the algorithm possesses highly satisfactory performance. However, there often exist missing values when analyzing real data. We therefore extend the time-constrained cost-sensitive tree induction to handle the missing values simultaneously, in which two methods are employed to deal with incomplete data. The first one is to apply the active feature acquisition (AFA) approach, and the other is the model-based imputation methods. Through AFA, we acquire the true feature values for those missing data at a cost that we have to take into account in the tree-inducting process. While imputing the missing values based on available data is a more statistical strategy, which may require little cost and time, but it leads to the issue of misclassification. The proposed strategies incorporate AFA and imputations with the time-constrained cost-sensitive tree induction for different scenarios. We conduct a simulation study and real-world data analysis to examine the performance of the proposed algorithm.

E0821: **Automatic algorithms for univariate time series forecasting using SARIMA and hybrid Wavelet-ARIMA-neural networks models**

Presenter: **Dedi Rosadi**, Universitas Gadjah Mada, Indonesia

In some application of time series modeling, it is necessary to obtain forecast of various types of data automatically and possibly in real-time, for instance, to do a (near) real-time processing of the satellite data. Various automatic algorithms for modeling univariate time series data are available in the literature. One class of the methods is the automatic algorithm to model and to forecast univariate SARIMA models. We discuss three methods of these types of algorithms, one of them based on a combination between the best exponential smoothing models to obtain the forecast, together with state-space approach of the underlying model to obtain the prediction interval. Other method, which is more advanced method, is based on X-13-ARIMA-SEATS. Other method use more heuristic approaches, namely the genetic algorithms. We also consider other general approaches and they are not restricted only to model the class of SARIMA models. One of the methods is based on automatic neural network method. The other (new) automatic algorithms are hybrid methods, which we called automatic Zhang (which uses a decomposition of the data into linear and nonlinear part), automatic KAV method (which combine decomposition of the data using wavelet and Zhang approach) and other type of automatic KAV method (which uses different architecture of the model). All of these approaches are implemented in our R-GUI package, namely RcmdrPlugin.SPSS. We provide empirical application using real data.

E0812: **Information criteria for gradient boosted trees: Adaptive tree size and early stopping**
Presenter: **Berent Aanund Stroemnes Lunde**, University of Stavanger, Norway
Co-authors: Tore Selland Kleppe, Hans Skaug

In gradient tree boosting, the functional form of the ensemble repeatedly changes during training. To select a sensible functional complexity for the boosting ensemble, the leading implementations offer a high number of hyperparameters for regularization, available for manual tuning. This tuning typically require a combination of computationally costly cross validation, coupled with some expert knowledge. To combat this, we propose an information criterion for gradient boosted trees, applicable to both the learning of the structure of trees, and as a stopping criterion for the boosting algorithm. The resulting algorithm is adaptive to the training data at hand; it is largely automatic and with little worries of overfitting. Moreover, the computations for the criterion require little additional computational overhead, and, as the algorithm only has to run once, the computational cost is drastically reduced in comparison to implementations with manual tuning.

E0811: **Dynamic density forecasting using machine learning**
Presenter: **Lubos Hanus**, UTIA AV CR, v.v.i, Czech Republic
Co-authors: Jozef Barunik

The use of machine learning techniques is proposed to describe and forecast the conditional probability distribution of asset returns. We redefine the problem of forecasting of conditional probabilities looking from a different perspective than traditional ordered binary choice models. Using deep learning methods, we offer a better description of asset returns distribution. The study on the most liquid U.S. stocks shows that predictive performance of machine learning methods is promising out-of-sample. We provide a comparison of machine learning methods to the unordered and order binary choice models used by the literature.

EC353 Room U502 CONTRIBUTIONS IN METHODOLOGICAL STATISTICS Chair: Ori Davidov

E0172: **Doubly robust estimation and causal inference for recurrent event data**
Presenter: **Chien-Lin Su**, McGill University, Canada
Co-authors: Russel Steele, Ian Shrier

Recurrent events are frequently observed in many biomedical longitudinal studies.The interest is to estimate the average causal effects for recurrent event data in the presence of confounders. We propose a doubly robust estimator which combines the weighted Nelson-Aalen estimator and regression estimator based on an assumed semiparametric multiplicative rate model for recurrent event data. The proposed estimators are shown to be consistent and asymptotically normal. In addition, a model diagnostic plot of residuals is presented to assess the adequacy of the semiparametric model. The finite sample behavior of the proposed estimators is evaluated through simulation studies. The proposed methodologies are illustrated via an injury database for circus artists.

E0257: **Discrete wavelet packet transform based whole genome sequences classification**
Presenter: **Hsin-Hsiung Huang**, University of Central Florida, United States

In recent years, alignment-free methods have been widely applied in comparing genome sequences, as these methods compute efficiently and provide desirable phylogenetic analysis results. These methods have been successfully combined with hierarchical clustering methods for finding phylogenetic trees. However, it may not be suitable to apply these alignment-free methods directly to existing statistical classification methods, because an appropriate statistical classification theory for integrating with the alignment-free representation methods is still lacking. We propose a discriminant analysis method which uses the discrete wavelet packet transform to classify whole genome sequences. The proposed alignment-free representation statistics of features follow a joint normal distribution asymptotically. The data analysis results indicate that the proposed method provides satisfactory classification results in real time.

E0218: **Non-parametric hypothesis testing with a nuisance parameter: A permutation test approach**
Presenter: **EunYi Chung**, University of Illinois at Urbana Champaign, United States

A classical problem in statistics is studied: testing goodness of fit in the presence of a nuisance parameter. The main contribution is a novel permutation test for this testing problem that is asymptotically valid under fairly weak assumptions, while still providing an exact error control

in finite samples under more restrictive conditions. In addition, the permutation test presented has finite- and large-sample properties comparable to those existing in the literature. The main result relies on the martingale transformation of the empirical process previously introduced. A noteworthy application of this testing problem is the one of testing for heterogeneous treatment effect in a randomized experiment. In this context, the null hypothesis implies that the distribution of the treatment and control groups are a constant shift apart. Moreover, the proposed method can be extended to testing the joint null hypothesis that treatment effects are constant within individual subgroups, while allowing for varying average treatment effects across subgroups. As a result, this test is able to detect treatment effect heterogeneity within individual subgroups even if the average treatment effects are identical across subgroups.

E0830: **Multi-resolution geographically weighted regression**

Presenter: **Yu-Ting Fan**, National Chiao Tung University, Taiwan

Geographically weighted regression (GWR) is concerned about regression for spatial data, where the regression coefficients are allowed to vary in space. The performance of GWR depends on some weighting matrices, which in some situations are difficult to determine. We propose a new approach to estimate the regression coefficients by representing them utilizing a class of multi-resolution spline basis functions. The proposed method not only provides a multi-resolution representation of the coefficient surfaces, but it also simplifies the estimation and inference problem. Some numerical examples are provided to demonstrate the effectiveness of the proposed method.

Wednesday 26.06.2019 10:45 - 12:25 Parallel Session G – EcoSta2019

EO091 Room UB99(B1) FLEXIBLE MODELING OF LATENT VARIABLES AND CENSORED DATA Chair: Victor Hugo Lachos Davila

E0297: **Estimation for partially linear censored regression models based on heavy tailed distributions**
Presenter: **Larissa Matos**, Campinas State University - UNICAMP, Brazil
Co-authors: Marcela Lemus, Victor Hugo Lachos Davila, Christian Galarza
In many studies, limited or censored data are collected. This occurs, in several practical situations, for reasons such as limitations of measuring instruments or due to experimental design. So, the responses can be either left, interval or right censored. On the other hand, partially linear models are considered as a flexible generalizations of linear regression models by including a nonparametric component of some covariates in the linear predictor. We discuss an estimation procedure in partially linear censored regression models with errors following a scale mixture of normal (SMN) distributions. This family of distributions contains a group of well-known heavy-tailed distributions that are often used for robust inference of symmetrical data, such as Student t, slash and contaminated normal, among others. A simple EM type algorithm for iteratively computing maximum penalized likelihood (MPL) estimates of the parameters is presented. We evaluate the finite sample performance of the algorithm and the asymptotic properties of the MPL estimates through empirical experiments. An application to a real dataset is presented to illustrate the effectiveness of the proposed methods.

E0822: **Bayesian analysis of cognitive diagnostic models for continuous response data**
Presenter: **Xiaojing Wang**, University of Connecticut, United States
Co-authors: Eduardo Schneider Bueno de Oliveira, Jorge Luis Bazan
The use of nondichotomous response models for assessment is increasing each time. Different scoring methods may be used to evaluate aspects of interest in many research fields, with the continuous responses being one of these. In the cognitive diagnosis models literature, much effort has been done to develop dichotomous and polytomous models in the past but, recently, the possibility of using continuous responses has been brought to discussion. However, there is no Bayesian approach for this class of models considering continuous responses. A first Bayesian framework for the continuous DINA is proposed. The good performance for parameter recovery is shown through a simulation study and also an application with continuous responses for risk perception is presented.

E0305: **Birnbaum-Saunders linear mixed-effects models with censored data: Bayesian MCMC implementation**
Presenter: **Filidor Vilca**, University of Campinas - Unicamp, Brazil, Brazil
The use of mixed effects models where responses are clustered around some random effects is usual in analysis of correlated data. The focus is on the Bayesian inference for Birnbaum-Saunders linear mixed models for censored data, which is inspired in previous work from a frequentist viewpoint. Specifically, the use of the Markov chain Monte Carlo method is explored to develop the Bayesian analysis, by using an acceleration convergence procedure. This approach provides an alternative to that developed under frequentist viewpoint that depends on the approximated likelihood function. Bayesian mechanisms for parameter estimation, residual analysis and influence diagnostics are developed. In order to examine the usefulness of this approach, we perform simulation studies taking in account the MCMC algorithm and its modified version. Also, an analysis to real dataset is considered to illustrate the proposed approach.

E0651: **Finite mixture modeling of censored data using the multivariate skew-normal distribution**
Presenter: **Victor Hugo Lachos Davila**, University of Connecticut, United States
Finite mixture models have been widely used for the modeling and analysis of data from a heterogeneous population. Moreover, data of this kind can be subject to some upper and or lower detection limits because of the restriction of experimental apparatus. Another complication arises when measures of each population depart significantly from normality, for instance, asymmetric data. For such data structures, we propose a robust model for censored data based on finite mixtures of multivariate skew-normal distributions. This approach allows us to model data with great flexibility, accommodating multimodality and skewness depending on the structure of the mixture components. We develop an analytically simple, yet efficient, EM-type algorithm for conducting maximum likelihood estimation of the parameters. The algorithm has closed-form expressions at the E-step that rely on formulas for the mean and variance of the multivariate truncated skew-normal distributions. Further, a general information-based method for approximating the asymptotic covariance matrix of the estimators is also presented. Results obtained from the analysis of both simulated and real datasets are reported to demonstrate the effectiveness of the proposed methodology.

EO117 Room S101 RECENT ADVANCES IN STATISTICAL LEARNING Chair: Lo-Bin Chang

E0410: **Robust mislabel logistic regression without modeling mislabel probabilities**
Presenter: **Su-Yun Huang**, Academia Sinica, Taiwan
Co-authors: Hung Hung, Zhi-Yu Jou
Logistic regression is among the most widely used statistical methods for linear discriminant analysis. In many applications, we only observe possibly mislabeled responses. Fitting a conventional logistic regression can then lead to biased estimation. One common resolution is to fit a mislabel logistic regression model, which takes into consideration of mislabeled responses. Another common method is to adopt a robust M-estimation by down-weighting suspected instances. We propose a new robust mislabel logistic regression based on gamma-divergence, which is also known as the density power divergence of type zero. The proposal possesses two advantageous features: (1) It does not need to model the mislabel probabilities. (2) The minimum gamma-divergence estimation leads to a weighted estimating equation without the need to include any bias correction term, i.e., it is automatically bias-corrected. These features make the proposed gamma-logistic regression more robust in model fitting and more intuitive for model interpretation through a simple weighting scheme. The method is also easy to implement, and two types of algorithms are included. Simulation studies and real data application will be presented.

E0416: **Robust linear discriminant analysis based on gamma-divergence**
Presenter: **Ting-Li Chen**, Academia Sinica, Taiwan
Fishers linear discriminant analysis (LDA), a traditional method in linear classification, is widely used in pattern recognition and dimension reduction. In many practical cases, some data points are mislabeled, which may badly affect the LDA results. We will first introduce gamma-divergence which is a more robust dispersion measure than the widely used Kullback-Leibler divergence. Based on minimum gamma-divergence, we propose a more robust LDA type method. Instead of sample mean and sample covariance derived by minimum K-L divergence, weighted sample mean and weighted sample covariance from gamma-divergence can successfully reduce the effects of incorrect labels. In the end, we will demonstrate the strength of our proposed method by simulation studies and applications on face data sets.

E0824: **Penalized orthogonal iteration for sparse estimation of generalized eigenvalue problems**
Presenter: **Yongho Jeon**, Yonsei University, Korea, South
A new algorithm is proposed for sparse estimation of eigenvectors in generalized eigenvalue problems (GEP). The GEP arises in a number of modern data-analytic situations and statistical methods, including principal component analysis (PCA), multiclass linear discriminant analysis

(LDA), canonical correlation analysis (CCA), sufficient dimension reduction (SDR) and invariant co-ordinate selection. We propose to modify the standard generalized orthogonal iteration with a sparsity-inducing penalty for the eigenvectors. To achieve this goal, we generalize the equation-solving step of orthogonal iteration to a penalized convex optimization problem. The resulting algorithm, called penalized orthogonal iteration, provides accurate estimation of the true eigenspace, when it is sparse. Also proposed is a computationally more efficient alternative, which works well for PCA and LDA problems. Numerical studies reveal that the proposed algorithms are competitive, and that our tuning procedure works well.

E0414: **Variable selection consistency in quantile regression by cross-validation**
Presenter: **Yoonsuh Jung**, Korea University, Korea, South
Co-authors: Sarang Lee
The problem of choosing the best predictive quantile regression model by cross-validation is considered. Although cross-validation is commonly used in quantile regression for model selection, its theoretical justification has not been verified yet. We prove that the cross-validation with check loss function can lead to variable-selection consistency in quantile regression. Specifically, we investigate the asymptotic properties of cross-validation in linear quantile regression model and its penalized versions under both the fixed and diverging number of parameters. One of the crucial requirements is that the sample size for model validation should be asymptotically equivalent to the total sample size, which is also required in the conditional mean regression.

EO157 Room S102 SPATIO-TEMPORAL MODELING OF INFECTIOUS DISEASE AND ONE HEALTH Chair: Andrew Lawson

E0179: **A geostatistical framework for combining spatially referenced disease prevalence data from multiple diagnostics**
Presenter: **Emanuele Giorgi**, Lancaster University, United Kingdom
Co-authors: Benjamin Amoah, Peter Diggle
Multiple diagnostic tests are often used due to limited resources or because they provide complementary information on the epidemiology of a disease under investigation. Existing statistical methods to combine prevalence data from multiple diagnostics ignore the potential over-dispersion induced by the spatial correlations in the data. To address this issue, we develop a geostatistical framework that allows for joint modelling of data from multiple diagnostics by considering two main classes of inferential problems: (1) to predict prevalence for a gold-standard diagnostic using low-cost and potentially biased alternative tests; (2) to carry out joint prediction of prevalence from multiple tests. We apply the proposed framework to two case studies: mapping Loa loa prevalence in Central and West Africa, using miscroscopy and a questionnaire-based test called RAPLOA; mapping Plasmodium falciparum malaria prevalence in the highlands of Western Kenya using polymerase chain reaction and a rapid diagnostic test. We also develop a Monte Carlo procedure based on the variogram in order to identify parsimonious geostatistical models that are compatible with the data. Our study highlights (i) the importance of accounting for diagnostic-specific residual spatial variation and (ii) the benefits accrued from joint geostatistical modelling so as to deliver more reliable and precise inferences on disease prevalence.

E0187: **Computational approaches to spatio-temporal surveillance of small area health data**
Presenter: **Andrew Lawson**, Medical University of South Carolina, United States
Some approaches to ST surveillance when MC sampling is used for posterior characterization are reviewed. With infectious disease as the main focus, we will consider the use of a variety of metrics, which are either residual based, or based on posterior functionals, such as SCPO, SKL, and surveillance residuals. We propose using new combinations of these metrics which include partial prediction using two stage models. We also explore the use of directional resultants in attempting to predict the spatial pathways for future infection in an infectious disease context. If time permits, we will also contrast the use of sequential MC as compared to conventional McMC when metrics are used.

E0307: **Parameterization via emulation: Spatial models of infectious disease transmission**
Presenter: **Rob Deardon**, University of Calgary, Canada
Statistical inference for spatial models of infectious disease spread is often very computationally expensive. Such models are generally fitted in a Bayesian Markov chain Monte Carlo (MCMC) framework, which requires multiple calculation of what is often a computationally cumbersome likelihood function. This problem is especially severe when there are large numbers of latent variables to compute. Here, we propose two methods of inference based on so-called emulation techniques. One method consists of approximating the likelihood via a Gaussian process built using "ABC-style" summary statistics. The second method consists of approximating the likelihood directly with the Gaussian process, but using pseudo-marginal approximations to allow for latent variables such as infectious periods. These methods are set in a Bayesian MCMC context, but avoid calculation of the computationally expensive likelihood function by replacing it via the aforementioned Gaussian processes. We show that such methods can be used to infer the model parameters and underlying characteristics of spatial disease systems, and that this can be done in much more computationally efficient manner than full Bayesian MCMC allows.

E0438: **Real time decision making for infectious disease outbreaks**
Presenter: **Michael Tildesley**, University of Warwick, United Kingdom
In the event of outbreaks of infectious diseases, mathematical models can be used to inform decision makers regarding the likely spread of disease and the impact of control strategies. However, in the early stages of novel outbreaks, there can often be significant uncertainty regarding the spatiotemporal spread of disease and the likely impact of any intervention policy. However, policy makers do not often have the luxury to wait for any uncertainty to resolve before introducing an intervention, so it is crucial that models are developed to take account of the most up to date information available. We analyse historical outbreaks of foot-and-mouth disease and avian influenza and investigate the predictability of infectious disease models in the early stages of these epidemics to determine how any predictions change as more data are accrued. Our results indicate that the substantial epidemiological uncertainty at epidemic onset can lead to misleading forecasts of the impact of any intervention policy. However, robust predictions can be obtained after the first two to three weeks owing to a resolution of uncertainty during this period. We conclude that real time information is vital to ensure that policy makers can select the most appropriate intervention policy to minimise the impact of any ongoing epidemic.

EO245 Room S104 RECENT DEVELOPMENTS IN HIGH DIMENSIONAL DATA Chair: Debashis Paul

E0375: **Sparse learning and structure identification for ultra-high-dimensional image-on-scalar regression**
Presenter: **Xinyi Li**, Statistical and Applied Mathematical Sciences Institute (SAMSI), United States
Co-authors: Lily Wang, Dan Nettleton, Huixia Judy Wang

High-dimensional image-on-scalar regression is proposed, where the spatial heterogeneity of covariate effects on imaging responses is investigated via a flexible partially linear spatially varying coefficient model. To tackle the challenges of spatial smoothing over the imaging response's complex domain consisting of regions of interest, the proposed method approximates the spatially varying coefficient functions via bivariate spline functions over triangulation. The aim is to first study estimation when the active constant coefficients and varying coefficient functions are known in advance, then a unified approach is developed for simultaneous sparse learning (i.e., variable selection) and model structure identification (i.e., determination of spatially varying vs. constant coefficients) in the presence of ultra-high-dimensional covariates. The proposed method can identify zero, nonzero constant and spatially varying components correctly and efficiently. The estimators of constant coefficients and varying coefficient functions are consistent and asymptotically normal. The method is evaluated by Monte Carlo simulation studies and applied to a dataset provided by the Alzheimer's Disease Neuroimaging Initiative (ADNI).

E0376: **Similarity scores for mixed types of data**
Presenter: **Keying Ye**, University of Texas at San Antonio, United States

In comparing similarity score or distance between two objects with many variables, studies in distance have been done for continuous, categorical and ordinal variables. A robust method of developing similarity or distance scores between different residential home properties with mixed types of data will be discussed. Such kind of scores can be used to find similar or dissimilar properties for comparisons, and more importantly, for selecting comps properties in home property appraisals.

E0244: **Parameterized matrix factorization with missing data via nonconvex optimization**
Presenter: **Xiaodong Li**, UC Davis, United States

Matrix factorization is one of the most foundational unsupervised learning techniques due to its power in exploration of informative structures in contemporary big data sets. We are particularly interested in parameterized matrix factorization with possible missing entries. Applications include reduced rank regression with possible missing entries in the responses, collaborative filtering with side information, pairwise ranking with aggregated comparison scores, to name just a few. We aim to establish a unified methodological and theoretical framework for nonconvex optimization based parameterized matrix factorization. In particular, we aim to implement a unified local minimum analysis in understanding the adaptivity of the method to the parameterized low-rank structures. For example, in the noiseless case we aim to study how to establish the required sampling complexity for exact recovery by the intrinsic dimension of the parameters; in the noisy case we are also interested in figuring out how the local-minimum based approximation depends on both the sample complexity and the intrinsic dimension.

E0613: **Minimax optimality of sparse predictive density estimates**
Presenter: **Gourab Mukherjee**, University of Southern California, United States

The problem of predictive density estimation is considered in a high-dimensional Gaussian model with sparsity constraints on the location parameters. The maximal risk of several spike-and-slab predictive density estimates is evaluated. Asymptotic minimaxity of these density estimates and the geometry of the least favorable prior distribution is discussed highlighting the contrasts with the minimax theory for point estimation of a multivariate normal mean under quadratic loss.

EO169 Room S106 STATISTICS FOR UNCONVENTIONAL, COMPLEX AND CHALLENGING DATASETS Chair: Anuradha Roy

E0278: **Asymptotic distribution of correlation matrix under blocked compound symmetric covariance structure**
Presenter: **Shinichi Tsukada**, Meisei University, Japan

The covariance matrix can have various specific structures in multivariate statistical analysis. One of these is the blocked compound symmetric (BCS) covariance structure. The BCS covariance structure for doubly multivariate observations is a multivariate generalization of the compound symmetric covariance structure for multivariate observations. The estimation and the hypothesis testing for the covariance matrix under the BCS covariance structure have been previously studied. Further analysis is to examine the correlation matrix. We investigate the asymptotic distribution of the correlation matrix and apply the result to the hypothesis testing.

E0429: **Two-sample correlation parameter testings in models with a Kronecker product covariance structure**
Presenter: **Chengcheng Hao**, Shanghai University of International Business and Economics, China
Co-authors: Yuli Liang

Under a model having a Kronecker product covariance structure with compound symmetry, two-sample hypothesis testing for a correlation is investigated. Several tests are suggested and practical recommendations are made based on their type I error probabilities and powers.

E0452: **Likelihood based analysis of longitudinal discrete data with overdispersion**
Presenter: **Justine Shults**, Perelman School of Medicine at University of Pennsylvania and Children's Hospital of Philadelphia, United States

The focus is on longitudinal discrete data that may be unequally spaced in time and may exhibit overdispersion, so that the variance of the outcome variable is inflated relative to its assumed distribution. We propose an approach that extends generalized linear models for analysis of longitudinal data and is likelihood based, in contrast to generalized estimating equations (GEE) that are semi-parametric. We demonstrate application of the method in an analysis of seizure counts and kidney transplant centers. Simulations for both studies show that the likelihood based approach outperforms GEE, especially as the degree of intra-subject correlation and over-dispersion in the outcome distribution increases.

E0728: **Mixture of multivariate t linear mixed models with missing information**
Presenter: **Tzy-Chy Lin**, Center for Drug Evaluation, Taiwan

Linear mixed-effects (LME) models have been widely used for longitudinal data analysis as they can account for both fixed and random effects, while simultaneously incorporating the variation on both within and between subjects. In clinical trials, some drugs may be more effective in Westerners than the Orientals. In this situation, such heterogeneity can be modeled by a finite mixture of LME models. The classical modeling approach for random effects and the errors parts are assumed to follow the normal distribution. However, the normal distribution is sensitive to outliers and intolerance of outliers may greatly affect the model estimation and inference. We propose a robust approach called the mixture of multivariate t LME models with missing information. To facilitate the computation and simplify the theoretical derivation, two auxiliary permutation matrices are incorporated into the model for the determination of observed and missing components of each observation. We describe a flexible hierarchical representation of the considered model and develop an efficient Expectation-Conditional Maximization Either (ECME) algorithm for carrying out maximum likelihood estimation. Simulation results and real data analysis are provided to illustrate the performance of the proposed methodology.

EO283 Room S1A01 NEW DEVELOPMENT IN FUNCTIONAL DATA ANALYSIS Chair: Adam Kashlak

E0176: **Sparse estimation of historical functional linear models with a nested group bridge approach**
Presenter: **Xiaolei Xun**, Fudan University, China
Co-authors: Jiguo Cao
The conventional historical functional linear model relates the current value of the functional response at time t to all past values of the functional covariate up to time t. Motivated by situations where it is more reasonable to assume that only recent, instead of all, past values of the functional covariate have an impact on the functional response, we investigate in this work the historical functional linear model with an unknown forward time lag into the history. Besides the common goal of estimating the bivariate regression coefficient function, we also aim to identify the historical time lag from the data, which is important in many applications. Tailored for this purpose, we propose an estimation procedure adopting the finite element method to conform naturally to the trapezoidal domain of the bivariate coefficient function. A nested group bridge penalty is developed to provide simultaneous estimation of the bivariate coefficient function and the historical lag. The method is demonstrated in a real data example investigating the effect of muscle activation recorded via the noninvasive electromyography (EMG) method on lip acceleration during speech production. The finite sample performance of our proposed method is examined via simulation studies in comparison with the conventional method.

E0502: **Symmetrisation for nonparametric functional ANOVA**
Presenter: **Adam Kashlak**, University of Alberta, Canada
Testing for equality of means and covariances among functional data groups has received a lot of attention from both parametric approaches via Gaussian processes and nonparametric ones reliant on permutation tests. We advance nonparametric testing by devising an exact test via a type of Khintchine inequality, a symmetrisation result for random variables in Banach spaces. This approach combines the computational speed of parametric methods with the distribution free benefits of permutation tests. The methodology is very general and can be extended to other data comparisons across categories.

E0775: **An approach for curve alignment and clustering**
Presenter: **Tzee-Ming Huang**, National Chengchi Univerisity, Taiwan
Co-authors: Yu-Hsiang Cheng, Su-Fen Yang
The focus is on the problem where we have several groups of curves, where curves in the same group have the same shape (up to a scale and location change) after being aligned. An approach will be presented that performs alignment and clustering simultaneously. The approach can handle misalignment due to nonlinear time transformation. Simulation results will be presented to demonstrate the performance of the proposed approach.

E0742: **Hierarchical ANOVA and F-tests for functional data with an application to highway traffic data in Taiwan**
Presenter: **Wei-Hsueh Huang**, National Tsing-Hua University, Taiwan
Co-authors: Li-Shan Huang
Some new tests for the one-way ANOVA problem for functional data are proposed. Based on local polynomial regression, exact local and global ANOVA expressions are obtained, forming hierarchical ANOVA structures for estimating individual curves, a mean curve within a group of functions, and an overall mean curve for all groups of functions. A local test taking account of functional structure is first developed to examining differences of functions in a neighborhood of a point. Combining local ANOVA quantities, a test statistic is formed to testing global functional ANOVA. We show that both the local and global test statistics have asymptotic F-distributions, and enjoy the "Wilks phenomenon". Simulation studies are presented to compare the proposed global ANOVA test with some tests in the literature. A real data example on traffic flows of Taiwan highways before and after a major earthquake in 2016 illustrates the proposed tests.

EO143 Room AT241 RISK ASSESSMENT ON NETWORK AND COMPLEX SYSTEMS Chair: Gaofeng Da

E0540: **Malicious data breach: Modeling and ratemaking**
Presenter: **Maochao Xu**, Illinois State University, United States
The data breach risk has become the most common and dangerous cyber risk nowadays, which can lead to the expose of sensitive and confidential information, and financial losses. Therefore, the study of data breach has attracted much attention in the literature, but mainly from the perspectives of information technology. We discuss a novel statistical model for modeling the malicious data breach, which can be further used for the ratemaking. In particular, a frequency-severity model will be developed for modeling the loss of data breach. Several interesting findings will be mentioned as well.

E0407: **A stochastic model of cyber attacks with imperfect detection**
Presenter: **Rui Fang**, Shantou University, China
A cyber security model with imperfect detection is introduced, in which one attacker launches multiple attacks against the target with adjusted strength based on the previous attacking outcome. Several sufficient conditions leading to the usual stochastic order on the first time to observe a truly compromised target, to observe a successful attack and to compromise the target are developed, respectively. The probability for the target to be truly compromised before observing some number of successful attacks is proved to increase (decrease) in the attacking (defense) strength. Monte Carlo simulations are also conducted to empirically illustrate the theoretical results.

E0356: **On total capacity of k-out-of-n systems with random weights**
Presenter: **Yiying Zhang**, Nankai University, China
Co-authors: Weiyong Ding, Peng Zhao
In engineering applications, many reliability systems can be modeled as k-out-of-n systems with components having random weights. Before putting such kind of system into a working state, it is of great significance for a system designer to find out the optimal assembly of the random weights to the components. We investigate the performance levels of k-out-of-n systems with random weights. Optimal assembly policies are obtained by maximizing the total capacity according to different criteria, including the usual stochastic order, the increasing convex [concave] order, and the expectation order. Based on the optimal assembly strategy derived by maximizing the expected total capacity, we further investigate stochastic properties of the resulting weighted system with respect to the vector of expectations of random weights. Numerical examples are provided to highlight our theoretical findings as well.

E0655: **On the assessment of systematic risk in networked systems with multiple types of nodes**
Presenter: **Gaofeng Da**, Nanjing University of Aeronautics and Astronautics, China
In a networked system, the systematic risks from an attack is a consequence of the network structure formed by connected nodes (e.g., individuals, businesses and computer systems). The most of previous modelling research on assessment of such risks only involves networks with one type of nodes (homogeneous case). We present a more general study on assessment of joint risks of networked systems with multiple types of nodes

(heterogeneous case). Under the L-Hop risk propagation model, we investigate the joint distribution for the numbers of compromised nodes with different types. An explicit formula of the joint distribution for a network with arbitrary structure is derived by using a reliability approach. In particular, some simplified formulas for special network topologies such as the complete, ER and complete bi-partite networks are provided. In addition, some useful observations on parameter effects are obtained by conducting a simulation study. Finally, the application in cyber insurance are discussed as well.

EO271 Room AT242 RECENT DEVELOPMENTS IN NON-PARAMETRIC METHODS Chair: Lan Xue

E0180: **Nonparametric kernel estimation of unrestricted distributions**
Presenter: **Carlos Martins-Filho**, University of Colorado at Boulder, United States
Co-authors: Kairat Mynbayev
Nonparametric estimation of an unrestricted distribution F which may, or may not, be absolutely continuous, is considered. First, for a point of continuity of F, x, we consider estimators that can be expressed as $\hat{F}_n(x) = (1/n)\sum_{i=1}^n U((X_i - x)/h)$, for a suitable choice of U and a bandwidth $h > 0$. We obtain the rates of convergence of these estimators to $F(x)$. Contrary to the extant literature, we make no restriction on the existence or smoothness of the derivatives of F. The key insight for the result is the use of Lebesgue-Stieltjes integrals. A special case of $\hat{F}_n(x)$, that reproduces the traditional kernel estimator, is given when $U(x) = \int_x^\infty K(u)du$ and K is a kernel. Second, for x that is either a point where F has a jump discontinuity, or isolated, we obtain rates of convergence for an estimator $\hat{J}(x) = (1/n)\sum_{i=1}^n W((X_i - x)/h)$ for the jump $F(x) - F(x-)$ and suitable choice of W. Once again, no restriction is imposed on F beyond right-continuity. A suitable choice is $W(x) = \int_{\mathbb{R}} e^{ixu}K(u)du$, the Fourier transform of a kernel K. The results are of significant practical use, as there are numerous examples of distributions that have mass points and singularities in Economics, Finance and Biomedicine.

E0262: **Kolmogorov-Smirnov simultaneous confidence bands for time series distribution function**
Presenter: **Jie Li**, Center for Statistical Science and Department of Industrial Engineering, Tsinghua University, China
Co-authors: Jiangyan Wang, Lijian Yang
Claims about distribution functions of time series are more often folklores than substantiated conclusions, due to lack of hypotheses testing tools. Kolmogorov-Smirnov type simultaneous confidence bands (SCBs) are constructed based on a simple random sample (SRS) drawn from a realization of time series, together with smooth SCBs using kernel distribution estimator (KDE). All SCBs are shown to enjoy the same limiting distribution as the standard Kolmogorov-Smirnov SCB for i.i.d. sample. This theoretical fact has been validated in simulation experiments performed on various time series. Hypotheses testing based on these SCBs has led to the unexpected finding that with proper rescaling, Gaussian distribution and most student's t-distributions are all acceptable alternatives of the S& P 500 daily returns' stationary distribution. This discovery challenges the long held belief that daily financial returns' distribution is fat-tailed and leptokurtic.

E0269: **Spline confidence bands for generalized regression models**
Presenter: **Jing Wang**, University of Illinois at Chicago, United States
A computational study of bootstrap confidence bands based on a free-knot spline regression is explored for the generalized linear models. In free-knot spline regression, the knot locations as additional parameters offers greater flexibility and the potential to better account for rapid shifts in slope and other important structures in the target function. However, the search for optimal solutions becomes very complicated because of freeing up the knots. In particular, the lethargy property in the objective function results in many local optima with replicate knot solutions. To prevent solutions with identical knots, a penalized Quasi-likelihood estimating equation is proposed that relies on both a Jupp transformation of knot locations and an added penalty on solutions with small minimal distances between knots. Focusing on logistic regression for binary outcome data, a parametric bootstrap is used to study the variability of the proposed estimator and to construct confidence bands for the unknown form of the logistic regression link function.

E0272: **Two-step estimation for time varying ARCH models**
Presenter: **Yuanyuan Zhang**, Tsinghua University, China
Co-authors: Rong Liu, Qin Shao, Lijian Yang
A time varying autoregressive conditional heteroskedasticity (ARCH) model is proposed to describe the changing volatility of a financial return series over long time horizon, along with two-step least squares and maximum likelihood estimation procedures. After preliminary estimation of the time-varying trend in volatility scale, approximations to the latent stationary ARCH series are obtained, which are used to compute the least squares estimator (LSE) and maximum likelihood estimator (MLE) of the ARCH coefficients. Under elementary and mild assumptions, oracle efficiency of the two-step LSE for ARCH coefficients is established, i.e., the two-step LSE is asymptotically as efficient as the infeasible LSE based on the unobserved ARCH series. As a matter of fact, the two-step LSE deviates from the infeasible LSE by $o_p(n^{-1/2})$. The two-step MLE,however, does not enjoy such efficiency, but $n^{1/2}$ asymptotic normality is established for both the two-step MLE as well as its deviation from the infeasible MLE. Simulation studies corroborate the asymptotic theory, and application to the S&P 500 index daily returns from 1950 to 2018 indicates significant change in volatility scale over time.

EO295 Room AT335 EAC-ISBA SESSION: HIGH DIMENSIONAL BAYESIAN METHODS IN DATA SCIENCE Chair: Wanjie Wang

E0283: **Oracle type posterior contraction rates in Bayesian inverse problems**
Presenter: **Shuai Lu**, Fudan University, Shanghai, China, China
Bayesian inverse problems in Hilbert spaces are discussed. The focus is on a fast concentration of the posterior probability around the unknown true solution as expressed in the concept of posterior contraction rates. This concentration is dominated by a parameter which controls the variance of the prior distribution. Previous results determine posterior contraction rates based on known solution smoothness. We show that an oracle type parameter choice is possible. This is done by relating the posterior contraction rate to the root mean squared estimation error. In addition we show that the tail probability, which usually is bounded by using the Chebyshev inequality, has exponential decay, at least for a priori parameter choices. These results implement the exponential concentration of Gaussian measures in Hilbert spaces.

E0323: **High-dimensional scaling limit of Monte Carlo methods**
Presenter: **Kengo Kamatani**, Osaka University, Japan
The focus will be on the dimensionality effect of some Monte Carlo methods. High-dimensional asymptotic analysis is becoming popular. Some recent results on the random-walk Metropolis algorithm, the MpCN algorithm, and piecewise deterministic Markov processes will be discussed.

E0505: **Robust estimation of causal effects under data combination using instrumental variables**
Presenter: **BaoLuo Sun**, National University of Singapore, Singapore
Although instrumental variable (IV) methods are widely used to estimate causal effects in the presence of unmeasured confounding, the IVs, exposure and outcome are often not measured in the same sample due to complex data harvesting procedures. Following previous influential

articles, numerous empirical researchers have applied two-sample IV methods to perform joint estimation based on an IV-exposure sample and an IV-outcome sample. We develop a general semi-parametric framework for two-sample data combination models and propose new multiply robust locally efficient estimators of the causal effect of exposure on the outcome, and illustrate the methods through simulation and an econometric application on public housing projects.

E0524: **Ensemble Kalman inversion**
Presenter: **Xin Tong**, National University of Singapore, Singapore
Ensemble Kalman inversion is a parallelizable methodology for solving inverse or parameter estimation problems. Although it is based on ideas from Kalman filtering, it may be viewed as a derivative-free optimization method. In its most basic form it regularizes ill-posed inverse problems through the subspace property: the solution found is in the linear span of the initial ensemble employed. We demonstrate how further regularization can be imposed, incorporating prior information about the underlying unknown. In particular, we study how to impose Tikhonov-like Sobolev penalties. As well as introducing this modified ensemble Kalman inversion methodology, we also study its continuous-time limit, proving ensemble collapse; in the language of multi-agent optimization, this may be viewed as reaching consensus. We also conduct a suite of numerical experiments to highlight the benefits of Tikhonov regularization in the ensemble inversion context.

EO135 Room AT337 NEW DEVELOPMENTS IN DIMENSION REDUCTION WITH COMPLEX DATA Chair: Teng Zhang

E0544: **Nonlinear and additive principal component analysis for functional data**
Presenter: **Jun Song**, University of North Carolina at Charlotte, United States
A nonlinear and additive version of principal component analysis for vector-valued random functions defined on an interval is developed. This is a generalization of functional principal component analysis that allows the relations among the random functions involved to be nonlinear. The method is constructed via two additively nested Hilbert spaces of functions, in which the first space characterizes the functional nature of the data, and the second space captures the nonlinear dependence. In the meantime, additivity is imposed so that we can avoid high-dimensional kernels in the functional space, which causes the curse of dimensionality. Simulation results show that the new method performs better than functional principal component analysis when the relations among random functions are nonlinear. We apply the new method to online handwritten digits data sets.

E0426: **Functional graphical modeling via Karhunen-Loève expansions**
Presenter: **Kuang-Yao Lee**, Temple University, United States
Co-authors: Lexin Li, Hongyu Zhao, Dingjue Ji, Todd Constable
Network estimation for multivariate functional data is becoming increasingly important in a wide variety of applications. The changes of graph structures can often be attributed to external variables such as other phenotypes observed in the data, or a time variable such as the subject's age. The latter gives rise to the problem of dynamic graphical modeling. Most existing methods focus on the random variable setting, and estimate the graph by aggregating samples, sometimes according to the diagnostic groups, but largely ignore the subject-level heterogeneity. We target graphical modeling of multivariate random functions, and treat the external variables as the conditioning set. We propose a new class of conditional functional graphical model that allows the graph links to vary along with the external variables. We develop two linear operators, the conditional precision operator and the conditional partial correlation operator, which generalize the precision matrix and the partial correlation matrix from the random variable setting to both the conditional and functional settings, and based on which we can estimate the conditional functional graph. We establish the error bounds of the corresponding estimators and the consistency of the conditional graph estimation. We demonstrate the efficacy of the proposed method through both simulations and a study of brain functional connectivity network.

E0446: **Envelope-based sparse partial least squares**
Presenter: **Guangyu Zhu**, University of Rhode Island, United States
Co-authors: Zhihua Su
Sparse partial least squares is a widely used method that performs dimension reduction and variable selection simultaneously in linear regression. Despite its popularity in applied sciences, its theoretical properties are largely unknown. We use a connection between envelope models and partial least squares (PLS) to construct an envelope-based SPLS estimator and establish its consistency, oracle property and asymptotic normality. The large-sample scenario and high-dimensional scenario are both considered. We also develop the envelope-based SPLS estimators under the context of generalized linear models, and discuss its theoretical properties including consistency, oracle property and asymptotic distribution. Numerical experiments and examples show that the envelope-based SPLS estimator has better variable selection and prediction performance over the SPLS estimator.

E0292: **A comprehensive Bayesian framework for envelope models**
Presenter: **Zhihua Su**, University of Florida, United States
Co-authors: Saptarshi Chakraborty
The envelope model is a nascent construct that aims to increase efficiency in multivariate analysis. 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 showed the potential to provide substantial efficiency gains. Virtually all of these advances, however, have been made from a frequentist perspective, and the literature addressing envelope models from a Bayesian point of view is sparse. The objective is to propose a Bayesian framework that is applicable across various envelope model contexts. The proposed framework aids straightforward interpretation of model parameters and allows easy incorporation of prior information. We provide a simple block Metropolis-within-Gibbs MCMC sampler for practical implementation of our method. Simulations and data examples are included for illustration.

EO193 Room U301 STATISTICAL ANALYSIS OF STOCHASTIC PROCESSES Chair: Keisuke Yano

E0417: **Bootstrap confidence bands for spectral estimation of Lévy densities under high-frequency observations**
Presenter: **Daisuke Kurisu**, Tokyo Institute of Technology, Japan
Co-authors: Kengo Kato
Bootstrap methods are developed to construct uniform confidence bands for nonparametric spectral estimation of Lévy densities under high-frequency observations.We assume that we observe n discrete observations at frequency $1/\Delta > 0$, and work with the high-frequency setup where $\Delta = \Delta_n \to 0$ and $n\Delta \to \infty$ as $n \to \infty$. We employ a spectral (or Fourier-based) estimator of the Lévy density, and develop novel implementations of Gaussian multiplier (or wild) and empirical (or Efron's) bootstraps to construct confidence bands for the spectral estimator on a compact set that does not intersect the origin. We provide conditions under which the proposed confidence bands are asymptotically valid. Our confidence bands are shown to be asymptotically valid for a wide class of Lévy processes.We also develop a practical method for bandwidth selection, conduct simulation studies to investigate the finite sample performance of the proposed confidence bands, and present a small real data analysis.

E0558: **Noise estimation for ergodic Levy driven stochastic differential equation model**
Presenter: **Yuma Uehara**, The Institute of Statistical Mathematics, Japan
Co-authors: Hiroki Masuda
To describe non-Gaussian activity in high frequency data obtained from financial, biological, and technological phenomenon, Levy driven stochastic differential equations serve as good candidates. Since the closed form of its genuine likelihood is generally not obtained, the estimation of its driving noise is often done by empirical moment fittings with respect to its Levy measure. However, the measure sometimes takes complex form, and thus intractable. For such a problem, we consider the approximation of unit time increments of the driving noise based on the Euler residual. By making use of this approximation, we can conduct parametric estimation methods of the driving noise with bias correction. We will present its theoretical properties and show some numerical experiments.

E0518: **Penalized methods for quasi likelihood analysis with locally asymptotic quadratic properties**
Presenter: **Yoshiki Kinoshita**, The University of Tokyo, Japan
Co-authors: Nakahiro Yoshida
Penalized methods are applied to the quasi likelihood analysis in order to perform a variable selection in stochastic models. We consider locally asymptotic quadratic properties to describe a local structure of quasi likelihood functions. The polynomial type large deviation inequality is introduced to bound the tail of quasi likelihood functions. We estimate not only the limiting distribution of the estimator but also the probability that correct model is selected.

E0570: **Is volatility rough**
Presenter: **Tetsuya Takabatake**, Osaka university, Japan
Co-authors: Masaaki Fukasawa, Rebecca Westphal
In recent years, the rough volatility model, which is a kind of stochastic volatility models whose log-volatility process is driven by a fractional Brownian motion with the very small Hurst parameter, attracts much attention in the community of Mathematical Finance because the rough volatility model can reproduce many stylized facts in financial markets. From statistical point of view, the previous studies are, however, not satisfactory in several aspects. Among others, we would like to emphasize that there is no estimator for the roughness of volatility whose consistency is proven so far. We propose a quasi-likelihood estimator for the Hurst and diffusion parameters of the fractional Brownian motion driving the volatility process based on high frequently observed realized volatility data, and prove its consistency under high frequency asymptotics. Moreover, we examine the finite sample performance of our estimator by simulations, and apply our estimator to empirical analysis by using the realized volatility data provided by the Oxford-Man realized library. Our data analysis suggests that the volatility is indeed rough; actually it is much rougher than considered in previous studies.

EO205 Room U302 ACTUARIAL AND FINANCIAL RISK MANAGEMENT Chair: Boris Choy

E0425: **Analytic solution of a portfolio optimization problem in a mean-variance-skewness model**
Presenter: **Zinoviy Landsman**, University of Haifa, Israel
Co-authors: Udi Makov, Tomer Shushi
In portfolio theory, it is well-known that in most of the cases stocks follows a non-symmetric and unimodal distributions. Therefore, many researches have suggested considering the skew-normal distribution an accurate model in quantitative finance. From the fact that the portfolio of stocks is non-symmetric, the celebrated mean-variance theory fails to provide an optimal portfolio selection rule, which comes from the fact that the mean-variance model does not take into account the skewness of the stocks. We provide a novel approach that solves such a problem of optimal portfolio selection with non-symmetric stocks, by putting it into a framework of mean-variance-skewness measure. For example, we show an analytical portfolio optimization solution to the exponential utility of the well-known skew-normal distribution or, even more general, scale mixtures of skew-normal distribution. Moreover, our optimal solutions are explicit and has closed-forms, and therefore they can be investigated in comparison to other portfolio selection rules, such as the standard mean-variance model. The results are then illustrated numerically.

E0484: **Multivariate long memory cohort mortality models**
Presenter: **Jennifer Chan**, University of Sydney, Australia
The existence of long memory in mortality data improves the understandings of features of mortality data and provides a new approach for establishing mortality models. The findings of long memory phenomena in mortality data motivate us to develop new mortality models by extending the Lee-Carter (LC) model to death counts and incorporating long memory model structure. Furthermore, there is no identification issues arising in the proposed model class. Hence, the constraints which cause many computational issues in LC models are removed. The models are applied to analyse mortality death count data sets from three different countries divided according to genders. Bayesian inference with various selection criteria is applied to perform the model parameter estimation and mortality rate forecasting. Results show that multivariate long memory mortality model with long memory cohort effect (LMLM) model outperform multivariate extended LC cohort (MELCC) model in both in-sample fitting and outsample forecast. Increasing the accuracy of forecasting of mortality rates and improving the projection of life expectancy is an important consideration for insurance companies and governments since misleading predictions may result in insufficient funds for retirement and pension plans.

E0486: **Semi-parametric dynamic asymmetric Laplace models for tail risk forecasting, incorporating realized measures**
Presenter: **Chao Wang**, The University of Sydney, Australia
Co-authors: Richard Gerlach, Chao Wang
The joint Value-at-Risk (VaR) and expected shortfall (ES) quantile regression model is extended, via incorporating a realized measure to drive the tail risk dynamics, as a potentially more efficient driver than daily returns. Further, a new model for the dynamics of the ES component is proposed and tested. Both a maximum likelihood and an adaptive Bayesian Markov Chain Monte Carlo method are employed for estimation, whose properties are compared in a simulation study; results favour the Bayesian approach, subsequently employed in a forecasting study of seven financial market indices. The proposed models are compared to a range of parametric, non-parametric and semi-parametric competitors, including GARCH, Realized GARCH, Extreme Value Theory method and the joint VaR and ES models, in terms of accuracy of one-day-ahead VaR and ES forecasts, over a long forecast sample period that includes the global financial crisis in 2007-2008. The results are favorable for the proposed models incorporating a realized measure, especially when employing the sub-sampled Realized Variance and the sub-sampled Realized Range.

E0521: **Actuarial applications of a new Weibull-type distribution**
Presenter: **Boris Choy**, University of Sydney, Australia
An extension to the class of contaminated Weibull distribution is proposed for statistical modelling. The extra parameter in the proposed distribution allows for more flexibility in the estimation of moments and the modelling of the tail behaviour. The properties of the distribution are presented and compared with the existing contaminated Weibull distribution. Bayesian simulation-based methods will be used for statistical inference. We shall demonstrate the adequacy of the proposed distribution in modelling positively-valued data in actuarial applications.

EO191 Room U414 ECONOMETRIC MODELLING: METHODOLOGY AND APPLICATION Chair: Hung-pin Lai

E0258: **A generalized focused information criterion for extremum estimators**
Presenter: **Chu-An Liu**, Academia Sinica, Taiwan
Co-authors: Xinyu Zhang
Unlike the traditional model selection criterion, which picks a single model based on the global fit of the model, the Focused Information Criterion (FIC) is tailored to the parameter of interest and aims to select a model based on the parameter under focus. The aim is to develop FIC for different models in an unified theoretical framework and to propose a Generalized Focused Information Criterion (GFIC) for extremum estimators. To illustrate GFIC, we apply our theoretical results to several examples, including the maximum likelihood estimator, the nonlinear least squares estimator, the generalized method of moments estimator, and the minimum distance estimator.

E0367: **Monotonicity test for local average treatment effects under regression discontinuity**
Presenter: **Yu-Chin Hsu**, Academia Sinica, Taiwan
Researchers are often interested in the relationship between treatment effects and observed individual heterogeneity. The first nonparametric monotonicity test under the popular regression discontinuity framework is proposed. The proposed test examines whether the average treatment effect or the local average treatment effect has a monotonic relationship with some of the observed individual characteristics. We show consistency and asymptotic uniform size control of the proposed test. We apply the test to study the heterogeneous effect of attending a more selective high school with respect to peer quality.

E0487: **Estimating nonparametric Berkson measurement error models**
Presenter: **Ji-Liang Shiu**, Jinan University, China
Non-parametric identification and estimation of Berkson measurement error models with conditional mean independent regression error is shown. The identification result is achieved without requiring side information and specifying the distribution of the measurement error. We first apply techniques of integral operators associated with the distributions of observable and unobservable variables to identify the moments of the measurement error. Based on the identification of the moments, the moments of the regression function can be uniquely determined.Additional identification restrictions include conditional distribution $f_{Y|X*}$ is complete and non-vanishing Fourier transforms of the regression function. We then use the identification result to construct a sieve minimum distance (MD) estimator to estimate the regression function and the distribution of the measurement error. We investigate the finite sample properties of the proposed sieve MD estimator through a Monte Carlo study.

E0637: **Averaging estimators for heterogeneous dynamic panel regressions with weakly exogenous regressors and multifactor error**
Presenter: **Chang-Ching Lin**, National Cheng Kung University, Taiwan
Co-authors: Shou-Yung Yin
Model averaging is considered in heterogeneous dynamic panel regressions with weak exogenous regressors and a multifactor error structure. Under a local to zero framework, we show that the common correlated effects mean group (CCEMG) estimator exists three different components of the asymptotic bias, the fundamental bias from ignoring variables, time series bias and the bias from the truncation of number of lags of augmented regressors used to approximate unobserved factor structure. We then propose a focused information criterion and a plug-in averaging estimator based on the half-panel jackknife bias-corrected CCEMG estimators for the full model and all submodels. Since the fundamental bias and the bias from the truncation cannot be corrected, the trade-off between bias and variance exists, and the proposed methods can minimize the asymptotic mean squared errors. Monte Carlo simulations show that the proposed averaging method generally outperforms other methods. An empirical study on the relationship between the commodity price volatility and the economic growth is considered as an application.

EO161 Room U502 MODERN DEVELOPMENTS IN CHANGE-POINT DETECTION ANALYSIS Chair: Jun Li

E0209: **Inference for change points in high dimensional data**
Presenter: **Xiaofeng Shao**, University of Illinois at Urbana-Champaign, United States
The aim is to present some recent work on change point testing and estimation for high dimensional data via self-normalization, which was developed for low dimensional time series recently. In the case of testing for a mean shift, we propose a new test which is based on U-statistics and utilizes the self-normalization principle. Our test targets dense alternatives in the high dimensional setting and involves no tuning parameters. We show the weak convergence of a sequential U-statistic based process to derive the pivotal limit under the null and also obtain the asymptotic power under the local alternatives. In addition, we illustrate how our approach can be used in combination with wild binary segmentation to estimate the number and location of multiple unknown change points.

E0211: **A self-normalized approach to sequential change-point detection for time series**
Presenter: **Wai Leong Ng**, The Chinese University of Hong Kong, Hong Kong
Co-authors: Chun Yip Yau, NH Chan
A self-normalization sequential change-point detection method for time series is proposed. In testing for parameter changes, most of the traditional sequential monitoring tests utilize a CUSUM-based test statistic, which involves a long-run variance estimator. However, the commonly used long-run variance estimators require the choice of bandwidth parameter which could be sensitive to the performance. Moreover, the traditional tests usually suffer from severe size distortion due to the slow convergence rate to the limiting distribution in the early monitoring stage. A self-normalization approach is implemented to tackle these issues. We establish null asymptotic and the consistency of the proposed sequential change-point test under general regularity conditions. Simulation experiments and empirical application to railway bearing temperature data are conducted for illustrations.

E0261: **Finite sample change point inference and identification for high-dimensional mean vectors**
Presenter: **Xiaohui Chen**, University of Illinois at Urbana-Champaign, United States
Co-authors: Mengjia Yu
Cumulative sum (CUSUM) statistics are widely used in the change point inference and identification. We discuss two problems for high-dimensional mean vectors based on the max-norm of CUSUM statistics. For the problem of testing for the existence of a change point in an independent sample generated from the mean-shift model, we propose a Gaussian multiplier bootstrap to calibrate critical values of the CUSUM test statistics in high dimensions. The proposed bootstrap CUSUM test is fully data-dependent and it has strong theoretical guarantees under arbitrary dependence structures and mild moment conditions. Specifically, we show that with a boundary removal parameter the bootstrap CUSUM test enjoys the uniform validity in size under the null and it achieves the minimax separation rate under the sparse alternatives when the dimension p can be larger than the sample size n. Once a change point is detected, we estimate the change point location by maximizing the max-norm of the CUSUM statistics at two different weighting scales. The first estimator is based on the covariance stationary CUSUM statistics, and the second estimator is based on non-stationary CUSUM statistics assigning less weights to the boundary data points. In the latter case, we show that it achieves the nearly best possible rate of convergence. In both cases, dimension impacts the rate of convergence only through the logarithm factors, and consistency of the CUSUM location estimators is possible when $p >> n$.

E0678: **Sequential subspace change-point detection**
Presenter: **Yao Xie**, Georgia Institute of Technology, United States
Co-authors: George Moustakides, Liyan Xie
The focus is on the sequential changepoint detection problem of detecting changes that are characterized by a subspace structure which is manifested in the covariance matrix. In particular, the covariance structure changes from an identity matrix to an unknown spiked covariance model. We consider three sequential changepoint detection procedures: The exact cumulative sum (CUSUM) that assumes knowledge of all parameters, the largest eigenvalue procedure and a novel Subspace-CUSUM algorithm with the last two being used for the case when unknown parameters are present. By leveraging the extreme eigenvalue distribution from random matrix theory and modeling the non-negligible temporal correlation in the sequence of detection statistics due to the sliding window approach, we provide theoretical approximations to the average run length (ARL) and the expected detection delay (EDD) for the largest eigenvalue procedure. The three methods are compared to each other using simulations.

EO123 Room U517 INFERENCE FOR NONSTATIONARY AND NONSTANDARD TIME SERIES MODELS Chair: Fumiya Akashi

E0200: **Goodness-of-fit tests for Markovian processes based on marked empirical processes**
Presenter: **Koji Tsukuda**, The University of Tokyo, Japan
Co-authors: Yoichi Nishiyama
Weak convergences of marked empirical processes in $L^2(\mathbb{R},\nu)$ and their applications to statistical goodness-of-fit tests are provided, where $L^2(\mathbb{R},\nu)$ is the set of equivalence classes of the square integrable functions on $\mathbb{R}$ with respect to a finite Borel measure ν. The results obtained in our framework of weak convergences are, in the topological sense, weaker than those in previous works. However, our results have the following merits: (1) avoiding conditions which do not suit for our purpose; (2) treating a weight function which make us possible to propose an Anderson–Darling type test statistics for goodness-of-fit tests. Indeed, applications are novel.

E0316: **Robust causality test of infinite variance processes**
Presenter: **Fumiya Akashi**, The University of Tokyo, Japan
Co-authors: Masanobu Taniguchi, Anna Clara Monti
A robust causality test is developed for time series models with infinite variance innovation processes. First, we introduce a measure of dependence for vector nonparametric linear processes and derive the limit distribution of a previous test statistic in the infinite variance case. Second, we construct a weighted-version of the generalized empirical likelihood (GEL) test statistic, called the self-weighted GEL statistic in time domain. The limit distribution of the self-weighted GEL test statistic is shown to be a standard chi-squared one regardless of whether the model has finite variance or not. Some simulation experiments illustrate desired finite sample performance of the proposed method.

E0515: **Weak convergence of the partial sum of $I(d)$ process to a fractional Brownian motion in finite interval representation**
Presenter: **Junichi Hirukawa**, Niigata University, Japan
Co-authors: Kou Fujimori
An integral transformation which changes a fractional Brownian motion to a process with independent increments has been given. A representation of a fractional Brownian motion through a standard Brownian motion on a finite interval has also been given. On the other hand, it is known that the partial sum of the discrete time fractionally integrated process ($I(d)$ process) weakly converges to a fractional Brownian motion in infinite interval representation. We derive the weak convergence of the partial sum of $I(d)$ process to a fractional Brownian motion in finite interval representation.

E0569: **A numerically efficient closed-form representation of mean-variance hedging for exponential additive processes**
Presenter: **Yuto Imai**, Tokyo Metropolitan University, Japan
Co-authors: Takuji Arai
The focus is on a mean-variance hedging problem for models whose asset price follows an exponential additive process. Some representations of mean-variance hedging strategies for jump-type models have already been suggested, but none is suited to develop numerical methods of the values of strategies for any given time up to the maturity. We aim to derive a new explicit closed-form representation, which enables us to develop an efficient numerical method using the fast Fourier transforms. Note that our representation is described in terms of Malliavin derivatives. In addition, we illustrate numerical results for exponential Levy models.

Wednesday 26.06.2019 14:00 - 15:40 Parallel Session H – EcoSta2019

EI009 Room UB99(B1) ECOSTA JOURNAL INVITED SESSION Chair: Erricos Kontoghiorghes

E0704: **Bias correction for local linear regression estimation using asymmetric kernels via the skewing method**
Presenter: **Masayuki Hirukawa**, Ryukoku University, Japan
Co-authors: Benedikt Funke
The skewing method that has been originally proposed as a bias correction device is extended for local linear regression estimation using standard symmetric kernels to the cases of asymmetric kernels. The method is defined as a convex combination of three local linear estimators. It is demonstrated that the skewed estimator using asymmetric kernels with properly chosen weights can accelerate the bias convergence from $O(b)$ to $O(b^2)$ as $b- > 0$ under sufficient smoothness of the unknown regression curve while not inflating the variance in an order of magnitude, where b is the smoothing parameter and the regressor is assumed to have at least one boundary. As a consequence, the estimator has optimal pointwise convergence of $n^{4/9}$ when best implemented, where n is the sample size. It is noteworthy that these properties are the same as those for a local cubic regression estimator. Finite-sample properties of the skewed estimator are assessed in comparison with local linear and local cubic estimators. An application of the skewed estimation to real data is also considered.

E0245: **Joint analysis of mixed types of outcomes with latent variables**
Presenter: **Xinyuan Song**, Chinese University of Hong Kong, Hong Kong
A joint modeling approach is proposed to investigate the observed and latent risk factors of mixed types of outcomes. The proposed model comprises three parts. The first part is an exploratory factor analysis model that summarizes latent factors through multiple observed variables. The second part is a proportional hazards model that examines the observed and latent risk factors of multivariate time-to-event outcomes. The third part is a linear regression model that investigates the determinants of a continuous outcome. We develop a Bayesian approach coupled with efficient MCMC methods to determine the number of latent factors, the association between latent and observed variables, and the important risk factors of different types of outcomes. A modified stochastic search item selection algorithm that introduces normal-mixture-inverse gamma priors to factor loadings and regression coefficients is developed for simultaneous model selection and parameter estimation. The proposed method is subjected to simulation studies for empirical performance assessment and then applied to a study concerning the risk factors of type 2 diabetes and the associated complications.

E0686: **Copula-based models of productivity and efficiency**
Presenter: **Artem Prokhorov**, University of Sydney, Australia
The aim is to discuss how to use copulas to measure firm efficiency and productivity. Copulas are mathematical concepts that model dependence between random variables. Traditional econometric analysis of efficiency and productivity based on parametric production functions (commonly referred to as stochastic frontier analysis) has ignored important aspects of production where dependence plays a key role. The dependence influences simultaneous decisions on how much to produce and what proportions of inputs to use. It influences how factors outside our control such as extreme weather or new legislation, affect how inefficient we are. Appropriate copulas capturing these dependencies allow for robust estimation and testing of production models and achieve remarkable improvements in productivity and efficiency. The focus will be on copula estimation for stochastic frontier analysis. The use of copulas to capture various dependencies that have not been accounted for before during production will be considered. One such important case study is endogenous choice of production inputs. Such endogeneity, if ignored, leads to biased estimates of productivity and return to scale and may understate inefficiency. Another is dependence overtime in a panel data setting, which allows for a more precise estimation of technical inefficiencies. Yet another is the case of modelling joint patterns in technical and allocative inefficiencies of firms.

EO306 Room S101 STATISTICAL LEARNING FOR DATA WITH DISTINCT CHARACTERISTICS Chair: Peng Wang

E0592: **Smooth collaborative recommender system**
Presenter: **Junhui Wang**, City University of Hong Kong, Hong Kong
In recent years, there has been a growing demand to develop efficient recommender systems which track users' preferences and recommend potential items of interest to users. We will present a smooth collaborative recommender system to utilize dependency information among users and items which share similar characteristics under the singular value decomposition framework. The proposed method incorporates the neighborhood structure among user-item pairs by exploiting covariates to improve the prediction performance. One key advantage of the proposed method is that it leads to more effective recommendation for "cold-start" users and items, whose preference information is completely missing from the training set. As this type of data involves large-scale customer records, efficient scheme will be proposed to achieve scalable computing. The advantage is confirmed in a variety of simulated experiments as well as one large-scale real example on Last.fm music listening counts. If time permits, the asymptotic properties will also be discussed.

E0607: **Weak signal identification and inference in penalized likelihood models**
Presenter: **Yuexia Zhang**, Fudan University, China
Co-authors: Annie Qu, Zhongyi Zhu, Peibei Shi
Penalized model selection is important when the number of covariates is large and the sample size is not large enough in the data. When the signal is weak, the existing model selection approaches have some limitations. For example, the estimator of coefficient is likely to shrink to zero and the confidence interval tends to be under-coverage. For the linear regression model, there have been some methods to deal with these problems, but the extension of these methods to general likelihood model is not easy. To estimate the coefficients in general likelihood model, the one-step penalized likelihood method is used. To identify the signal strength level, a new indicator is proposed. After signal identification, a two-step inference procedure is developed to construct confidence interval for the estimator of coefficients. Both finite sample theory and numerical study indicate that the proposed method leads to better confidence coverage for weak signals, compared with several existing methods. In the end, the proposed method is applied to the study of Practice Fusion Diabetes dataset for illustration.

E0631: **Determinants of corporate bankruptcy: Identification and uncertainty**
Presenter: **Yan Yu**, University of Cincinnati, United States
Corporate bankruptcy is both socially influential and economically detrimental. Hence, research in corporate bankruptcy is critically important. Some natural questions arise: 1) What are important determinants of corporate bankruptcy? 2) How reliable are the determinants identified? the aim is to identify important determinants and investigate their uncertainly. We build a comprehensive bankruptcy database including as many determinants proposed in the literature as possible merging both accounting statements and market information. We explore various variable selection tools and find that important determinants include both market variables and accounting ratios. More importantly, we construct a model confidence bound on the selected determinants. We further propose a graphical representation of the model uncertainty and define deviation and skewness measures to facilitate an understanding on how reliable the identified important determinants are. We find that among eight predictive

variables, stock price is identified with high confidence, while stock excess return is included within the model confidence bound, yet market to book ratio is outside the model confidence bound.

E0645: **Machine learning for spatially aggregated data**
Presenter: **Peng Wang**, University of Cincinnati, United States
Co-authors: Bo Li
In recent years, statistical machine learning approaches has been extremely popular largely due to its superior performance in prediction. Of all the commonly used machine learning tools, the gradient boosting tree is usually the favored vehicle for many practitioners. On the popular data analytics competition platform Kaggle, gradient boosting is the winning algorithm for almost every structured data. Besides its superior prediction performance, the gradient boosting trees also enjoys the interpretablility of a non-parametric additive model and its fitting algorithm can be paralleled. We extend this powerful machine learning technique to the realm of spatial data analysis. The proposed approach does not require any parametric assumption on the spatial correlations and enjoy all the advantages of gradient boosting. We illustrate the potential of the data with application on prediction of HIV new diagnose rates for all counties of the United States.

EO163 Room S102 CHALLENGES OF STATISTICS AND HEALTH ECONOMICS RESEARCH IN ONCOLOGY Chair: Yu Shen

E0251: **Conditions for valid estimation of cancer overdiagnosis using excess incidence from screening trials**
Presenter: **Roman Gulati**, Fred Hutchinson Cancer Research Center, United States
Co-authors: Ruth Etzioni
The primary harm of cancer screening is the detection of cancers that would not have been diagnosed during the patient's lifetime without screening. A common approach to estimating the frequency of these so-called overdiagnosed cancers uses the excess incidence in screened relative to unscreened individuals. However, conditions for validity of this approach have not been established. We develop a transparent model of the effects of screening on cancer incidence and use the model to examine dynamics of excess incidence in stop-screening and continued-screening trials with and without overdiagnosis. We find that excess incidence yields an unbiased estimate of overdiagnosis only if (1) follow-up exceeds the maximum preclinical period after screening stabilizes and (2) incidence is appropriately quantified with respect to the trial design–i.e., using cumulative incidence for stop-screening trials and annual incidence for continued-screening trials. We illustrate these points using simulations and use the results to assess the validity of published overdiagnosis estimates from cancer screening trials. We discuss the bias-variance tradeoff and consider extensions to account for dependence of disease risk on age. We conclude that, without quantitative knowledge about underlying cancer natural history, estimates of cancer overdiagnosis using excess incidence are frequently unreliable in practice.

E0252: **Model-assisted adaptive designs in oncology**
Presenter: **J Jack Lee**, University of Texas MD Anderson Cancer Center, United States
Oncology drug development is a long, arduous, and expensive process. Recent report shows that the cost of developing a drug is at $2.6 billion. The overall success rate from Phase I trials to FDA approval was 9.6% with oncology drug at 5.1%. Bayesian adaptive designs have been shown to increase efficiency, allow more flexible trial conduct, and treat more patients with more effective treatments. However, model-based Bayesian adaptive designs have not gained much traction largely because computation burdens. Although software tools are available to ease the design and implementation, many such designs are still being viewed as too complicated and too difficult to do. Conversely, algorithm-based methods, such as the 3+3 design for Phase I studies, remain popular in spite of inferior operating characteristics. In viewing of this quandary, several model-assisted designs have been developed to attain superior statistical properties while provide simple rules for trial conduct. Easy-to-use tools for trial design and conduct have also been introduced. These include the Bayesian optimal interval designs for single and combination agents (Phase I) and the Bayesian optimal Phase II design for simple and complex endpoints. The model-assisted design fulfilled the new Keep it Simple and Smart (KISS) principle. This new class of model-assisted design can be easily and widely applied to treat patients better and to improve the success rate of drug development.

E0306: **Impact of new technology diffusion on the costs of renal cell cancer**
Presenter: **Ya-Chen Tina Shih**, University of Texas MD Anderson Cancer Center, United States
The impact of new oncologic technologies on the costs of renal cell cancer (RCC) is examined. We used the linked Surveillance, Epidemiology, and End Results (SEER)-Medicare database and employed both prevalence and incidence costing approaches. We conducted longitudinal analysis of cost data per patient per month (PPPM) for a prevalence cohort of patients to determine which category of new technology (surgery, radiation, or chemotherapy) was the major cost driver for RCC. We then applied the incidence costing approach to estimate cost related to RCC by care phases (initial, continuing, and terminal) and compared costs between two incidence cohorts to examine how new technology affected RCC costs over time. After controlling for demographic factors, clinical characteristics, neighborhood socioeconomic status, and time trend, we found that rising PPPM costs was driven primarily by new technologies in chemotherapy. Incidence-based analysis showed the annual cost (2018 US$) for distant stage RCC patients diagnosed between 2002 and 2006 was $51,639, $19,025, $76,603, and $29,045, for initial, continuing (year 1), terminal (died from RCC), and terminal (died from other causes) care phase, respectively. Costs increased to $70,703, $34,716, $107,989, and $47,538, respectively, for the incidence cohort diagnosed between 2007 and 2011.

E0538: **Recent development of statistical methods in cancer screening studies**
Presenter: **Yu Shen**, UT MD Anderson Cancer Center, United States
Statistical modeling is an effective tool to estimate medical costs as well as effectiveness in both cancer screening programs and cancer treatment, which is an important top in health policy and health economics research. Over last decades, many of cancer screening trials have been conducted for breast, lung, colon and prostate. These trials generate data, which may be used to estimate the preclinical sojourn time distribution and screening sensitivity and other quantities of interest from the screening cohort. The information on the natural history of cancer is critical in designing optimal screening programs and assessing screening benefit. We will review some existing statistical approaches in this area and the implications.

EO027 Room S104 RECENT DEVELOMENTS IN HIGH-DIMENSIONAL STATISTICAL ANALYSIS Chair: Makoto Aoshima

E0354: **Variable selection for high-dimensional regression models with time series and heteroscedastic errors**
Presenter: **Hai-Tang Chiou**, National Tsing Hua University, Taiwan
Co-authors: Meihui Guo, Ching-Kang Ing
Although existing literature on high-dimensional regression models is rich, the vast majority of studies have focused on independent and homogeneous error terms. We consider the problem of selecting high-dimensional regression models with heteroscedastic and time series errors, which have broad applications in economics, quantitative finance, environmental science, and many other fields. The error term in our model is not only allowed to be short- or long-range dependent, but also contains a high-dimensional dispersion function accounting for heteroscedasticity. By making use of the orthogonal greedy algorithm and the high-dimensional information criterion, we propose a new model selection procedure that can consistently choose the relevant variables in both the regression and the dispersion functions. The finite sample performance of the proposed procedure is also illustrated via simulations and real data analysis.

E0475: **Inference on mean vectors for high-dimensional data with the strongly spiked eigenstructure**
Presenter: **Aki Ishii**, Tokyo University of Science, Japan
Co-authors: Kazuyoshi Yata, Makoto Aoshima
Constructing theories and methodologies for high-dimensional data has become increasingly important in many fields. It is known that high-dimensional data include strongly spiked noise and the noise is troublesome when we analyze high-dimensional data. A lot of conventional methods are heavily influenced by such huge noise and cannot claim accuracy. We note that such huge noise generates a strongly spiked eigenstructure. In order to remove the strongly spiked noise, a data transformation technique has been previously developed. First, we consider one-sample test by using the data transformation. We construct a new test procedure and show that our test procedure works well both in theory and simulation. We also apply the test procedure to multisample problem. Finally, we give some demonstrations by using famous microarray data sets.

E0444: **A high-dimensional quadratic classifier by data transformation for strongly spiked eigenvalue models**
Presenter: **Kazuyoshi Yata**, University of Tsukuba, Japan
Co-authors: Aki Ishii, Makoto Aoshima
High-dimensional classification is considered. Any high-dimensional data is classified into two disjoint models: the strongly spiked eigenvalue (SSE) model and the non-SSE (NSSE) model. In actual high-dimensional data, one often finds a non-sparse structure which contains strongly spiked eigenvalues. That structure fits the SSE model. There are several studies providing high-dimensional classifiers. However, it should be noted that one cannot usually obtain a consistency property of the classifiers under the SSE model. It is because the classifiers are heavily influenced by the strongly spiked eigenvalues. In order to overcome the difficulty, a data transformation technique that transform the SSE model to the non-SSE model has been previously developed. We propose a quadratic classification procedure by using the data transformation. We prove that our proposed classification procedure has a consistency property for misclassification rates under the SSE model. We discuss performances of our classification procedure in simulations and real data analyses using microarray data sets.

E0511: **Towards a sparse, scalable, and stably positive definite (inverse) covariance estimator**
Presenter: **Joong-Ho Won**, Seoul National University, Korea, South
High dimensional covariance and inverse covariance matrix estimation is notoriously difficult as the traditional estimate is not even positive definite. An important line of research in this regard is to shrink the extreme spectrum of the covariance matrix estimators. A separate line of research has considered sparse inverse covariance estimation which in turn gives rise to graphical models. In practice, however, a sparse covariance or inverse covariance matrix which is simultaneously well-conditioned and at the same time computationally tractable is desired. We consider imposing a condition number constraint to various types of losses used in covariance and inverse covariance matrix estimation. When the loss function can be decomposed as a sum of an orthogonally invariant function of the estimate and its inner product with a function of the sample covariance matrix, we show that a complete solution path can be obtained, involving a series of ordinary differential equations. An important finding is that the proximal operator for the condition number constraint, which turns out to be very useful in regularizing loss functions that are not orthogonally invariant and may yield non-positive-definite estimates, can be efficiently computed by this path algorithm.

EO355 Room S106 STATISTICAL ANALYSIS OF COMPLEX STRUCTURED DATA Chair: Wenqing He

E0329: **Solar flare predictions with statistical learning**
Presenter: **Yang Chen**, University of Michigan, United States
Over the space age, extensive knowledge has been accumulated about the regions of space surrounding the Earth and the Sun, and the governing physical processes controlling space weather in these regions. However, this knowledge has not been translated into an operational forecast capability. By combining our expertise in space weather modeling and data science/machine learning we can not only address the "holy grail" of space weather prediction and extend the forecast horizon from minutes to days, but also transition the results to space weather operations. The current space weather predictive capabilities are either short term and/or not accurate and reliable. We initiated a research program that will (hopefully) answer these questions using a unique combination of modeling, computation, and massive amounts of space weather data collected by satellites that will be used to train state-of-the-art machine learning algorithms.

E0332: **Data adaptive support vector machine with imbalanced observations**
Presenter: **Wenqing He**, University of Western Ontario, Canada
Support vector machines (SVM) have been widely used as classifiers in various settings. However, such methods are faced with newly emerging challenges such as imbalanced observations and noise data. We will present an SVM method using a data-adaptive kernel to feature imbalanced observations by considering the location of support vectors in the feature space and thereby generates more accurate classification results. The performance of the proposed method is compared with existing methods using numerical studies and illustrated through a prostate cancer image example.

E0514: **A support vector machine based semiparametric mixture cure model**
Presenter: **Yingwei Peng**, Queenś University, Canada
The mixture cure model is an extension of standard survival models to analyze survival data with a cured fraction. Many developments in recent years focus on the latency part of the model to allow more flexible modeling strategies for the distribution of uncured subjects, and fewer studies focus on the incidence part to model the probability of being uncured/cured. We propose a new mixture cure model that employs the support vector machine (SVM) to model the covariate effects in the incidence part of the cure model. The new model inherits the features of the SVM to provide a flexible model to assess the effects of covariates on the incidence. Unlike the existing nonparametric approaches for the incidence part, the SVM method also allows for potentially high-dimensional covariates in the incidence part. Semiparametric models are also allowed in the latency part of the proposed model. We develop an estimation method to estimate the cure model and conduct a simulation study to show that the proposed model outperforms existing cure models, particularly in incidence estimation. An illustrative example using data from leukemia patients is given.

E0531: **Complexities in the analysis of electronic health records: An illustrative case of a chronic kidney disease study**
Presenter: **Guofen Yan**, University of Virginia, United States
While randomization is generally considered the gold standard approach to address clinical questions such as efficacy of an experimental treatment, some questions may only be addressed using observational data, such as the relationship of race and outcomes. Chronic kidney disease (CKD) is a major public health problem in the US and worldwide, with the current estimated prevalence of 10-12% in the US. Many questions remain unanswered, from simple facts like the distribution of onset-age to more involved issues such as the differential survival rates among racial and ethnic groups. Utilizing the U.S. veteran electronic health record (EHR) database for more than 8.9 million veterans – the largest EHR database among U.S. health care systems, we started to address many important questions. We present our project and share our experience, with emphasis on the study design, modelling and analysis, and accompanying challenges such as how to identify the onset time of CKD and issues relate to irregular and informative visits and confounding.

EO077 Room S1A01 METHODS FOR FUNCTIONAL DATA ANALYSIS Chair: Ci-Ren Jiang

E0344: **Functional clustering and missing value imputation of longitudinal data**
Presenter: **Pai-Ling Li**, Tamkang University, Taiwan
A functional data approach is proposed for clustering and missing value imputation for incomplete longitudinal data. We adopt the notion of subspace-projected functional data clustering that each observed trajectory is viewed as a realization of a random function and is drawn from a mixture of stochastic processes, where each subprocess represents a cluster with a cluster-specific mean function and covariance function. The proposed algorithm comprises the probabilistic functional clustering (PFC) and the missing value imputation based on clustering results obtained from the PFC. The performance of the proposed method is demonstrated through a data example.

E0382: **Partial separability and graphical models for multivariate functional data**
Presenter: **Alexander Petersen**, University of California Santa Barbara, United States
Co-authors: Sang-Yun Oh, Javier Zapata
Graphical models are a ubiquitous tool for identifying dependencies among components of high-dimensional multivariate data. Recently, these tools have been extended to estimate dependencies between components of multivariate functional data by applying multivariate methods to the coefficients of truncated basis expansions. A key difficulty compared to multivariate data is that the covariance operator is compact, and thus not invertible. We present two important developments in this area for multivariate Gaussian processes. The first is to identify sufficient conditions under which absence of an edge in the conditional independence graph can be shown to correspond with zeros in a suitable inverse covariance operator. We then propose a new notion of partial separability as a useful tool for simplifying estimation, and show that the estimators are robust to certain types of model misspecification. Finally, we will demonstrate the empirical findings of our method through simulation and analysis of functional brain connectivity during a motor task.

E0529: **Inference on level sets in functional linear regression**
Presenter: **Masaaki Imaizumi**, The Institute of Statistical Mathematics, Japan
An inference method on active domains of functional data is developed via a functional linear regression and a principal component analysis (PCA) based estimator. In a linear regression model with a functional covariate and a scalar response variable, an active domain of a functional covariate is defined as a subset of a domain on which a functional data has a positive effect on outputs. Based on a functional linear regression model, an active domain is regarded as a level set of a slope function of the regression model. We propose an estimator for an active set by combining the PCA-based estimator and a kernel convolution approach. Also, we provide a multiplier bootstrap method for confidence analysis for an active set based on the high-dimensional Gaussian approximation technique. Our confidence analysis is shown to be valid asymptotically with ordinal conditions for a PCA-based estimator. We also propose a practical selection method for hyperparameters such as a cut-off level for basis functions and kernel width. The experimental analysis supports the validity of our method.

E0551: **Identifying essence codings and effects in functional linear models**
Presenter: **Shao-Wei Cheng**, National Tsing Hua University, Taiwan
The functional linear model $Y(t) = \beta_0(t) + \boldsymbol{X}\boldsymbol{\beta}(t) + \varepsilon(t)$ is considered. We assume that the vector of coefficient functions $\boldsymbol{\beta}(t)$ is a linear combination of some unknown essence codings $\boldsymbol{\Phi}(t)$. That is, there exists a matrix of parameters $\boldsymbol{\Gamma}$ such that $\boldsymbol{\beta}(t) = \boldsymbol{\Gamma}\boldsymbol{\Phi}(t)$. The parameters in $\boldsymbol{\Gamma}$ are called essence effects. We are interested in identifying meaningful and important essence codings and effects. Because the equation $\boldsymbol{\beta}(t) = \boldsymbol{\Gamma}\boldsymbol{\Phi}(t)$ has infinitely many solutions of $\boldsymbol{\Gamma}$ and $\boldsymbol{\Phi}(t)$, we propose a criterion for this equation to be uniquely defined, and therefore $\boldsymbol{\Gamma}$ and $\boldsymbol{\Phi}(t)$ become identifiable. This criterion treats essence codings as projection directions. By projecting $Y(t)$ onto every essence codings, the functional linear model is transformed into a uni-variate linear model. We obtain a set of orthogonal estimators $\hat{\boldsymbol{\Phi}}(t)$ of essence codings by sequentially maximizing the amount of variation in the response of this uni-variate linear model explained by the model matrix $\boldsymbol{X}$. For the analysis about the essence effects in $\boldsymbol{\Gamma}$, including estimation and testing, we suggest performing them by conditioning on $\hat{\boldsymbol{\Phi}}(t)$. Finally, the methods developed are applied on some functional data of wafer thickness to estimate essence codings and identify important essence effects.

EO127 Room AT241 NEW DEVELOPMENT IN DESIGN OF EXPERIMENTS Chair: Ming-Chung Chang

E0374: **Exploiting variance reduction potential in local Gaussian process search**
Presenter: **Chih-Li Sung**, Michigan State University, United States
Gaussian process models are commonly used as emulators for computer experiments. However, developing a Gaussian process emulator can be computationally prohibitive when the number of experimental samples is even moderately large. Local Gaussian process approximation was proposed as an accurate and computationally feasible emulation alternative. Constructing local sub-designs specific to predictions at a particular location of interest remains a substantial computational bottleneck to the technique. Two computationally efficient neighborhood search limiting techniques are proposed, a maximum distance method and a feature approximation method. Two examples demonstrate that the proposed methods indeed save substantial computation while retaining emulation accuracy.

E0285: **Detection of location and dispersion effects from partially replicated two-level factorial designs**
Presenter: **Shin-Fu Tsai**, National Taiwan University, Taiwan
Co-authors: Chen-Tuo Liao
Identifying active location and dispersion effects is an important issue at the early stage of a quality improvement process. After understanding the various impacts of factorial effects on the system response, a quality engineer can improve the system performance by adjusting the levels of identified factors. We will introduce a new testing procedure for screening active location effects from partially replicated two-level factorial designs. In addition, a two-stage procedure will be introduced for integrating the analyses of location and dispersion effects. Some numerical examples will be presented for illustrating the proposed method.

E0319: **A systematic construction of cost-efficient designs for order-of-addition experiments**
Presenter: **Jing-Wen Huang**, National Tsing-Hua University, institute of statistics, Taiwan
Co-authors: Frederick Kin Hing Phoa, Yuan-Lung Lin
An order-of-addition (OofA) experiment aims at investigating how the order of factor inputs affects the experimental response,which is recently of great interest among practitioners in clinical trials and industrial processes. Although the initial framework was established for more than 70 years, recent studies in the design construction of OofA experiments focused on their properties of algebraic optimality rather than cost-efficiency. The latter is more practical in the sense that some experiments, like cancer treatments, may not easily have adequate number of observations. We propose a systematic construction method for designs in OofA experiments from cost-efficient perspective. In specific, our designs take the effect of two successive treatments into consideration. To be cost-efficient, each pair of level settings from two different factors in our design matrix appears exactly once. Compared to recent studies in OofA experiments, our designs not only handle experiments of one-level factors (i.e. all

factors are mandatorily considered), but also factors of two or more levels, so practitioners may insert placebo or choose different dose when our designs are used in an OofA experiment in clinical trials for example.

E0773: **Spatial modeling of ground-level PM2.5 in Taiwan based on two types of data**
Presenter: **Chi-Wei Lai**, Academia Sinica, Taiwan
Co-authors: Hsin-Cheng Huang, ShengLi Tzeng
There are two systems to monitor fine particulate matter (PM2.5) in Taiwan. One consists of 77 monitoring stations of the Environmental Protection Administration, which provides high-quality measurements. The other one involves a large number of low-cost internet-of-things devices called AirBoxes, which produce less precise measurements but with much broader coverage. We propose a spatial model to obtain spatial prediction at any location in Taiwan by combining these two types of data. In addition, we develop a Shiny application that automatically identifies unusual measurements and shows the current PM2.5 concentration map with uncertainty quantification based on the proposed method.

EO061 Room AT242 RECENT DEVELOPMENTS IN STATISTICAL ANALYSIS FOR SURVIVAL DATA Chair: Chyong-Mei Chen

E0280: **Statistical analysis of hierarchical clustered data**
Presenter: **Weijing Wang**, National Chiao Tung U., Taiwan
A new model is proposed for right-censored survival data with multilevel clustering based on the hierarchical Kendall copula model with Archimedean clusters. This model easily accommodates clusters of unequal size and multiple clustering levels, without any structural conditions on the parameters or on the copulas used at various levels of the hierarchy. A step-wise estimation procedure is proposed and shown to yield consistent and asymptotically Gaussian estimates under mild regularity conditions. The model fitting is based on multiple imputation, given that the censoring rate increases with the level of the hierarchy. To check the model assumption of Archimedean dependence, a goodness-of test is developed. The finite-sample performance of the proposed estimators and of the goodness-of-fit test is investigated through simulations. The new model is applied to data from the study of chronic granulomatous disease.

E0286: **A class of general pretest estimators for the normal means**
Presenter: **Jia-Han Shih**, National Central University, Taiwan
Co-authors: Yoshihiko Konno, Genso-Yuan Tsung Watanabe-Chang, Takeshi Emura
For estimating a large number of mean parameters, univariate analyses remain the most popular approach in real applications due to its simplicity. Such analyses typically perform some preliminary tests (e.g. t-tests) to reduce the number of variables and impose some sparsity assumptions to shrink the estimates. To handle these tasks simultaneously, we propose a class of general pretest estimators that include many existing pretest, shrinkage, Bayes, and empirical Bayes estimators as special cases. We adopt the idea of randomized tests to construct a class of general pretest estimators, where the randomization probability is related to a shrinkage parameter. Theoretical properties of the proposed pretest estimator such as the exact distribution, bias, and mean squared error are derived. Our new expressions for the bias and mean squared error are simpler and more straightforward than the existing ones. We illustrate the use of the proposed estimator through the analysis of high-dimensional gene-expressions.

E0406: **Causal mediation of semi-competing risks**
Presenter: **Yen-Tsung Huang**, Academia Sinica, Taiwan
The semi-competing risk problem arises when one is interested in the effect of an exposure or treatment on both intermediate (e.g., having cancer) and primary events (e.g., death) where the intermediate event may be censored by the primary event, but not vice versa. Here we propose a nonparametric approach casting the semi-competing risks problem in the framework of causal mediation modeling. We set up a mediation model with the intermediate and primary events, respectively as the mediator and the outcome, and define indirect effect (IE) as the effect of the exposure on the primary event mediated by the intermediate event and direct effect (DE) as that not mediated by the intermediate event. A Nelson-Aalen type of estimator with time-varying weights is proposed for direct and indirect effects where the counting process at time t of the primary event $N_{2n_1}(t)$ and its compensator $A_{n_1}(t)$ are both defined conditional on the status of the intermediated event right before t, $N_1(t^-) = n_1$. We show that $N_{2n_1}(t) - A_{n_1}(t)$ is a zero-mean martingale. Based on this, we further establish the asymptotic unbiasedness, consistency and asymptotic normality for the proposed estimators. Numerical studies including simulation and data application are presented to illustrate the finite sample performance and utility of the proposed method.

E0466: **Confidence interval for the difference between two median survival times with semiparametric transformation models**
Presenter: **Yu-Mei Chang**, Tunghai University, Taiwan
In medical studies, the focus is usually on comparing the treatment effects of the drug according to the difference of two median survival times. We consider the problem of constructing conditional confidence interval for the difference of two median survival times given the covariates under a general class of the semiparametric transformation models with right-censored data. We propose two methods for constructing the conditional confidence intervals. One is based on the estimating equations (EE) estimator and the other on the nonparametric maximum likelihood estimator. Simulation results indicate that both methods provide satisfactory coverages for finite sample. We illustrate the proposed method using a real data set in a two-arm non-small cell lung cancer study.

EO330 Room AT335 BAYESIAN ANALYSIS AND ITS APPLICATIONS Chair: Charlotte Wang

E0506: **A Bayesian method to prioritizing candidate pathways association and gene ranking**
Presenter: **Shu-Ju Lin**, National Taiwan University, Taiwan
Co-authors: Tzu-Pin Lu, Chuhsing Kate Hsiao
Cancer is an important topic of global concern. Some cancer are closely related to genetic aberrations. In order to reduce costs, providing a prioritized list may help to find key genes. Currently, methods for screening genes associated with diseases are roughly classified into three types, such as the single marker test, gene set analysis methods, and pathway analysis, to provide candidate genes or candidate gene sets. However, due to the large number of bio markers, scientists need to face the issue of multiple testing with single marker test. Gene sets found may be a good indicator to cluster, but it is difficult to explain in biology. To overcome the limitations, we consider simultaneously several competing pathways, incorporate the relationship between pathways, and account for the relationship between genes under pathway information. In the simulation, our method controls the type I error well and correctly find true key genes. This novel method identifies the primary pathway of breast cancer as the Jak-STAT signaling pathway, and further identifies 37 key genes in this pathway. In the glioblastoma multiforme study, this method identifies Long-term potentiation as the primary pathway, and from which four key genes are identified.

E0567: **Use Bayesian conditional logistic regression to build up a localized medical-meteorological index**
Presenter: **Charlotte Wang**, Tamkang University, Taiwan
The environmental factors like meteorological factors and air pollutants have been recognized as important factors for human health, where mortality and morbidity of certain diseases may be related to the abrupt climate change or the air pollutant concentration. We use the National Health Insurance database and define the people aged 51-90 years who were free from cerebrovascular disease (ICD9: 430-438) or ischemic heart disease

(ICD9: 410-414) in 1996-2002 as the susceptible group. We then adopted the case-crossover study design and used a Bayesian conditional logistic regression to predict personal risks for suffering cerebrovascular diseases or ischemic heart diseases via environmental factors. Based on the predicted odds ratios, we defined the interval of the alert for the disease risks and evaluate the performance of the interval of the alert for the disease risks. We also explored the association between meteorological factors and two diseases in six areas in Taiwan. The results show that people living in different areas of Taiwan have different risk levels of two diseases and the intervals of the alert for the disease risks vary in six areas. In addition, health risk index, provided by this personal risk prediction model, can be a reference for weather bureaus to issue health warnings in the future. With the early warnings, the susceptible group will be able to prevent from suffering the diseases when meteorological conditions change.

E0593: **Evaluation of outlier detection algorithms in linear regression for temperature validation**
Presenter: **Mei-Hsien Lee**, University of Taipei, Taiwan
Co-authors: Yu-Chung Wei

Outlier detection is an important part of data quality control. Meteorological data verification has a significant impact on the future accurate construction of forecasting systems and other related industries. Erroneous observations are detected through the comparison with estimated references. Linear regression model is generally adopted to construct the relationship between observations and references. And then the potential false data points regarded as outliers are identified. Outlier detection methods via linear regression model are evaluated in temperature data from severs instrument stations. For frequentist approaches, studentized residuals are the easy way to detect the erroneous observations. DFFITS and Cooks distance are inappropriate because the data points resulted from extreme climatic rather than false observations are detected. Moreover, Bayesian predictive discordancy test and extreme posterior probabilities of random error with conjugate prior involved station specific information make identify outliers genuinely from the former meteorological dataset. An easy understanding of the statisticians on the application of meteorological verification is provided, as well as a reference for the selection of appropriate statistical models for the calibration of meteorological data in the meteorological field.

E0832: **A Bayesian sparse latent factor model for identification of cancer subgroups with data integration**
Presenter: **Dongjun Chung**, Medical University of South Carolina, United States
Co-authors: Zequn Sun, Brian Neelon

Identification of cancer subgroups is of critical importance for the development of precise therapeutic strategies for various types of cancer. The Cancer Genome Atlas (TCGA) have generated tremendous amount of high throughput genomic data, which profiles somatic mutation, copy number alteration, DNA methylation, gene expression for each patient. This large-scale cancer genomic data provides unprecedented opportunity to investigate cancer subgroups using integrative approaches based on multiple types of genomic data. We will discuss our recent work on a Bayesian sparse latent factor model for simultaneous identification of cancer subgroups (clustering) and key molecular features (variable selection), based on a joint analysis of continuous, binary, and count data. In addition, by utilizing pathway (variable group) information, this approach does not only improve accuracy and robustness in identification of cancer subgroups and key molecular features, but also facilitates biological understanding of novel findings generated with this approach. Finally, in order to facilitate efficient posterior sampling, a heavy-tailed prior is specified for continuous data while alternative Gibbs samplers are proposed based on Polya-Gamma mixtures of Normal densities for binary and count data. We will illustrate the proposed statistical model with simulation studies and its application to the TCGA data.

EO179 Room AT337 STATISTICAL INNOVATIONS IN PSYCHOMETRICS Chair: Chia-Yi Chiu

E0270: **Sufficient and necessary conditions for the identifiability of the Q-matrix**
Presenter: **Gongjun Xu**, University of Michigan, United States

Restricted latent class models (RLCMs) have recently gained prominence in educational assessment, psychiatric evaluation, and medical diagnosis. Different from conventional latent class models, restrictions on RLCM model parameters are imposed by a design matrix to respect practitioners' scientific assumptions. The design matrix, called the Q-matrix in cognitive diagnosis literature, is usually constructed by practitioners and domain experts, yet it is subjective and could be misspecified. To address this problem, researchers have proposed to estimate the design Q-matrix from the data. On the other hand, the fundamental learnability issue of the Q-matrix and model parameters remains underexplored and existing studies often impose stronger than needed or even impractical conditions. Sufficient and necessary conditions are proposed for the joint identifiability of the Q-matrix and RLCM model parameters. The developed identifiability conditions only depend on the design matrix and therefore is easy to verify in practice.

E0405: **Estimation accuracy in analyzing multilevel data with extremely unbalanced design and highly correlated structure**
Presenter: **Hsiu-Ting Yu**, National Chengchi University, Taiwan

Hierarchical Linear Models (HLM) have been widely used as the data analytical methods in psychological research to deal with the dependency naturally existed in data with a multilevel structure. However, several important methodological issues in HLMs are still lack of systematic and comprehensive investigations. Two structural properties in multilevel data possibly found in empirical research are studied: extremely unbalanced and highly correlated data structure. Systematic Monte Carlo simulation studies are conducted to examine the accuracy of parameter estimates in fixed- and random-effects. Factors examined include the number of groups, the mean group-sizes, patterns of group-sizes, and degrees of dependency in data. Results suggest that the unbalanced data structure compensate the inaccuracy in the estimates of fixed-effects parameters under the same total sample size. The unbalanced data structure also has more impact on the stability of estimates for random-effects parameters. Moreover, the number of groups affects the accuracy of parameter estimation more than the sizes of the group. Higher degrees of data correlation also lead to more underestimation of model parameters. The effects of extremely small samples are also analyzed and compared. Complete findings and possible implications will be reported and discussed.

E0564: **Multidimensional computerized adaptive testing for non-compensatory test structure**
Presenter: **Chia-Ling Hsu**, The Education University of Hong Kong, Hong Kong
Co-authors: Ming Ming Chiu

Current multidimensional computerized adaptive testing (MCAT) is limited only to linked multiple abilities that can compensate for one another rather than non-compensatory ones. In recognition of the usefulness of MCAT and the complications associated with non-compensatory test structure, we propose extending this model to other types of abilities (non-compensatory ones) and evaluating their performance. Three popular item selection methods are used and compared, namely, the Fisher information method, the mutual information method, and the Kullback-Leibler information method. The simulation results showed that the Fisher information and mutual information methods performed similarly, and both outperformed the Kullback-Leibler information method. Furthermore, it was found that the more stringent the termination criterion and the higher the correlation between the latent traits, the higher the resulting measurement precision and test reliability. Test reliability was very similar across the dimensions, regardless of the correlation between the latent traits and termination criterion. Generally, the difficulties of the administered items were found to be at a lower level than the examinees abilities, which shed light on item bank construction for non-compensatory items.

E0612: **Attribute hierarchy models in cognitive diagnosis: Identifiability of the latent attribute space and the Q-matrix**
Presenter: **Hans Friedrich Koehn**, University of Illinois, Urbana-Champaign, United States

Educational researchers have argued that a realistic view of the role of attributes in cognitively diagnostic modeling should account for the possibility that attributes are not isolated entities but interdependent in their effect on test performance. ("Attributes" is a collective term in cognitive diagnosis for any cognitive characteristic required to perform tasks.) Different approaches to modeling possible attribute interdependency have been discussed in the literature; among them the proposition to impose a hierarchical structure so that mastery of certain attributes is a prerequisite of mastering one or more other attributes. A hierarchical organization of attributes constrains the latent attribute space such that several proficiency classes, as they exist if attributes are not hierarchically organized, are no longer defined because the corresponding attribute combinations cannot occur with the given attribute hierarchy. Hence, the identification of the latent attribute space is often difficult–especially, if the number of attributes is large. As an additional complication, constructing a complete Q-matrix may not at all be straightforward if the attributes underlying the test items are supposed to have a hierarchical structure. A framework based on lattice theory is proposed for examining the conditions of identifiability of the latent space and of completeness of the Q-matrix if attributes are hierarchically organized.

EO334 Room U301 STRUCTURAL INSTABILITIES IN HIGH-DIMENSIONAL DATA I Chair: Matus Maciak

E0439: **Goodness-of-fit test for innovation copula in multivariate nonparametric time series**
Presenter: **Sarka Hudecova**, Charles University, Prague, Czech Republic
Co-authors: Natalie Neumeyer, Marek Omelka

Copula-based models have become popular for modelling multivariate econometric time series. Within these applications, the multivariate models are usually built from univariate models via a copula function and the conditional version of Sklar's theorem. The whole model has three separate components: conditional model (conditional mean and conditional variance), marginal distributions of innovations, and the innovation copula. We consider nonparametric estimation of the conditional model and the marginal distributions of the innovations. For the estimation of the innovation copula, both nonparametric and parametric estimators based on the estimated residuals are considered. We show that under some regularity assumptions, these copula estimators are asymptotically equivalent to estimators that would be based on the unobserved innovations, i.e. the asymptotic distribution is not affected by the necessary pre-estimation of the mean and variance functions. Furthermore, we propose a goodness-of-fit test for correct specification of the innovation copula, which is based on a comparison between the parametric and nonparametric copula estimator.

E0451: **Detection of changes in panel data models with stationary regressors**
Presenter: **Charl Pretorius**, Charles University, Czech Republic
Co-authors: Marie Huskova

A panel regression model with cross-sectional dimension N is considered. The aim is to test, based on T observations, whether the N intercepts in the model remain unchanged throughout the observation period. The test procedure involves the use of a CUSUM-type statistic derived from a pseudo-likelihood argument. We present asymptotic results of the test statistic in the case where both N and T are allowed to become large. The asymptotic results are valid under strong mixing and stationarity assumptions on the error and regressor sequences. Monte Carlo results will be presented that indicate that the tests work in the case of small to moderate sample sizes. The talk ends with an illustrative application of the procedure to financial data.

E0479: **Sparse regression and structure hunting in the periodogram analysis of unequally spaced time series**
Presenter: **Ivan Mizera**, University of Alberta, Canada
Co-authors: Li Zhang

The periodogram methodology is revised along the classical lines, with special attention to (substantially) unequally spaced time series. On the basis of this revision, we propose modifications along the lines of various techniques of sparse regression. Among those figure prominently l1/l2 regularization solved via second-order convex programming; but we also consider possible ramifications, in particular in view of possible theoretical objections to the application of l1/l2 regularization in this specific setting. We pay particular attention to the choice of tuning parameters determining the correct amount of regularization leading to the appropriate model selection: we try several recent structure-oriented approaches (like stability selection, for instance), and propose few others.

E0266: **Nuisance parameters free changepoint detection in non-stationary series**
Presenter: **Michal Pesta**, Charles University, Faculty of Mathematics and Physics, Czech Republic
Co-authors: Martin Wendler

Many changepoint detection procedures rely on the estimation of nuisance parameters (like long-run variance). If a change has occurred, estimators might be biased and data-adaptive rules for the choice of tuning parameters might not work as expected. If the data is not stationary, this becomes more challenging. The aim is to present two changepoint tests, which involve neither nuisance nor tuning parameters. This is achieved by combing self-normalization and wild bootstrap. We investigate the asymptotic behavior and show the consistency of the bootstrap under the hypothesis as well as under the alternative, assuming mild conditions on the weak dependence of the time series. As a by-product, a changepoint estimator is introduced and its consistency is proved. The results are illustrated through a simulation study. The new completely data-driven tests are applied to a real data example from finance.

EO069 Room U302 RECENT ADVANCES IN ECONOMETRICS AND FINANCIAL STATISTICS Chair: Li-Hsien Sun

E0338: **A nonparametric Bayesian approach to simultaneous subject and cell heterogeneity discovery for single cell RNA-seq data**
Presenter: **Xiangyu Luo**, Renmin University of China, China
Co-authors: Qiuyu Wu

The advent of the single cell sequencing era opens new avenues for the personalized treatment. The first but important step is discovering the subject heterogeneity at the single cell resolution. We address the two-level-clustering problem of simultaneous subject subgroup discovery (subject level) and cell type detection (cell level) based on the single cell RNA sequencing (scRNA-seq) data from multiple subjects. However, current approaches either cluster cells without considering the subject heterogeneity or group subjects not using the single cell information. We develop a solid nonparametric Bayesian model SCSC (Subject and Cell clustering for Single-Cell data) to achieve subject and cell grouping at the same time without pre-specifying the subject subgroup number or the cell type number. An efficient blocked Gibbs sampler is then proposed for the posterior inference. The simulation study and the real application demonstrate the good performance of our model.

E0397: **A simple and efficient estimation of the average treatment effect in the presence of unmeasured confounders**
Presenter: **Zheng Zhang**, Renmin University of China, China
Co-authors: Chunrong Ai, Lukang Huang

Identification and estimation of the average treatment effect has been studied previously when some confounders are unmeasured. Under an identification condition, it has been shown that the semiparametric efficient influence function depends on five unknown functionals. It has been

proposed to parameterize all functionals and estimate the average treatment effect from the efficient influence function by replacing the unknown functionals with estimated functionals. The proposed estimator has been established to be consistent when certain functionals are correctly specified and attains the semi-parametric efficiency bound when all functionals are correctly specified. In applications, it is likely that those functionals could all be misspecified. Consequently, the estimator could be inconsistent or consistent but not efficient. An alternative estimator is proposed which does not require parameterization of any of the functionals. We establish that the proposed estimator is always consistent and always attains the semiparametric efficiency bound. A simple and intuitive estimator of the asymptotic variance is presented, and a small scale simulation study reveals that the proposed estimation outperforms the existing alternatives in finite samples.

E0476: **Some results on a portfolio optimization problem with delays**
Presenter: **Sheng-Jhih Wu**, Soochow University, China

A Merton type investment-consumption problem is considered in which the dynamics of the risky asset depends on its own past. The problem is formulated as a stochastic control problem with certain types of delays in the state system. The objective is to choose trading and consumption strategies so that the expected discounted utility is maximized. We first derive solutions to the associated Hamilton-Jacobi-Bellman equations under some frequently used utility functions in the literature. We then establish verification theorems and derive the optimal trading and consumption strategies.

E0640: **Systemic risk measures: SRISK and CoVaR under dynamic volatility matrix models**
Presenter: **Chuan-Hsiang Han**, National Tsing-Hua University, Taiwan

Systemic risk is an important measure for financial stability in the situation of financial crisis. A two-step procedure is proposed for systemic risk estimation under the stochastic volatility matrix models. The first step utilizes Fourier transform method for dynamic volatility matrix estimation. The second step conducts a rigorous proof of asymptotic optimal importance sampling estimators for rare event simulations. For empirical results, we find that the systemic risk can be useful to measure the stability of financial system. The ranking of systemically important financial institutions (SIFIs) identified by common systemic risk measures such as SRISK and CoVaR are compared in the framework of Spearman correlation. Systemic risks of markets in the US, China and Taiwan during the financial crisis period between 2008 and 2010 are presented.

EO053 Room U414 NONPARAMETRIC AND VARIABLE SELECTION METHODS FOR MODERN COMPLEX DATA Chair: Guanyu Hu

E0456: **Kernel density matching and its application for accurate alignment of single-cell RNA-seq samples**
Presenter: **Mengjie Chen**, University of Chicago, United States

With technologies improved dramatically over recent years, single cell RNA-seq (scRNA-seq) has been transformative in studies of gene regulation, cellular differentiation, and cellular diversity. As the number of scRNA-seq datasets increases, a major challenge will be the standardization of measurements from multiple different scRNA-seq experiments enabling integrative and comparative analyses. However, scRNA-seq data can be confounded by severe batch effects and technical artifact. In addition, scRNA-seq experiments generally capture multiple cell-types with only partial overlaps across experiments making comparison and integration particularly challenging. To overcome these problems, we have developed a method, dmatch, which can both remove unwanted technical variation and assign the same cell(s) from one scRNA-seq dataset to their corresponding cell(s) in another dataset. By design, our approach can overcome compositional heterogeneity and partial overlap of cell types in scRNA-seq data. We further show that our method can align scRNA-seq data accurately across tissues biopsies.

E0617: **Feature selection with survival outcome data**
Presenter: **Hyokyoung Grace Hong**, Michigan State University, United States

Detecting biomarkers that are relevant to patients' survival outcome is crucial for precision medicine. Dimension reduction is key in the process. Although regularization methods have been used for dimension reduction, they cannot handle a large number of candidate biomarkers generated by modern bio-techniques. Variable screening, which has been widely used for handling exceedingly large numbers of variables, is however much underdeveloped for censored outcome data. A series of new feature screening procedures for survival data with ultrahigh dimensional covariates is introduced. These methods include conditional screening, integrated powered density (IPOD) screening, L_q-norm learning, and forward regression with partial likelihood. We will discuss the intuition behind and demonstrate their utilities through real data analyses.

E0671: **Optimal variance estimation under random design**
Presenter: **Chao Gao**, University of Chicago, United States

The problem of estimating the variance, or variance function, in nonparametric regression with random design is considered. We show that the minimax rate of estimation is different from that in a fixed design setting. To be specific, for variance, or variance function, estimation, the minimax rate is faster in a random design setting than that in a fixed design setting. This phenomenon is unique to the variance estimation problem, and is different from the mean function estimation problem where the minimax rates are the same. To achieve the minimax rate, we develop a U-statistic-based local polynomial estimator and a lower bound is carefully constructed based on a result on the maximum frequency of sparse multinomial distribution and the moment-matching techniques in the literature.

E0738: **Variance component testing and selection for a longitudinal microbiome study**
Presenter: **Jin Zhou**, University of Arizona, United States

High-throughput sequencing technology has enabled population-based studies of the role of the human microbiome in disease etiology and exposure response. Due to the high cost of sequencing technology such studies usually have limited sample sizes. We study the association of microbiome composition and clinical phenotypes by testing the nullity of variance components. When the null model has more than one variance parameters and sample sizes are limited, such as in longitudinal metagenomics studies, testing zero variance components remains an open challenge. We first introduce a series of efficient exact tests (score test, likelihood ratio test, and restricted likelihood ratio test) of testing zero variance components in presence of multiple variance components. The approach does not rely on the asymptotic theory, thus significantly boosts the power in small samples. Furthermore, to further conquer limited sample size and high dimensional features of metagenomics data, we introduce a variance component selection scheme with lasso penalization. We propose an minorization-maximization (MM) algorithm for the difficult optimization problem. Extensive simulations demonstrate the superiority of our methods vs existing methods. Finally, we apply our method to a longitudinal microbiome study of HIV infected patients.

EO302 Room U501 STATISTICAL MODELING: NEW METHODOLOGIES AND APPLICATIONS Chair: Wan-Lun Wang

E0170: **Statistical modeling for adaptive trait evolution in randomly evolving environment**
Presenter: **Dwueng-Chwuan Jhwueng**, Feng-Chia University, Taiwan
In past decades, Gaussian processes has been widely applied in studying trait evolution using phylogenetic comparative analysis. In particular, two members of Gaussian processes: Brownian motion(BM) and Ornstein-Uhlenbeck(OU) process, have been frequently used to describe continuous trait evolution. Under the assumption of adaptive evolution, several models have been created around OU process where the optimum of a single trait is influenced with predictor. Since in general the dynamics of rate of evolution of trait could adopt a pertinent process, we extend models of adaptive evolution by considering the rate of evolution following the Cox-Ingersoll-Ross(CIR) process. We provide a heuristic Monte Carlo simulation scheme to simulate trait along the phylogeny as a structure of dependence among specie. We add a framework to incorporate multiple regression with interaction between optimum of the trait and its potential predictors. Since the likelihood function of our models are intractable, we propose the use of Approximate Bayesian Computation(ABC) for parameter estimation and inference. Simulation as well as empirical study using the proposed models are also performed and carried out to validate our models and for practical application.

E0389: **A new approach for analyzing response style data**
Presenter: **Yu-Wei Chang**, Department of Statistics, Feng Chia University, Taiwan
In survey questionnaires studies, latent factors underlying Likert-type data are often the quantities of interest. Many empirical studies have shown that response style is an issue in survey questionnaires. For example, people from East Asia are more conservative than people from Latin America, and the former are less likely to answer "very agree" or "very disagree" in survey questionnaires, given that they have the same latent factors. Multiple-group factor analysis (M-FA) models are one of the common practice for taking the group difference into account so that we could have better estimates for the latent factors. However, the Likert-type data are treated as continuous in M-FA models. To better accommodate the ordinal categorical feature of the data, we suggest using Multiple-group categorical confirmatory factor analysis (MC-FA) models for the response style modeling. As other multiple-group analysis, we need one or some anchor item(s) for a fair comparison between groups. More specifically, before fitting a multiple-group model, we have to find one or some item(s) which functions the same between groups, and then parameters of the item(s) are restricted to be the same between groups. We provide a strategy for the anchor selection while MC-FA models are applied to the response style data. A justification of the procedure will also be given. Finally, the proposed procedure is applied to a real data set for illustration.

E0431: **Bregman divergence to generalize Bayesian influence measures for data analysis**
Presenter: **Mauricio Castro**, Pontificia Universidad Catolica de Chile, Chile
For existing Bayesian cross-validated measure of influence of each observation on the posterior distribution, a generalization using the Bregman Divergence (BD) is considered. We investigate various practically useful and desirable properties of these BD based measures to demonstrate the superiority of these measures compared to existing Bayesian measures of influence and Bayesian residual based diagnostics. We provide a practical and easily comprehensible method for calibrating these BD based measures. Also, we show how to compute our BD based measure via Monte Carlo Markov Chain (MCMC) samples from a single posterior based on the full data. Using a Bayesian Meta-analysis of clinical trials, we illustrate how our new measures of influence of observations have more useful practical roles for data analysis than popular Bayesian residual analysis tools.

E0358: **Parametrizing the Kepler exoplanet period-radius distribution with the bivariate normal inverse Gaussian distribution**
Presenter: **Wen-Liang Hung**, National Tsing Hua University, Taiwan
A simple and robust method is presented for obtaining a comprehensive understanding of the joint period and radius distribution in Kepler exoplanets. The proposed method is based on particle swarm optimization and bivariate Normal Inverse Gaussian distribution. Furthermore, in the construction of the probability density function, planet-host stars with the GK-type are selected. The injecting approach is also employed to solve the survey completeness of sample. The resulting occurrence rate of Earth analogs is 0.025 with a 95% bootstrap confidence interval between 0.023 and 0.032.

EO281 Room U502 STATISTICAL LEARNING IN GENOMIC APPLICATIONS Chair: Xinlei Wang

E0281: **Efficient coordinate descent algorithm for sparse precision matrix estimation via scaled Lasso**
Presenter: **Donghyeon Yu**, Inha University, Korea, South
Sparse precision matrix plays an important role in Gaussian graphical model due to the fact that a zero off-diagonal element denotes the conditional independence of two corresponding variables given others. In the Gaussian graphical model, lots of methods have been developed and their theoretical properties are given as well. Among them, the sparse precision matrix estimation via scaled lasso (SPMESL) has an attractive feature that automatically sets the penalty level to achieve the optimal convergence rate under the sparsity and the invertibility conditions, while other methods have to search for optimal tuning parameters. Yet, despite its advantage, the SPMESL is not widely used due to its expensive computational cost and the restricted assumptions. The coordinate descent (CD) algorithm for the SPMESL is considered, and it is shown to be numerically more efficient than the least angle regression (LARS). In addition, we develop a parallel CD algorithm using graphics processing units for scalability. Numerical study is also conducted to investigate the sensitivity of the SPMESL to the theoretical assumptions, and shows that the SPMESL has the smallest false discovery rate for all cases and the best performance in cases where the level of sparsity of columns is high.

E0322: **Bayesian data integration in Cancer genomics**
Presenter: **Francesco Stingo**, University of Florence, Italy
Identifying patient-specific prognostic biomarkers is of critical importance in developing personalized treatment for clinically and molecularly heterogeneous diseases such as cancer. We propose a novel regression framework, Bayesian hierarchical varying-sparsity regression models to select clinically relevant disease markers by integrating proteogenomic (proteomic+genomic) and clinical data. Our methods allow flexible modeling of protein-gene relationships as well as induces sparsity in both protein-gene and protein-survival relationships, to select genomically driven prognostic protein markers at the patient-level. We apply the proposed method to The Cancer Genome Atlas (TCGA) proteogenomic pan-cancer data and find several interesting prognostic proteins and pathways that are shared across multiple cancers and some that exclusively pertain to specific cancers.

E0499: **Relaxation rate of gene expression kinetics reveals the feedback sign of auto-regulatory gene network**
Presenter: **Min Chen**, University of Texas at Dallas, United States
Co-authors: Chen Jia, Hong Qian, Michael Zhang
The transient response to a stimulus and subsequent recovery to a steady state is a fundamental characteristics of a dynamic organism. We study the relaxation kinetics of auto-regulatory gene networks based on the Delbruck-Gillespie process description of single-cell stochastic gene expression. We report a novel relation between the rate of relaxation, characterized by spectral gap of the process, and the sign of feedback loop in the gene regulation. When a network has no feedback, the relaxation rate is precisely the degradation rate of the protein. We show that positive feedback always decreases the relaxation rate while negative feedback always increases it. Numerical simulations demonstrate that this relation provides

an effective method for the inference of gross feedback topology of the underlying gene regulatory network by using time-series data of gene expression from either single cells or cell populations.

E0628: **Bayesian jackknife empirical likelihood**
Presenter: **Yichen Cheng**, Georgia State University, United States
Empirical likelihood is a very powerful nonparametric tool that does not require any distributional assumptions. It has been previously shown that, if you replace the usual likelihood component in the Bayesian posterior likelihood with the empirical likelihood, then posterior inference is still valid when the functional of interest is a smooth function of the posterior mean. However, it is not clear whether similar conclusions can be obtained for parameters defined in terms of U-statistics. We propose the so-called Bayesian jackknife empirical likelihood, which replaces the likelihood component with the jackknife empirical likelihood. We show, both theoretically and empirically, the validity of the proposed method as a general tool for Bayesian inference. Empirical analysis shows the small sample performance of the proposed method is better than its frequentist counterpart. Analysis of a case-control study for pancreatic cancer is used to illustrate the new approach.

EO310 Room U517 ADVANCES IN ANALYSIS OF COMPLEX TIME SERIES DATA Chair: Raymond Ka Wai Wong

E0167: **Asymptotically constant risk estimator of the time-average variance constant**
Presenter: **Chun Yip Yau**, Chinese University of Hong Kong, Hong Kong
Co-authors: Kin Wai Chan
Estimation of the asymptotic variance of a time-average, which is known as the time-average variance constant (TAVC), or long run variance, is important for many statistical procedures involving dependent data. However, the estimation of TAVC is difficult as its performance relies heavily on the choice of a bandwidth parameter. Specifically, the optimal choices of bandwidth of all existing estimators depend on the TAVC itself and another unknown parameters which is very difficult to estimate. Thus, the optimal estimation of TAVC is not achievable. By introducing a novel concept of converging kernel, we develop a new class of TAVC estimators in which the optimal bandwidth is free of unknown parameters and hence can be computed easily. Moreover, we prove that the new estimator has a constant risk asymptotically, in contrast to the exploding risk in the existing estimators.

E0175: **Consistent estimation of high-dimensional factor models when the factor number is over-estimated**
Presenter: **Haeran Cho**, University of Bristol, United Kingdom
Co-authors: Matteo Barigozzi
A high-dimensional r-factor model for an n-dimensional vector time series is characterised by the presence of a large eigengap (increasing with n) between the r-th and the $(r+1)$-th largest eigenvalues of the covariance matrix. Consequently, Principal Component Analysis (PCA) is a popular estimation method for factor models and its consistency, when r is correctly estimated, is well-established in the literature. However, various factor number estimators often suffer from the lack of an obvious eigengap in empirical eigenvalues. We show that they tend to over-estimate the factor number in the presence of moderate correlations in the idiosyncratic (not factor-driven) components which, in turn, leads to non-negligible errors in the PCA estimators. To remedy this problem, we propose two new estimators based on capping or scaling the entries of the sample eigenvectors, which are less sensitive than the PCA estimator to the over-estimation of r without knowing the true factor number. We show both theoretically and empirically that the two estimators successfully controls for the over-estimation error, and demonstrate their good performance on macroeconomics and financial time series datasets.

E0457: **Segmentation of Japanese rescue dog behavior data: A Markov-switching model approach**
Presenter: **Rex Cheung**, San Francisco State University, United States
Co-authors: Ryunosuke Hamada
The use of Markov Switching Vector Autoregressive Model (MSVAR) is proposed to analyze Japanese search and rescue dog behavior. As search and rescue dogs have been a popular companion in many police operations in many countries, they are rarely included in any search operations in Japan, due to the lack of data evidence to show the efficiency of the rescue dogs. Therefore, experiments have been done by mounting sensors and video camera to capture the movement signals of the dog, and to analyze the activity and behavior during a search operation. The aim is to apply the MSVAR model to analyze the sensor signals, by dividing the observed signal into homogenous regions such that each region corresponds to a unique, non-overlapping activity of the dog (such as run, sniff, etc.). To encourage flexibility, a sparse-lag penalty will be imposed to perform automatic lag selection and sparsification of autoregressive coefficients.

E0495: **Elastic-band transform: A new way for multiscale method**
Presenter: **Hee-Seok Oh**, Seoul National University, Korea, South
Co-authors: Guebin Choi
A new multiscale transformation method is presented for the statistical analysis of one-dimensional data such as time series and functional data under the concept of the scale-space approach. The proposed method uses regular observations (eye scanning) with a range of different intervals. The results, termed 'elastic-band transform' can be considered as a collection of observations over various intervals (length of elastic-band) of viewing. It is motivated by a way that people look at an object such as a sequence of data repeatedly in order to overview a global structure of it as well as find some specific features of it. Some measures based on elastic-bands are considered for describing characteristics of data, and multiscale visualizations induced by the measures are developed for an understanding of data and detecting important structures of them. The proposed transform holds inherently some strengths for analyzing periodic signals because of its definition induced by a collection of regular observations; hence, statistical applications such as detection and signal extraction of periodic signals are studied.

Wednesday 26.06.2019 16:10 - 17:50 Parallel Session I – EcoSta2019

EI007 Room UB99(B1) ADVANCES IN FINITE MIXTURE MODELS **Chair: Geoffrey McLachlan**

E0160: **A unified tool for the root selection and the hypothesis testing for mixture models**
Presenter: **Weixin Yao**, UC Riverside, United States

The aim is to show how to apply goodness of fit (GOF) statistics to choose a consistent root for finite mixture models. Our new method inherits both the consistency properties of distance estimators and the efficiency of the MLE. The new method is simple to use and its computation can be easily done using existing R packages for mixture models. In addition, we will also introduce how to apply the GOF test statistics to perform the hypothesis testing and model selection for finite mixture models. The limiting distribution of test statistics is simulated based on a bootstrap method. It is demonstrated through extensive empirical studies that a simple application of GOF test statistics to finite mixture models can provide comparable or even superior hypothesis testing performance compared to some existing cutting-edge testing methods.

E0161: **A mixture cure model with multilevel frailties for analysing recurrent event data**
Presenter: **Shu-Kay Ng**, Griffith University, Australia
Co-authors: Richard Tawiah

In the field of medical and health sciences, researchers frequently encounter multivariate survival data consisting of multiple failure outcomes of (recurrent) events (e.g. tumour relapses) from multi-centre studies. Data collection in these studies often exhibits a multilevel structure, inducing strong methodological challenges in modelling intra-subject and between-subject correlation. The complexity in modelling increases further by the presence of long-term survivors who respond favourably to treatment and are thus insusceptible to tumour relapses (a characterised feature is the marginal survival being levelled off to non-zero probability). A new mixture frailty model is developed to analyse multilevel recurrent event data within the context of cure models, in which multivariate random effects with an AR(1) structure are used to impose serial dependence between gap times to recurrent events and another set of random effects are used to model main centre effects and treatment-by-centre interaction effects for uncured patients. The cure probability is modelled by a logistic mixed model with random (patient, centre, and treatment-by-centre) effects. Estimation inference is developed via an EM-type algorithm based on residual maximum likelihood (REML) through the generalised linear mixed model (GLMM) methodology. The method is illustrated using simulated data and a publicly-available dataset concerning a randomised multi-centre trial of rhDNase for treating cystic fibrosis.

E0162: **Skew distributions in model-based clustering**
Presenter: **Sharon Lee**, University of Adelaide, Australia

The past decade has seen increasing use of flexible distributions that can handle skewness in the data. The literature now offers a wide variety of non-normal and asymmetric distributions with different characterizations and properties. These distributions are motivated by and are suitable for different applications. The aim is to present a survey of some popular skew distributions adopted in the model-based clustering literature, including the class of skew symmetric distributions, the multiple scaled distributions, as well as those obtained by transformation. Their formulations, properties, and advantages will be briefly discussed. Some simulations will be presented to illustrate their abilities in modelling distinct types of skewness in the data.

EO073 Room S101 RECENT ADVANCES IN LEARNING THEORY AND APPLICATIONS **Chair: Qiang Wu**

E0282: **Exploring the power of source reliability in information integration**
Presenter: **Houping Xiao**, Robinson College of Business/GSU, United States
Co-authors: Shiyu Wang

In the era of Big Data, data entries, even describing the same objects or events, can come from a variety of sources. There are some sources that typically provide accurate information, but due to various reasons such as recording errors, device malfunction, background noise, or even intent to manipulate the data, some other sources may contain noisy or even erroneous information. Therefore, during information integration, it is critical to identify reliable sources that more often provide accurate information. Unfortunately, there is no oracle telling us which information source is more reliable a priori. Novel information integration methods are developed that incorporate the estimation of source reliability in both data-level and model-level information integration. In both works, we prove some nice properties of the proposed approaches via theoretical analysis and demonstrate their impact on some real applications such as indoor floorplan construction and crowdsourced question answering.

E0289: **Distributed regression learning with coefficient and partial coefficients regularization**
Presenter: **Hongwei Sun**, University of Jinan, China

Distributed regression learning with a coefficient regularization scheme in a reproducing kernel Hilbert space (RKHS) is studied. The algorithm randomly partitions the sample set z_i, $i = 1,2,...,N$ into m disjoint sample subsets of equal size, applies coefficient regularization scheme to each sample subset to produce an output function, and averages the individual output functions to get the final global estimator. We deduce the error bound in expectation in the L^2-metric and prove the asymptotic convergence for this distributed coefficient regularization learning. Satisfied learning rates are derived under a very mild regularity condition on the regression function, which reveals an interesting phenomenon that when $m < N^s$ and s is small enough, this distributed learning has the same convergence rate compared with the algorithm processing the whole data in one single machine. In order to reduce the complexity of algorithms, we also study a new distributed coefficient regularization scheme, which apply a partial coefficients regularization to each sample subset to produce an output function, and average the individual output functions to get the final global estimator. The error bound in the L^2-metric is deduced and the asymptotic convergence for this distributed learning with partial coefficients regularization is proved by integral operator technique. Satisfactory learning rates are then derived under a standard regularity condition on the regression function.

E0291: **Privacy friendly learning with empirical feature-based summary statistics**
Presenter: **Xin Guo**, The Hong Kong Polytechnic University, Hong Kong
Co-authors: Huihui Qin

Nowadays the extensive collecting and analyzing of data is stimulating widespread privacy concerns, and therefore is increasing tensions between the potential sources of data and researchers. Obviously, a privacy-friendly learning framework can help to ease the tension, and to boost data-related research. We propose a new algorithm of regression learning with empirical features, which uses only summary statistics instead of raw data. The selection of empirical features serves as a trade-off between prediction precision and the protection of privacy. Mathematical analysis of the convergence of the framework is provided, which covers also the scenario where data sets are collected from different sources respectively.

E0561: **Mathematical foundations of learning with information theoretic criteria**
Presenter: **Qiang Wu**, Middle Tennessee State University, United States
Co-authors: Yunlong Feng
Learning with information theoretic criteria, namely, the maximum correntropy criterion (MCC) and the minimum error entropy (MEE), has been shown successful in a variety of applications. It is particularly powerful to handle data contaminated by outlier or heavy tailed noises. To theoretically justify the effectiveness of these two information theoretic criteria, we studied the consistency of MCC and MEE based machine learning algorithms. We showed that, with appropriate parameter selection strategies, these algorithms can effectively learn both mean regression function and modal regression function. We also proved some no-free-lunch theorems which indicate that, in some scenarios, there must be some sacrifice of information contained in the data in order to achieve high prediction accuracy.

EO332 Room S102 STATISTICAL INNOVATIONS IN THE ANALYSIS OF MICROBIOME DATA Chair: Jin Zhou

E0634: **Strain-GeMS: Optimization subspecies identification from microbiome data based on accurate variant modeling**
Presenter: **Xinping Cui**, University of California, Riverside, United States
Subspecies identification is one of the most critical issue in microbiome studies, as it is directly related to the functions of the species as well as the whole microbial communities in response to the environmental stress and their feedbacks. However, identification of subspecies remains a challenge largely due to the small variation between different strains within the same species. Accurate identification of subspecies primarily rely on variant identification and categorization through microbiome data. However current SNP calling through microbiome data remain underdeveloped. We have proposed Strain-GeMS for subspecies identification from microbiome data, based on SNP calling with solid statistical model, as well as optimized subspecies identification procedure. Results on simulated, ab initio and in vivo datasets have shown that Strain-GeMS could always outperform other subspecies identification methods in terms of accuracy and coverage of the strains. With the rapidly increasing amount of microbiome samples, and the needs for subspecies identification, we believe that Strain-GeMS could become a key tool towards elucidating of subtle differences among subspecies in a microbial community.

E0455: **A novel normalization and differential abundance test framework for microbiome data**
Presenter: **Hongmei Jiang**, Northwestern University, United States
Microbial communities have been proved to have close relationship with many diseases. The identification of differentially abundant microbial species is clinically meaningful for finding disease-related pathogenic or probiotic bacteria. However, certain characteristics of microbiome data have hurdled the accuracy and effectiveness of differential abundance analysis. We develop a novel framework for differential abundance analysis on sparse high-dimensional marker gene microbiome data. The methodology relies on a network-based normalization technique and a two stage zero-inflated mixture count regression model (RioNorm2). Our novel network-based normalization method aims to find a group of relatively invariant species across samples and environments in order to construct size factors. It does not make any assumption on count distributions. Our testing approach can take into consideration under-sampling and over-dispersion with flexibility by separating microbiome species into different subgroups and model them separately. Through comprehensive simulation studies, the performance of our method is consistently powerful and robust across different settings with different sample sizes, library sizes and effect sizes. We also demonstrate the effectiveness of our novel framework using a published dataset of Metastatic Melanoma and find biological insights from the results.

E0608: **A novel statistical approach on trace evidence using human microbiome**
Presenter: **Lingling An**, University of Arizona, United States
Microbial forensics is a fast growing field. New research opportunities are emerging in the examination of trace evidence and human identification using microbiome. However, there is a lack of statistical and computational methods for analyzing the microbiome data for this purpose. The authors develop rigorous statistical methods and computational algorithms to estimate the probability of individuals presence at the scene of crime based on the microbial materials collected at the scene. Specifically, they can accurately identify sources or individuals from noisy metagenomic data and can even detect if there is any individual who was at the scene of a crime but is omitted from being considered as a suspect.

E0353: **Development of microbiome-based prediction models using co-informative information**
Presenter: **Michael Wu**, Fred Hutchinson Cancer Research Center, United States
The low practical and financial cost of collecting microbiome data has spurred interest in the development of microbiome-based models for predicting health outcomes. Many of these studies also collect additional types of genomic data that are co-informative with the microbiome. However, such data are expensive such that the objective is to use these data to facilitate development of a prediction model that only requires collecting microbiome data in the future. To do this, we propose a framework for summarizing the co-informative information within a kernel which can be used to modify the usual LDA loss function. We further consider the imposition of additional practical and biologically informed penalties. Simulations and data applications are used to illustrate the approach.

EO239 Room S104 HIGH-DIMENSIONAL STATISTICAL METHODS WITH APPLICATIONS Chair: Tongtong Wu

E0226: **Constrained maximum entropy models to select genotype interactions associated with censored failure times**
Presenter: **Qing Pan**, George Washington University, United States
A novel screening method targeting genotype interactions associated with disease risks is proposed. The proposed method extends the maximum entropy conditional probability model to address disease occurrences over time. Continuous occurrence times are grouped into intervals. The model estimates the conditional distribution over the disease occurrence intervals given individual genotypes by maximizing the corresponding entropy subject to constraints linking genotype interactions to time intervals. The EM algorithm is employed to handle observations with uncertainty, for which the disease occurrence is censored. Stepwise greedy search is proposed to screen a large number of candidate constraints. The minimum description length is employed to select the optimal set of constraints. Extensive simulations show that five or so quantile-dependent intervals are sufficient to categorize disease outcomes into different risk groups. Performance depends on sample size, number of genotypes, and minor allele frequencies. The proposed method outperforms the likelihood ratio test, Lasso, and a previous maximum entropy method with only binary(disease occurrence, non-occurrence) outcomes. Finally, a GWAS study for type 1 diabetes patients is used to illustrate our method. Novel one-genotype and two-genotype interactions associated with neuropathy are identified.

E0227: **Empirical frequency band analysis of nonstationary time series**
Presenter: **Scott Bruce**, George Mason University, United States
Co-authors: Cheng Yong Tang, Martica Hall, Robert Krafty
The time-varying power spectrum of a time series process is a bivariate function that quantifies the magnitude of oscillations at different frequencies and times. To obtain low-dimensional, parsimonious measures from this functional parameter, applied researchers consider collapsed measures of power within local bands that partition the frequency space. Frequency bands commonly used in the scientific literature were historically derived from manual inspection and are not guaranteed to be optimal or justified for adequately summarizing information from a given time series process under current study. There is a dearth of methods for empirically constructing statistically optimal bands for a given signal. The goal is to discuss

a standardized, unifying approach for deriving and analyzing customized frequency bands. A consistent, frequency-domain, iterative cumulative sum based scanning procedure is formulated to identify frequency bands that best preserve nonstationary information. A formal hypothesis testing procedure is also dedicatedly developed to test which, if any, frequency bands remain stationary. The proposed method is used to analyze heart rate variability of a patient during sleep and uncovers a refined partition of frequency bands that best summarize the time-varying power spectrum.

E0228: **Empirical likelihood in high dimensionality with application to TAAG**
Presenter: **Tongtong Wu**, University of Rochester, United States
Co-authors: Cheng Yong Tang, Jinyuan Chang
Empirical likelihood (EL) methods, a nonparametric counterpart of likelihood methods, are appealing and effective, especially in conjunction with estimating equations through which useful information can be adaptively and flexibly incorporated. It is also known in the literature that EL approaches encounter substantial difficulties when dealing with problems having high-dimensional model parameters and estimating equations. To answer the challenges, we propose a new penalized EL by applying two penalty functions respectively on the model parameters and the associated Lagrange multipliers in the optimizations of EL. Allowing both the dimensionalities of model parameters and estimating equations growing exponentially with the sample size, the estimator from our new penalized EL is sparse and consistent with asymptotically normally distributed nonzero components. We also design a statistical inference procedure for low-dimensional components of the model parameters, by linearly mapping the original estimating equations to a low-dimensional space. Nest coordinate descent algorithm, along with additional steps, can be used to tackle the well-known difficulties in EL computation, especially in high-dimensional settings. Simulation studies provide numeric evidence that the new penalized EL works well in high dimensionality. This method was applied to the TAAG data to examine multi-level factors related to the MVPA levels over time for girls from adolescence into young adulthood.

E0230: **Adaptive Bayesian factor spectral analysis of high–dimensional nonstationary time series**
Presenter: **Zeda Li**, Baruch College, City University of New York, United States
Co-authors: Ori Rosen, Robert Krafty
A frequency-domain factor model is proposed to spectral analysis of high-dimensional nonstationary time series. The model provides a general framework for estimating both real- and complex-valued spectra by allowing a time series acts simultaneously and propagates in a lagged fashion. Real and imaginary parts of the factor loading matrix are modeled independently by tensor products of penalized splines and multiplicative gamma process shrinkage priors, which allows infinitely many factors with the loadings increasingly shrunk towards a constant function as the column index increases. Formulated in a fully Bayesian framework, a conditional Whittle likelihood-based Gibbs sampler is developed for efficient model fitting. By using stochastic approximation Monte Carlo (SAMC) and partitioning a time series into an unknown number of approximately stationary segments, the approach automatically and adaptively estimates the power spectrum of both stationary and nonstationary high-dimensional time series.

E0837: **Matching methods for obtaining survival functions to estimate the effect of a time-dependent treatment**
Presenter: **Yun Li**, University of Michigan, United States
In observational studies of survival time featuring a binary time-dependent treatment, the hazard ratio (an instantaneous measure) is often used to represent the treatment effect. However, investigators are often more interested in the difference in survival functions. We propose semiparametric methods to estimate the causal effect of treatment among the treated with respect to survival probability. The objective is to compare post-treatment survival with the survival function that would have been observed in the absence of treatment.For each patient, we compute a prognostic score (based on the pre-treatment death hazard) and a propensity score(based on the treatment hazard). Each treated patient is then matched with an alive, uncensored and not-yet-treated patient withsimilar prognostic and/or propensity scores. The experience of each treated and matched patient is weighted using a variant of Inverse Probability of Censoring Weighting to account for the impact of censoring. We propose estimators of the treatment-specific survival functions (and their difference), computed through weighted Nelson-Aalen estimators. Closed-form variance estimators are proposed which take into consideration the potential replication of subjects across matched sets. The proposed methods are evaluated through simulation, then applied to estimate the effect of kidney transplantation on survival among end-stage renal disease patients using data from a national organ failure registry.

EO105 Room S106 MODELING AND CLASSIFICATION OF LARGE-SCALED, COMPLEX DATA Chair: Hyokyoung Grace Hong

E0303: **Time-varying copula models for longitudinal data**
Presenter: **Esra Kurum**, University of California, Riverside, United States
Co-authors: John Hughes, Runze Li, Saul Shiffman
A copula-based joint modeling framework for mixed longitudinal responses is proposed. The approach permits all model parameters to vary with time and thus will enable researchers to reveal dynamic response-predictor relationships and response-response associations. We call the new class of models timecop because we model dependence using a time-varying copula. We develop a one-step estimation procedure for the timecop parameter vector, and also describe how to estimate standard errors. We investigate the finite sample performance of our procedure via simulation studies, one of which shows that our procedure performs well under ignorable missingness. We also illustrate the applicability of our approach by analyzing binary and continuous responses from the Women's Interagency HIV Study.

E0488: **Self-semi-supervised clustering for large scaled-data with a massive null cluster**
Presenter: **Johan Lim**, Seoul National University, Korea, South
Co-authors: Soohyun Ahn, hyungwon Choi, Kyeong Eun Lee
Self-semi-supervised clustering, a new clustering method for large scale data with a massive null group, is proposed. Self-semi-supervised clustering is a two-stage procedure: preselect a part of "null" group from the data in the first stage and apply semi-supervised clustering to the rest of the data in the second stage, allowing them to be assigned to the null group. We evaluate the performance of the proposed method using a simulation study and demonstrate the method in the analysis of time course gene expression data from a longitudinal study of Influenza A virus infection.

E0620: **High dimensional classification with artificial neural network**
Presenter: **Taps Maiti**, Michigan State University, United States
High-dimensional models with correlated predictors are commonly seen in practice. Most proposed statistical models works well either in the low-dimensional correlated case, or in the high-dimensional independent case. Few methods deals with high-dimensional correlated predictors. Neural networks have been applied in practice for years, which have a good performance in correlated predictors due to the non-linearity brought by the activation functions. However, it may have too many parameters in high-dimensional case. With regularization, we are able to apply the neural network to high-dimensional correlated predictors case and obtain a parsimonious model with fairly good theoretical and numerical performance.

E0633: **Drawing inferences for high-dimensional linear models: A selection-assisted partial regression and smoothing approach**
Presenter: **Yi Li**, University of Michigan, United States
Drawing inferences for high-dimensional models is challenging as regular asymptotic theories are not applicable. A new framework of simultaneous estimation and inferences for high-dimensional linear models is proposed. By smoothing over partial regression estimates based on a given

variable selection scheme, we reduce the problem to a low-dimensional least squares estimation. The procedure, termed as Selection-assisted Partial Regression and Smoothing (SPARES), utilizes data splitting along with variable selection and partial regression. We show that the SPARES estimator is asymptotically unbiased and normal, and derive its variance via a nonparametric delta method. The utility of the procedure is evaluated under various simulation scenarios and via comparisons with the de-biased LASSO estimators, a major competitor. We apply the method to analyze a genomic dataset and obtain biologically meaningful results.

EO115 Room S1A01 RECENT ADVANCES IN MODEL SELECTION Chair: Garth Tarr

E0250: **GEE-assisted variable selection for latent variable models: Making the most of zero consistency**
Presenter: **Francis Hui**, The Australian National University, Australia
Co-authors: Samuel Mueller, Alan Welsh
In many disciplines, it is becoming common to collect and analyze multivariate or multi-response data. For example, the Southern Ocean Continuous Plankton Recorder (SO-CPR) survey is an annual survey which collect presence-absence observations on zooplankton assemblages in the Southern Ocean, with a primary goal being to identify important environmental factors driving the communities distribution while accounting for biotic affects such as species interactions. An increasingly popular approach for analyzing multivariate data in ecology is generalized linear latent variable models (GLLVMs), which utilizes latent variables to parsimoniously account for residual between species correlations. However, estimation let alone variable selection for GLLVMs presents a major computational challenge, since the marginal likelihood does not possess a closed form. To overcome this problem, we propose utilizing marginal generalized estimation equations (GEEs) to perform inference on GLLVMs. Focusing on multivariate binary data, we show that GEEs are zero consistent for GLLVMs. This then motivates us to propose two GEE-assisted selection methods: 1) information criteria based on score and Wald statistics; 2) penalized GEEs based on exploiting the grouped structure of the marginal coefficients. Both methods are asymptotically selection consistent for GLLVMs, with simulations studies demonstrating their computational efficiency and strong finite sample performance.

E0260: **Fast and approximate exhaustive variable selection for GLMs with APES**
Presenter: **Kevin Wang**, The University of Sydney, Australia
Co-authors: Garth Tarr, Jean Yang, Samuel Mueller
Obtaining maximum likelihood estimates for generalised linear models (GLMs) is computationally intensive and remains as the major obstacle for performing exhaustive variable selection. On the other hand, efficient algorithms for exhaustive searches do exist for linear models, most notably the leaps and bound algorithm and, more recently, the mixed integer optimisation algorithm. We present APES (APproximated Exhaustive Search), a new method that approximates all subset selection for a given GLM by reformulating the problem as a linear model. The method works by learning from observational weights in a correct/saturated generalised linear regression model. APES can be used in partnership with any other state- of-the-art linear model selection algorithm, thus enabling (approximate) exhaustive model exploration in dimensions much higher than previously feasible. We will demonstrate that APES model selection is competitive against genuine exhaustive search via simulation studies and applications to health data. The APES method is made available in R through the mplot package.

E0421: **Visualising model stability information for better prognosis based network-type feature extraction**
Presenter: **Connor Smith**, University of Sydney, Australia
Co-authors: Samuel Mueller, Boris Guennewig
Findings to deliver new statistical approaches to identify various types of interpretable feature representations that are prognostically informative in classifying complex diseases are reported. Identifying key features and their regulatory relationships which underlie biological processes is the fundamental objective of much biological research; this includes the study of human diseases, with direct and important implications in the development of target therapeutics. We present new and robust ways to visualise valuable information from the thousands of resamples in modern selection methods that use repeated subsampling to identify what features predict best disease progression. The new method VIVID learns from feature importance measures via pairwise feature comparisons to identify significant features. We will show how the selected features are repeatedly ranked higher and are more stable than other features. We take advantage of cluster analysis to first construct a set of nested feature groups and to then select an optimal group of features. We highlight the computational speed and requirements of VIVID and how it is able to deal with data where the number of features is continually increasing.

E0447: **On bivariate extreme value copulas with polynomial dependence functions**
Presenter: **Berwin Turlach**, The University of Western Australia, Australia
The aim is to discuss how the family of bivariate mixed model extreme value copula and the family of bivariate asymmetric mixed model extreme value copula can be extended to bivariate extreme value copulas with polynomial dependence function of arbitrary degree. We will discuss how extreme value copulas with polynomial dependence functions can be fitted to data and how the degree of the polynomial can be chosen.

EO304 Room AT241 STATISTICAL QUALITY TECHNOLOGIES Chair: Yuhlong Lio

E0164: **Planning accelerated life tests via GLMs**
Presenter: **Rong Pan**, Arizona State University, United States
A generalized linear model (GLM) approach to reliability data analysis is presented. Specifically, the failure time or censoring time data that are generated from an accelerated life test can be formulated by the GLM that defines the relationship between these test response data and the stress factor that was applied in the test. Through this modeling approach, it is easy to derive the test plan that can maximize its statistical efficiency in terms of the D-optimality or U-optimality criterion, etc. We also developed an R package, ALTopt, which can help practitioners to obtain optimal test plans even with multiple stress factors.

E0327: **Planning of accelerated degradation tests**
Presenter: **I-Chen Lee**, National Cheng Kung University, Taiwan
The accelerated degradation test (ADT) is widely used for assessing the lifetime information of highly reliable products. Because conducting an ADT is very expensive, how to plan an efficient ADT is a challenging issue for reliability analysts. By taking the experimental cost into consideration, an algorithm is proposed to determine the total sample size, testing stress levels, the measurement frequencies, and the number of measurements based on a class of exponential dispersion (ED) degradation models. For an ADT plan, the proposed method provides some design insights, and we further compare the planning strategies under different assumptions.

E0196: **Sequential process monitoring using empirical likelihood methods**
Presenter: **Cornelis Potgieter**, Southern Methodist University, United States
Many sequential monitoring procedures rely on strong parametric assumptions such as normality of the observations. When the parametric assumptions are not met, the procedure can have an inflated false signal rate. We construct sequential monitoring techniques using an empirical likelihood

approach. These techniques can be used to monitor for a change in mean and/or variance, and can also be used to monitor more robust measures of location/scale. The performance of the proposed methods is compared to some procedures commonly used in practice.

E0268: **Decision theoretic sampling plan for competing risks model**
Presenter: **Debasis Kundu**, Indian Institute of Technology Kanpur, India
An extensive amount of work has been done on the decision theoretic sampling plan based on censored lifetime data when the failure time is observed but no cause of failure is observed. We mainly discuss the decision theoretic sampling plans (DSP) for Type-II censored data when along with the failure time, the cause of failure is also observed. It is assumed that a lot is accepted provided the expected failure time for each cause is greater than a certain threshold. We define a suitable loss function and it is used to obtain the Bayes risk of the DSP. A finite algorithm is used to obtain the optimum DSP by minimizing the Bayes risk.

EO059 Room AT242 RECENT DEVELOPMENTS ON LATENT VARIABLE MODELS Chair: Xinyuan Song

E0255: **A Bayesian flexible joint model of multivariate longitudinal and survival data**
Presenter: **Kai Kang**, The Chinese University of Hong Kong, Hong Kong
Co-authors: Xinyuan Song
Joint models for analyzing longitudinal and survival data are widely used to investigate the relationship between a failure time process and time-variant predictors. A common assumption in conventional joint models in the survival analysis literature is that all predictors are observable. However, this assumption may not always be supported because unobservable traits, namely, latent variables, which are indirectly observable and should be measured through multiple observed variables, are commonly encountered in the medical, behavioral, and financial research settings. We propose a novel joint modeling approach to deal with this feature. The proposed model comprises three parts. The first part is a dynamic factor analysis model for characterizing latent variables through multiple observed indicators over time. The second part is a random coefficient trajectory model for describing the individual trajectories of latent variables. The third part is a proportional hazard model for examining the effects of time-invariant predictors and the longitudinal trajectories of time-variant latent risk factors on hazards of interest. We develop a Bayesian approach coupled with a Markov chain Monte Carlo algorithm to perform statistical inference. We conduct simulation studies to assess the empirical performance of the developed methodology. An application of the proposed joint model to a study on the Alzheimer's Disease Neuroimaging Initiative is presented.

E0203: **Bayesian analysis of hidden Markov structural equation models with an unknown number of hidden states**
Presenter: **Hefei Liu**, Qujing Normal University, China
Hidden Markov models (HMMs) are widely used to analyze heterogeneous longitudinal data owing to their capability to model dynamic heterogeneity. Early advancements in HMMs have mainly assumed that the number of hidden states is fixed and pre-determined based on the knowledge of the subjects or a certain criterion. However, as a limitation, this approach determines the number of hidden states on a pairwise basis, which becomes increasingly tedious when the state space is enlarged. Moreover, criterion-based statistics tend to select complex models with overestimated numbers of components in mixture modeling. A full Bayesian approach is developed to analyze hidden Markov structural equation models with an unknown number of hidden states. An efficient and hybrid algorithm that combines the reversible jump Markov chain Monte Carlo (RJMCMC) algorithm, the forward filtering and backward sampling scheme, and the Metropolis-Hastings algorithm is also developed to simultaneously select the number of hidden states and perform parameter estimation. The simulation study shows the satisfactory performance of the proposed method. A real data example concerning the prevention of cocaine use is also presented.

E0369: **Bayesian quantile scalar on image regression with non-ignorable non-response**
Presenter: **Qi Yang**, Department of Statistics, the Chinese University of Hong Kong, Hong Kong
Co-authors: Xinyuan Song
Exploiting imaging information is usually difficult because imaging data have ultrahigh dimensions. Functional principal component analysis can be used to reduce the dimensionality of imaging data, such that the reduced imaging information can be incorporated into conventional regression as predictors. However, a mean regression analysis generally does not provide a comprehensive understanding of the relationships between covariates and responses of interest. Instead, quantile regression enables the model to assess such relationships at different quantiles and simultaneously provides robust estimation results regardless of response distributions. We consider a quantile scalar on image regression model in the presence of nonignorable missing data. The proposed model includes ultrahigh dimensional imaging data as predictors to examine the effects of medical images and other covariates on the scalar response of interest. We develop a Bayesian approach with Markov chain Monte Carlo algorithms to conduct statistical inference. Simulation studies demonstrate that the proposed method performs satisfactorily. An application to a study of the Alzheimer's Disease Neuroimaging Initiative dataset is presented.

E0703: **Two-part hidden Markov model with semicontinuous longitudinal data**
Presenter: **Xiaoxiao Zhou**, The Chinese University of HongKong, Hong Kong
A two-part hidden Markov Model is developed for the analysis of semicontinuous longitudinal data. The two-part model manages a semicontinuous variable by splitting it into two random variables, a binary indicator to determine the occurrence of excess zeros at all occasions, and a continuous random variable to determine its actual level. For the continuous longitudinal response, a hidden Markov model (HMM) is proposed further to describe the relationship between the observation process and the unobservable finite-state transition process. The HMM consists of two major components. The first component is a transition model for investigating how potential covariates influence the probabilities of transitioning from one hidden state to another. The second component is a conditional regression model for examining the effects of covariates on the response. A full Bayesian approach together with efficient Markov chain Monte Carlo methods is developed for statistical inference in the presence of missing covariates. The proposed methodology is applied to a study on the Alzheimers Disease Neuroimaging data set. New sights into the pathology of Alzheimers disease and its potential risk factors are obtained.

EO285 Room AT335 ADVANCES IN HIDDEN MARKOV MODELS: THEORY AND APPLICATIONS Chair: Yang Chen

E0171: **A merging algorithm for hidden Markov models with unknown number of states**
Presenter: **Chu-Lan Kao**, National Chiao-Tung University, Taiwan
Co-authors: Cheng-Der Fuh, Yang Chen
Most inference techniques for hidden Markov models (HMM) require a prespecified number of hidden states. Traditional approaches follow two steps, order selection followed by inference, which potentially suffer from model misspecification. Another related field for time trajectory modeling, the change-point detection, allows us to cut the data into segments. These segments are likely to represent different states. We propose a merging algorithm that uses the change-point detection technique to conduct inference on HMM without requiring a predetermined number of states. The proposed algorithm connects these two highly investigated fields in statistics. Both theoretical and computational performances of the proposed algorithm are given.

E0343: **Predicting time series with abrupt changes and smooth evolutions**
Presenter: **Jie Ding**, University of Minnesota, United States
A methodology (referred to as kinetic prediction) is introduced for predicting time series undergoing unknown abrupt changes or smooth evolutions in their data generating distributions. Based on Kolmogorov-Tikhomirovs epsilon-entropy, we propose a concept called epsilon-predictability that quantifies the size of a model class and the maximal number of structural changes that guarantee the achievability of asymptotically optimal prediction. Moreover, for parametric distribution families, the aforementioned kinetic prediction with discretized function spaces is extended to its counterpart with continuous function spaces, which naturally leads to an efficient sequential Monte Carlo implementation. Wide applicability of the proposed methodology will be illustrated by its applications to time-varying cointegration, time-varying volatility models, and case studies in finance.

E0368: **Modeling dependence on the superposition of Markov chains: An application to ion channels**
Presenter: **Laura Jula vanegas**, University of Goettingen, Germany
Hidden Markov Models (HMM) are widely used for modeling temporal data in biostatistics (particularly for Ion channels), among other fields. Recent work has shown that the long-held belief that multiple ion channels in a membrane behave independently is often false. Models for dependence of multivariate Markov chains usually rely on the observation of each individual chain, whereas in our application we only have recordings for the superposition (sum) of the chains. We developed a coupled Hidden Markov Model for multiple dependent Markov chains, where the only information needed comes from the superposition of the chains. The model can explain a wide range of behavior, including negative and positive coupling. Our work shows the presence of negative coupling behavior in RyR2 Channels found in cardiac muscle.

E0461: **A joint HMM for RFID-based object tracking in a complex environment**
Presenter: **Kerby Shedden**, Statistics, United States
Co-authors: Yang Chen
RFID technology is an inexpensive way to track the movement of objects through an environment. We present a study in which patients and health care providers in an ophthalmology clinic were issued RFID tags, allowing their motions to be tracked at high temporal resolution. The main goals were to assess the amount of time each patient spends with providers, and to quantify time spent waiting at each stage of a clinic visit. Since RFID signals are weak, and the clinic environment contains many signal-distorting physical obstacles, the raw RFID data has many gaps, and frequently places patients and providers in impossible locations or on very unlikely trajectories of motion. We developed a joint HMM for patients and providers that produces realistic reconstructions of their joint motion. The model considers both within-target behavior (e.g. the rate of transition between rooms, statistical tendencies of different patients and providers to follow certain trajectories), and between-target behavior (e.g. the number and type of targets that can simultaneously occupy a location). New algorithms were developed to jointly fit a collection of interacting HMMs, incorporating the unique properties of the RFID signal distributions and domain-based probabilistic structure. We demonstrate that this approach produces accurate and informative summaries of patient and provider trajectories and removes most of the artifacts that are present in the raw data.

EO328 Room AT337 MULTIVARIATE METHODS FOR ANALYZING COMPLEX AND HIGH NOISY DATA Chair: Yoshikazu Terada

E0412: **Canonical dependency analysis by using chi-square matrix**
Presenter: **Jun Tsuchida**, Tokyo University of Science, Japan
Co-authors: Hiroshi Yadohisa
Canonical correlation analysis(CCA) is a popular method for investigating the relationship between datasets. CCA assumes that the relationship is represented as linear function. Therefore, it is not suitable to apply CCA to datasets whose relationship are non-linear. To achieve this problem, canonical dependency analysis(CDA) has been proposed by researchers. Many canonical dependency analyses adopt K-L divergence and kernel-based method as measure of dependency. Although these measures are useful for CDA, calculation cost is higher. Moreover, canonical variables maximize dependency between canonical variables. Thus, it is not as if the result of CDA is a summary of dependency between datasets. To overcome this problem, we propose CDA by using chi-square matrix. Chi-square matrix is calculated from datasets as similar manner of correlation matrix. Hence, using chi-square matrix, the proposed method summarizes dependency relationships between datasets. In addition, the calculation cost is not so high, because chi-square statistic is simple. As result of the numerical example, the proposed method has the best results in the sense of estimation accuracy of loadings matrix.

E0553: **Regularized interaction models for function-on-function regression**
Presenter: **Hidetoshi Matsui**, Shiga University, Japan
A regression model with a functional predictor and a functional response is considered. A functional quadratic model is an extension of a functional linear model and includes the quadratic term that takes the interaction between two different time points of the functional data into consideration. Predictor and the coefficient functions in the model are supposed to be expressed by basis expansions, and then parameters included in the model are estimated by the penalized likelihood method assuming that the error function follows a Gaussian process. Model selection criteria for evaluating the functional quadratic model are also derived using the idea of information-theoretic and Bayesian approach. The proposed method is applied to the analysis of meteorological data and the results are explored.

E0584: **Functional canonical correlation analysis for multivariate stochastic processes**
Presenter: **Michio Yamamoto**, Okayama University, RIKEN AIP, Japan
Co-authors: Yoshikazu Terada
The aim is to study the extension of canonical correlation analysis from pairs of random functions to the case where a data sample consists of multivariate square integrable stochastic processes. We refer to this extension as the generalized functional canonical correlation analysis (GFCCA). In functional data analysis, the data space is essentially infinite. Thus, unlike the generalized canonical correlation analysis (GCCA) for multivariate data, the well-definedness of GFCCA cannot be ensured in general. To address this issue, we provide sufficient conditions under which GFCCA has a meaningful solution. In addition, we develop the functional version of homogeneity analysis, which is another formulation of GCCA for multivariate data. Interestingly, we show that, unlike in the case of finite-dimensional space, the functional homogeneity analysis is not necessarily equivalent to GFCCA.

E0644: **Revisiting factor analysis: Three types of its formulation**
Presenter: **Kohei Adachi**, Osaka University, Japan
The main goal of factor analysis (FA) is to explain the variation of multiple observed variables by the factors called common and unique: the common factors serve for explaining the variation of all variables, while the unique factors have one-to-one correspondences to the variables. According to how the common and unique factors are treated, the FA procedures can be classified into three types. A classic one of them can be called a latent variable FA (LVFA), in which the factors are regarded as random latent variables. The remaining two types were recently proposed, in which the factors are treated as fixed parameter matrices. Those two types can be called matrix decomposition FA (MDFA) and constrained uniqueness FA (CUFA), respectively, as the former algorithm consists entirely of matrix-algebraic computations, and the latter is a constrained

version of MDFA. Here, the constraint requires each unique factor to affect the corresponding variable in a completely exclusive manner, and CUFA can be regarded as equivalent to minimum rank FA. We compare the three types of FA theoretically and empirically.

EO338 Room U301 STRUCTURAL INSTABILITIES IN HIGH-DIMENSIONAL DATA II Chair: Michal Pesta

E0441: **Hierarchical outer power Archimedean copulas**
Presenter: **Ostap Okhrin**, Dresden University of Technology, Germany
Co-authors: Jan Gorecki, Marius Hofert

Distributions based on hierarchical Archimedean copulas (HACs) became popular as they enable one to model non-elliptical and non-exchangeable dependencies among random variables. Their practical applications reported in the literature are, however, mostly limited to the case in which all generator functions in a HAC are one-parametric, which implies that all properties (e.g., Kendall's tau and tail dependence coefficients) of each bivariate margin of such a HAC is given just by a single parameter. Involving so-called outer power transformations of Archimedean generators in such models, this limitation can be alleviated, which typically allows one to set Kendall's tau and upper-tail dependence coefficient independently of each other. The construction, sampling and estimation of the resulting so-called hierarchical outer power Archimedean copulas are addressed.

E0198: **A monitoring procedure for detecting structural breaks in factor copula models**
Presenter: **Florian Stark**, University of Cologne, Germany
Co-authors: Dominik Wied, Hans Manner

A new monitoring procedure is proposed based on moving sums (MOSUM) for detecting single or multiple structural breaks in factor copula models. The test compares parameter estimates from a rolling window to those from a historical data set and analyzes the behavior under the null hypothesis of no parameter change. The case of multiple breaks is also treated. In the model, the joint copula is given by the copula of random variables which arise from a factor model. This is particularly useful for analyzing data with high dimensions. Parameters are estimated with the simulated method of moments (SMM). We analyze the behavior of the monitoring procedure in Monte Carlo simulations and a real data application. We consider an online procedure for predicting the day-ahead Value-at-risk based on the suggested monitoring procedure.

E0320: **L1 regularized smoothing with changepoints applied to implied volatility surfaces**
Presenter: **Matus Maciak**, Charles University, Czech Republic

Nonparametric models with changepoints gain on their popularity because of the overall flexibility which they offer. On the other hand, the theoretical background of such models is quite challenging and many problems need to be solved to correctly apply these models in real life situations. The idea is to introduce a direct method to estimate the implied volatility function (the implied volatility surface respectively) in option pricing by adopting the nonparametric smoothing together with the sparsity principle and recent developments in the area of atomic pursuit techniques. We propose the L1 regularized nonparametric estimation with possible changepoints (structural breaks) in the underlying dependence structure. Moreover, conditional quantiles can be obtained instead of the conditional mean and various hierarchical concepts can be used to specify the unknown model and the occurrence of hypothetical changepoints in it. Some theoretical results are derived and the consistency of the proposed method is proved with respect to the model estimation and the changepoint detection too. Empirical performance is investigated via a simulation study and some real implied volatility surface estimation.

E0823: **High-dimensional matrix autoregressive model**
Presenter: **Di Wang**, University of Hong Kong, Hong Kong
Co-authors: Guodong Li

High-dimensional matrix-valued time series data are getting widely available in many fields, such as macroeconomics, finance and image processing. It is natural to vectorize the matrix-valued data and apply the classic vector autoregressive model, but the model will suffer from the curse of dimensionality for high-dimensional data. We propose a matrix autoregressive model by folding the parameter matrix to a fourth-order tensor and consider a multilinear low rank structure for the parameter tensor to achieve substantial dimension reduction. Under the fixed dimension scenario, we study the asymptotic properties of the least squares estimator with low-rank constraints. For the case with much higher dimension, a novel convex regularization approach is proposed for estimation and the oracle inequalities are established. The methods can be readily extended to the high-dimensional tensor autoregressive model.

EO045 Room U302 ADVANCES IN NONLINEAR FINANCIAL ECONOMETRICS Chair: Jeroen Rombouts

E0409: **Multivariate lasso-based forecast combinations for stock market volatility**
Presenter: **Jeroen Rombouts**, ESSEC Business School, France

Volatility forecasts indicate future risk and are key inputs in financial analysis. We forecast the realized variance, an observable measure of volatility, of several major international stock market indices and account for the differing predictive information present in jump, continuous, and option-implied variance components. We allow for volatility spillovers of different stock markets by using a multivariate modeling approach. To obtain the forecasts, we use an ordered lasso-based forecast approach. A simple forecast combination of the univariate and multivariate model outperforms the benchmark model for all considered forecast horizons and stock markets.

E0419: **Quantile co-movement in stock markets with production linkages of firms**
Presenter: **Tomohiro Ando**, Melbourne Business School, Australia
Co-authors: Lina Lu

A spatial panel quantile model with unobserved heterogeneity is introduced. The proposed model is capable of capturing high-dimensional cross-sectional dependence and allows heterogeneous regression coefficients. For estimating model parameters, a new estimation procedure is proposed. When both the time and cross-sectional dimensions of the panel go to infinity, the uniform consistency and the asymptotic normality of the estimated parameters are established. In order to determine the dimension of the interactive fixed effects, we propose a new information criterion. It is shown that the criterion asymptotically selects the true dimension. Monte Carlo simulations document the satisfactory performance of the proposed method. Finally, the method is applied to study the quantile co-movement structure of the U.S. stock market by taking into account the input-output linkages as firms are connected through the input-output production network.

E0420: **Option pricing with conditional GARCH models**
Presenter: **Lars Stentoft**, University of Western Ontario, Canada
Co-authors: Marcos Escobar, Javad Rastegari Koopaei

The class of conditional GARCH models is introduced. These models offer significantly added flexibility to accommodate features of financial asset returns, akin to that of normal variance-mean mixtures, while admitting closed-form solutions for the moment generating function, a variance dependent pricing kernel and, therefore, efficient option pricing in a realistic setting. The class of conditional Gaussian models naturally generalizes the Heston-Nandi model and combined with a Regime Switching model illustrates the flexibility of our methodology and demonstrates the importance in terms of option prices and Greeks of accommodating crisis periods and state dependency as well as priced variance risk.

E0443: **Dynamic properties and correlation structure of a large panel of cryptocurrencies**
Presenter: **Francesco Violante**, ENSAE ParisTech, France
Co-authors: Luc Bauwens, Jeroen Rombouts
The behaviour of a large portfolio of highly valued and most actively traded cryptocurrencies is studied. Unlike more traditional financial assets, the dynamic behaviour of cryptocurrencies returns is characterised by a particularly high level of volatility, by abnormally large variations, and is affected by extreme shocks to liquidity. We aim at investigating the dynamic properties of cryptocurrencies and particularly the correlation structure linking them, with the scope to identify whether and to what extent there exist diversification opportunities in these markets.

EO031 Room U414 ECONOMETRIC AND STATISTICAL MODELLING OF TIME SERIES AND SPATIAL PROCESSES Chair: Guodong Li

E0310: **Time-varying graphs by locally stationary Hawkes processes**
Presenter: **Hiroshi Shiraishi**, Keio University, Japan
Co-authors: Taiga Uno, Yu Izumisawa, Junichi Hirukawa
Hawkes Graphs have been recently introduced to grasp the branching structure of multivariate stationary Hawkes processes. However, the existing procedure cannot describe the time structural changes, since stationary Hawkes processes are a class of stationary processes. We introduce a multivariate locally stationary Hawkes (lsHawkes) process, which is a natural extension of the univariate lsHawkes process. We first consider an approximation of the lsHawkes process by a time-varying integer-valued autoregressive (tvINAR) process. Then, we propose an estimation procedure for the time varying parameters based on the local least-squares method. Finally, we propose time-varying Hawkes graphs (tvHawkes graphs) by using the estimated parameters.

E0473: **Model averaging marginal nonlinear logistic regressions for time series data**
Presenter: **rong peng**, University of Southampton, United Kingdom
Co-authors: Zudi Lu
A semi-parametric procedure named Model Averaging MArginal nonlinear LOgistic Regressions (MAMALOR) is proposed, which is flexible for forecasting of binary count time series data. It is motivated by applications in such scenarios of forecasting the price up/down direction in stock market and the default/non-default in credit scoring. Such binary time series exist in wide applications beyond finance though the considered financial application. The procedure can avoid the curse of dimensionality for high dimension d and be easily carried out by maximum likelihood methods. Our initial application shows that the MAMALOR procedure is promising in forecasting of the stock price direction, outperforming the linear logistic regression and the (logistic) generalised additive modelling. We comment that although this procedure is basically focused on binary time series data, the ideas and methodology as well as insights into applications, learned from this project, will help to further study other count time series data in a generalised setting.

E0665: **Multilinear low-rank vector autoregressive modeling via tensor decomposition**
Presenter: **Guodong Li**, University of Hong Kong, Hong Kong
The classic vector autoregressive model is a fundamental tool for multivariate time series analysis. However, it involves too many parameters for high-dimensional time series, and hence suffers from the curse of dimensionality. We rearrange the parameter matrices of a vector autoregressive model into a tensor form, and use the tensor decomposition to restrict the parameter space in three directions. Compared with the reduced-rank regression method, which can limit the parameter space in one direction only, the proposed method dramatically improves the capability of vector autoregressive models in handling high-dimensional time series. For this method, its asymptotic properties are studied and an alternating least squares algorithm is suggested. Moreover, for the case with much higher dimension, we further assume the sparsity of three loading matrices, and the regularization method is thus considered for estimation and variable selection. An ADMM-based algorithm is proposed for the regularized method and oracle inequalities for the global minimizer are established.

E0424: **Measure the generality of convolutional layers with projection correlation**
Presenter: **Yuan Ke**, University of Georgia, United States
The transfer learning problem is studied in image classification applications. A phase transition phenomenon has been empirically validated: the convolutional layer shifts from general to specific with respect to the target task as its depth increases. It is suggested that measuring the generality of convolutional layers through an easy to compute and tuning free quantity named projection correlation. The non-asymptotic upper bounds for the estimation error of the proposed generality measure has been provided. Based on this generality measure, a forward adding layer selection algorithm to select generable layers is proposed. The algorithm aims to find a cut-off in the pre-trained model according to where the phase transition from general to specific happens. Then, we propose to transfer only the generable layers as specific layers can cause overfitting issues and hence hurt the prediction performance. The proposed algorithm is computationally efficient and can consistently estimate the true location of phase transition under mild conditions. Its superior empirical performance has been justified by various numerical experiments.

EO293 Room U502 PROBABILITY TECHNIQUES IN STATISTICS, ECONOMICS OR FINANCE Chair: Aklilu Zeleke

E0202: **Time-changed Poisson processes of order k**
Presenter: **Neelesh Shankar Upadhye**, Indian Institute of Technology Madras, India
Co-authors: Ayushi Sengar, Aditya Maheshwari
The aim is to study the Poisson process of order k (PPoK) time-changed with an independent Lévy subordinator and its inverse, which we call respectively, as TCPPoK-I and TCPPoK-II, through various distributional properties, long-range dependence and limit theorems for the PPoK and the TCPPoK-I. Further, we study the governing difference-differential equations of the TCPPoK-I for the case inverse Gaussian subordinator. Similarly, we investigate the distributional properties, asymptotic moments and the governing difference-differential equation of TCPPoK-II. As an application to ruin theory, we give a governing differential equation of ruin probability in insurance ruin using these processes. Finally, we present some simulated sample paths of both the processes.

E0212: **A generalized normal distribution with applications to fit financial and economic data**
Presenter: **Carl Lee**, Central Michigan University, United States
The distributions of financial and economic data are often highly skewed. During the recent decades, some new methods have been developed for generating highly flexible distributions with four or more parameters. These flexible distributions, although are capable of fitting highly skewed data, they suffer a common weakness of requiring four or more parameters, which have no practical and meaningful interpretations. This article presents a three-parameter generalized normal distribution capable of fitting highly skewed data. One parameter characterizes the location, two parameters together characterize a very wide range of scale, skewness and kurtosis. The distribution is applied to fit the World Trade import and export data, and the time and cost to start a business of 200 countries in the world. Ten years of World Trade import and export data (2008 to 2017) are extracted from WTO (World Trade Organization) database. Ten years of time and cost to start a business (2008-2017) are extracted from the World Bank database. The goodness of fits are presented and the pattern of the distribution shift during the ten years period are investigated.

E0408: **Exponential order statistics and some combinatorial identities**
Presenter: **Aklilu Zeleke**, Michigan State University, United States
Co-authors: Palaniappan Vellaisamy
It is known that order statistics from exponential distribution have several interesting properties. We consider, without loss of generality, the exponential distribution with mean unity. For example, the k-th order statistic, $1 \leq k \leq n$, has the distribution of sum of independent exponential random variables (rvs) with different parameters. The usual proof of this result uses the transformation to the set of spacings from the set of order statistics and by applying Jacobian density theorem. We prove the above-mentioned result using the Laplace transform methods. The main purpose is to bring out the connections between exponential order statistics and several combinatorial identities. In fact, we give simpler proofs of several combinatorial/binomial identities by evaluating the Laplace transformation of the k-th exponential order statistic by two different ways and equating them. A probabilistic interpretation and some extensions of these combinatorial identities are also discussed.

E0512: **Copula-based structure of co-movement between oil and manufacturing outputs**
Presenter: **Ibrahim Abdalla Alfaki**, United Arab Emirates University, United Arab Emirates
The main objective is to investigate the co-movement and interdependency between macroeconomic variables; namely the oil and the manufacturing sectors' output volatility. Distributional properties and characteristics, including appropriate marginal distributions of both time series are illustrated. Copula-based multivariate analyses, including copula-based GARCH modeling strategy and its respective variants, are used to capture possible dependency structure. The discussion is further complemented by focusing on the case of the United Arab Emirates (UAE), a major producer of crude oil, ranked in the top 10 exporters. This is an interesting case, because understanding and modeling of dependence between oil and manufacturing outputs are crucial to the countrys strategy to diversify away from the dominance of a single commodity. The analysis Utilizes available annual data covering the period from 1975 to 2010, representing real UAE oil and manufacturing sector's output measured annually in the local UAE currency at 2001 constant prices.

Thursday 27.06.2019 10:20 - 12:25 Parallel Session K – EcoSta2019

EO275 Room S101 NEW ADVANCE IN LEARNING THEORY AND RELATED APPLICATION Chair: Zhigang Yao

E0772: **Principal boundary on Riemannian manifolds**
Presenter: **Zhigang Yao**, National University of Singapore, Singapore
Co-authors: Zhenyue Zhang
The classification problem is considered. The focus is on nonlinear methods for classification on manifolds. For multivariate datasets lying on an embedded nonlinear Riemannian manifold within the higher-dimensional ambient space, we aim to acquire a classification boundary for the classes with labels, using the intrinsic metric on the manifolds. Motivated by finding an optimal boundary between the two classes, we invent a novel approach – the principal boundary. From the perspective of classification, the principal boundary is defined as an optimal curve that moves in between the principal flows traced out from two classes of data, and at any point on the boundary, it maximizes the margin between the two classes. We estimate the boundary in quality with its direction, supervised by the two principal flows. We show that the principal boundary yields the usual decision boundary found by the support vector machine in the sense that locally, the two boundaries coincide. Some optimality and convergence properties of the random principal boundary and its population counterpart are also shown. We illustrate how to find, use and interpret the principal boundary with an application in real data.

E0681: **Feedback of spectral projection for subspace learning**
Presenter: **Yuqing Xia**, National University of Singapore, Singapore
Co-authors: Zhenyue Zhang
Representation based methods for subspace learning consist of two stages: Affinity learning and spectral clustering. A feedback strategy is proposed to softly combine the two separate stages together by simultaneously optimizing the affinity and spectral projection. The soft feedback strategy can strengthen the required block-diagonal structure of the affinity matrix for most of the existing state-of-art algorithms. Using the feedback strategy, a scalable and projection unified model is given for datasets in large scale. A fast and efficient algorithm is also given to solve this problem, based on active piece-wise sign updating. Experiments are reported to demonstrate the improvement for the existing algorithms and the effectiveness and efficiency of the proposed model on large scale datasets.

E0688: **Quantifying time-varying sources in magnetoencephalography: A discrete approach**
Presenter: **Zengyan Fan**, National University of Singapore, Singapore
Co-authors: Zhigang Yao, Masahito Hayashi, William Eddy
The aim is to discuss the distribution of brain sources from the most advanced brain imaging technique, Magnetoencephalography (MEG), which measures the magnetic fields outside the human head produced by the electrical activity inside the brain. Common time-varying source localization methods assume the source current with a time-varying structure and solve the MEG inverse problem by mainly estimating the source moment parameters. These methods use the fact that the magnetic fields linearly depend on the moment parameters of the source, and work well under the linear dynamic system. However, magnetic fields are known to be non-lineally related to the location parameters of the source. The existing work on estimating the time-varying unknown location parameters is limited. We are motivated to investigate the source distribution for the location parameters based on a dynamic framework, where the posterior distribution of the source is computed in a closed form discretely. Both a dynamic procedure and a switch procedure are proposed for the new discrete approach, balancing estimation accuracy and computational efficiency when multiple sources are present. Lastly, we will discuss the source localization for the Brain-Controlled Interfaces data and illustrate that the new method is able to provide comprehensive insight into the time evolution of the sources at different stages of the experiment.

E0736: **Manifold learning in ambient space**
Presenter: **Wee Chin Tan**, National University of Singapore, Singapore
Many real life data lie in high dimensional spaces. For example, a matrix of high dimension is used to represent an image. In the analysis high dimensional data, the curse of dimensionality arises. In recent years, there have been vast developments of linear and nonlinear dimension reduction techniques to overcome the curse of dimensionality. These techniques are sometimes called manifold learning. They assume that there is an underlying manifold of dimension much smaller than the dimension of the space which the data lie in. They aim is to overcome the curse of dimensionality by learning the behavior of the data from a lower dimensional space. In reality, data are very noisy, and it is not easy to extract the relevant information. The method which is proposed extends the idea of subsampling to noisy data sets in higher dimensional space, and utilizes the Moving Least Square method to approximate the underlying manifold. Subsampling is used in computer graphics to reduce the image size. The idea subsampling is to extract a small subset from the input data which preserves a large amount of information regarding the underlying manifold. This will greatly reduce the computation complexity of learning the manifold.

E0826: **Two-sample Mendelian randomization for summary statistics accounting linkage disequilibrium**
Presenter: **Qing Cheng**, Duke-NUS, Singapore
Obtaining a reliable causal relationship between risk exposures and disease outcomes from epidemiological studies remains an essential challenge. Proliferation of genome-wide association studies (GWAS) has prompted the use of two-sample Mendelian randomization (MR) with genetic variants as instrumental variables (IV) for data analysis and interpretation. However, most of existing methods assume that IVs are not in linkage disequilibrium(LD) which can lead to biased estimates and false-positive casual relationships. To overcome these limitations, we propose a probabilistic model thatleverages GWAS summary statistics in the presence of LD, as well as properly accounts for horizontal pleiotropy among genetic variants (MR-LDP). MR-LDP utilizes a computationally efficient variational Bayes expectation-maximization (VBEM) algorithm, calibrating evidence lower bound (ELBO) for a likelihood ratio test. We further conducted comprehensive simulation studies to demonstrate the advantages of MR-LDP over existing methods in terms of both type-I error control and point estimates. Moreover, we used two real expsoure-outcome pairs to validate results from MR-LDP in comparison with alternative methods, particularly showing our method is more efficient using all genetic variants in LD.

EO221 Room S102 RECENT ADVANCES IN COMPLEX BIOMETRIC DATA ANALYSIS Chair: Sangbum Choi

E0700: **Double-robust inference for differences in restricted mean lifetimes using pseudo-observations**
Presenter: **Sangbum Choi**, Korea University, Korea, South
When comparing survival times between two treatment groups, restricted mean lifetime, defined as the expectation of the survival function restricted to a prespecified time point, is often of direct interest, as it is easily understood by clinician investigators and does not require restrictive assumptions, such as proportionality. If the treatments are not randomized as in observational studies, covariate adjustment is needed to account for treatment imbalances in confounding factors. We propose a simple pseudo-value approach to estimate the difference of the restricted mean lifetime between two groups while accounting for confounders, which can be used as a metric for average causal effect (ACE). The proposed method combines two general approaches, (1) group-specific regression models (Q-model) for the time-to-event and covariate information, and (2) inverse probability of treatment assignment weights (A-model), where the restricted mean lifetimes are replaced by the corresponding pseudo-observations for survival outcomes. We show that the proposed estimator is double robust in the sense that it is consistent if at least one of the two working models remains correct. Simulation studies are conducted to assess its finite-sample performance and the method is applied to kidney transplant data.

E0737: **Cost-effective extreme case-control design using resampling method**
Presenter: **Young Min Kim**, Kyungpook National University, Korea, South
Nested case-control sampling design is a popular method in a cohort study whose events are often rare. The controls are randomly selected with or without the matching variable fully observed across all cohort samples to control confounding factors. We propose a new nested case-control sampling design incorporating both extreme case-control design and a resampling technique. This new algorithm has two main advantages with respect to the conventional nested case-control design. First, it inherits the strength of extreme case-control design such that it does not require the risk sets in each event time to be specified. Second, the target number of controls can only be determined by the budget and time constraints and the resampling method allows an under sampling design, which means that the total number of sampled controls can be smaller than the number of cases. A simulation study demonstrated that the proposed algorithm performs well even when we have a smaller number of controls compared to the number of cases. The proposed sampling algorithm is applied to a public data collected for Thorotrast Study.

E0770: **Multiple response logistic regression with structure constraints and its applications to Cancer Cell Line Encyclopedia**
Presenter: **Seyoung Park**, Sungkyunkwan University, Korea, South
Cancer cell lines (CCLs) play a critical role in enabling the screening of drug candidates at an early stage of drug discovery. Due to the complex interactions between cell lines and drug molecules, identifying the correlation between the variability in molecular profiles and pharmacological responses is challenging. We propose a multiple response logistic regression framework with structure constraints that exploits information on gene expression of CCLs and multiple drug response binary data to incorporate similarity among drugs. We identify a limited set of genes that might be directly involved in drug sensitivity or resistance.

E0783: **Permutation tests for general dependent truncation**
Presenter: **Sy Han Chiou**, University of Texas at Dallas, United States
Co-authors: Jing Qian, Elizabeth Mormino, Rebecca Betensky
Truncated survival data arise when the event time is observed only if it falls within a subject-specific region, known as the truncation set. Left-truncated data arise when there is delayed entry into a study, such that subjects are included only if their event time exceeds some other time. Quasi-independence of truncation and failure refers to factorization of their joint density in the observable region. Under quasi-independence, standard methods for survival data such as the Kaplan-Meier estimator and Cox regression can be applied after simple adjustments to the risk sets. Unlike the requisite assumption of independent censoring, quasi-independence can be tested, e.g., using a conditional Kendall's tau test. Current methods for testing for quasi-independence are powerful for monotone alternatives. Nonetheless, it is essential to detect any deviation from quasi-independence so as not to report a biased Kaplan-Meier estimator or regression effect, which would arise from applying the simple risk set adjustment when dependence holds. Nonparametric, minimum p-value tests that are powerful against non-monotone alternatives are developed to offer protection against erroneous assumptions of quasi-independence. The use of conditional and unconditional methods of permutation for evaluation of the proposed tests is investigated in simulation studies. The proposed tests are applied to a study on the cognitive and functional decline in aging.

E0829: **Induced-smoothed quantile regression analysis for competing risks data under case-cohort study**
Presenter: **Sangwook Kang**, Yonsei University, Korea, South
Co-authors: Dongjae Son, Sangbum Choi
Cohort sampling designs offer an economical and efficient way of investigating association between exposure variables and risk of disease outcomes. A case-cohort design is a cohort sampling design in which a disproportionate fractions of failures and censored subjects are sampled. We consider competing risks data arising from case-cohort studies and propose statistical inference procedures for fitting censored quantile regression models for such data. Estimation of regression parameters is based on an induced smoothing approach applied to nonsmooth weighted estimating equations. Two types of weight are considered - The inverse censoring probability and sampling probability weights are included to account for competing risks data and biased feature in case-cohort samplings, respectively. The proposed induced smoothed estimating functions are smooth in regression parameters enabling one to apply the standard numerical algorithms such as the Newton's method. An iterative algorithm is proposed to simultaneously estimate regression parameters and their variances. Asymptotic properties of the proposed estimators are established. Finite sample properties are investigated through extensive simulation studies. The proposed methods are illustrated with two real data sets.

EO099 Room S104 RECENT ADVANCES IN HIGH DIMENSIONAL STATISTICS Chair: Peter Radchenko

E0271: **Co-manifold learning on data matrices**
Presenter: **Eric Chi**, North Carolina State University, United States
Co-authors: Gal Mishne, Ronald Coifman
A new method is introduced for performing joint dimension reduction, or manifold learning, on the rows and columns of a data matrix. Our approach generalizes recent work on a convex formulation of the biclustering problem. Like convex biclustering, our co-manifold learning procedure possesses stability guarantees with respect to perturbations in the data. We illustrate how out method can identify coupled row and column geometries in simulated and real data examples.

E0556: **Grouped variable selection with discrete optimization**
Presenter: **Peter Radchenko**, University of Sydney, Australia
Co-authors: Rahul Mazumder, Hussein Hazimeh
The focus is on a new tractable framework for grouped variable selection with a cardinality constraint on the number of selected groups, leveraging tools in modern mathematical optimization. The proposed methodology covers both the case of high-dimensional linear regression and nonparametric sparse additive modelling. Computational experiments demonstrate the effectiveness of the proposal as an alternative method for

sparse grouped variable selection - in terms of better predictive accuracy and greater model sparsity, at the cost of increased, but still reasonable, computation times. Empirical and theoretical evidence shows that the proposed estimators outperform their Group Lasso type counterparts in a wide variety of regimes.

E0646: **GAP: A general framework for information pooling in two-sample high-dimensional regression models**
Presenter: **Wenguang Sun**, University of Southern California, United States
A general framework is discussed for exploiting the sparsity information in two-sample multiple testing problems. We propose to first construct a covariate sequence, in addition to the usual primary test statistics, to capture the sparsity structure, and then incorporate the auxiliary covariates in inference via a three-step algorithm consisting of grouping, adjusting and pooling (GAP). The GAP procedure provides a simple and effective framework for information pooling. An important advantage of GAP is its capability of handling various dependence structures such as those arise from high-dimensional linear regression, differential correlation analysis, and differential network analysis. We establish general conditions under which GAP is asymptotically valid for false discovery rate control, and show that these conditions are fulfilled in a range of settings, including testing high-dimensional linear regression, differential covariance or correlation matrices, and Gaussian graphical models. Numerical results demonstrate that existing methods can be significantly improved by the proposed framework. The GAP procedure is illustrated using a breast cancer study for identifying gene-gene interactions.

E0534: **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 on 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.

E0654: **Principal component reduction of a nonparametric additive model with variable selection**
Presenter: **Shiyuan He**, Renmin University of China, China
Co-authors: Kejun He
Additive models have been widely used as a flexible nonparametric regression method that can overcome the curse of dimensionality. By using sparsity-inducing penalty for variable selection, several authors have developed methods for fitting additive models when the number of predictors is very large, sometimes even larger than the sample size. However, despite good asymptotic properties, the finite sample performance of existing methods deteriorates considerably when the number of relevant predictors becomes moderately large. We propose to reduce the number of additive component functions to be estimated using principal components. To fit the reduced additive model to the data, we develop a novel algorithm to solve the penalized least squares on a fixed-rank manifold with a sparsity-inducing penalty. Our asymptotic theory shows that the resulting estimator has faster convergence rate than estimating without principal component reduction; and this is true even when the reduced model is only an approximation, provided that the approximation error is small. Moreover, the proposed method is able to consistently identify the relevant predictors. The advantage of the reduced additive model is also illustrated using a simulation study.

EO324 Room S106 ECOSTA JOURNAL: ECONOMETRICS AND STATISTICS Chair: Stella Hadjiantoni

E0698: **Partial derivative functional response model with application to temperature dynamic**
Presenter: **Catherine Liu**, The Hong Kong Polytechnic University, Hong Kong
Co-authors: Tao Zhang, Zhaohai Li, Jin Yang
A novel functional regression model is described where the response is the partial derivative of bivariate functional data and the predictor is a vector or scalar. An estimation procedure of the model is proposed. We apply this partial derivative functional response model to a dynamic analysis of temperature data to explore how the latitude, longitude, and their interaction impact the rate of change of the temperature over time (days or years) in the Midwest of USA. Simulation studies prove that the proposed method performs well in finite samples.

E0693: **Numerical strategies for the estimation of functional regression models**
Presenter: **Stella Hadjiantoni**, University of Kent, UK, United Kingdom
Co-authors: Ana Colubi, Erricos John Kontoghiorghes
In functional data analysis, the discrete observed data are converted to smooth functions and so they become infinite dimensional data objects. The analysis involves representing the functional data using a basis expansion and then truncating the basis in term of a finite number of basis elements. Choosing the number of basis elements is part of the data analysis. Therefore, the dimension of the basis expansion is an unknown parameter and investigation is required to determine its value. A recursive numerical method is examined for choosing the number of basis elements within the context of model selection. Penalised least squares and cross validation procedures are used in order to choose the number of basis elements that optimise the estimation of the functional regression model. The proposed numerical method is based on orthogonal and hyperbolic transformations.

E0788: **On structure testing for component covariance matrices of a high-dimensional mixture**
Presenter: **Jeff Yao**, The University of Hong Kong, Hong Kong
Co-authors: Weiming Li
By studying the family of p-dimensional scale mixtures, a non trivial example is shown where the eigenvalue distribution of the corresponding sample covariance matrix does not converge to the celebrated Marčenko-Pastur law. A different and new limit is found and characterized. The reasons of failure of the Marčenko-Pastur limit in this situation are found to be a strong dependence between the p-coordinates of the mixture. Next, we address the problem of testing whether the mixture has a spherical covariance matrix. To analyse the traditional John's type test we establish a novel and general CLT for linear statistics of eigenvalues of the sample covariance matrix. It is shown that the John's test and its recent high-dimensional extensions both fail for high-dimensional mixtures, precisely due to the different spectral limit above. As a remedy, a new test procedure is constructed afterwards for the sphericity hypothesis .This test is then applied to identify the covariance structure in model-based clustering. It is shown that the test has much higher power than the widely used ICL and BIC criteria in detecting non-spherical component covariance matrices of a high-dimensional mixture.

E0201: **Intertemporal similarity of economic time series: An application of dynamic time warping**
Presenter: **Philip Hans Franses**, Erasmus School of Economics, Netherlands
The non-parametric Dynamic Time Warping (DTW) technique is adapted to an application to examine the temporal alignment and similarity across economic time series. DTW has important advantages over existing measures in economics as it alleviates concerns regarding a pre-defined fixed temporal alignment of series. For example, in contrast to current methods, DTW can capture alternations between leading and lagging relationships of series. We illustrate DTW in a study of US states business cycles around the Great Recession, and find considerable evidence that temporal alignments across states dynamic. Trough cluster analysis, we further document state-varying recoveries from the recession.

E0722: **News and intraday jumps: A big data approach**
Presenter: **Massimiliano Caporin**, University of Padova, Italy
Co-authors: Francesco Poli
The aim is to study how news provokes intraday price jumps in the S&P 100 constituents. We build high-frequency indicators that go beyond the mere release of news and investigate their association with jumps by applying penalised logistic regression and by dealing with the rare nature of jumps with appropriate techniques. Relevant causes of jumps are found to be EPS, rate decisions, bad macro-news and company-specific news with specific topics or negative sentiment. Market players sometimes act before the public release of information. Finally, we find that news influences the economic significance of jumps in terms of returns predictability and volatility persistence.

EO223 Room S1A01 NEW METHODS IN THE MODELING OF FUNCTIONAL AND HIGH-DIMENSIONAL DATA Chair: Shu-Chuan Chen

E0330: **Bayesian Ising sparse nonparametric model**
Presenter: **Inyoung Kim**, Virginia Tech, United States
A Bayesian Ising Sparse nonparametric model is proposed for variable selection via graphical and Ising models for the ordered categorical clinical outcome. The Bayesian variable problem can be considered as a complete graph and described by an Ising model with random interactions. There are several advantages of our approach: it is applicable to (1) problems with small sample sizes and a larger number of variables and (2) any nonparametric regression models; and easy to (3) incorporate graphical prior information. The results indicate that the best prior for the model coefficients in terms of variable selection should place substantial weight on small, nonzero shrinkage. We also discuss the relationship between the tempering algorithms for the Ising model and the global-local shrinkage approach, showing that the shrinkage parameter plays a tempering role. The methods are illustrated with simulated and real data.

E0337: **Predicting one-day-ahead wind power capacity factor via functional inverse regression**
Presenter: **Ci-Ren Jiang**, Academia Sinica, Taiwan
Co-authors: Lu-Hung Chen
Inverse regression is an appealing dimension reduction method for regression models with multivariate covariates. Recently, it has been extended to the cases with functional or longitudinal covariates. However, the extensions simply focus on one single functional or longitudinal covariate. Motivated by a real application, we extend functional inverse regression to the cases with multiple functional covariates, whose domains could be different. The asymptotic properties of the proposed estimators are investigated. The computational issues are taken care with data binning, the fast Fourier transformation and random projections on a multi-core computation platform. In addition to simulation studies, the proposed approach is applied to predict the one-day-ahead wind power capacity factors in Germany from 2016 to 2017. Both demonstrate the good performance of our method.

E0472: **Nonparametric logistic regression in high dimension**
Presenter: **Jong Soo Lee**, University of Massachusetts Lowell, United States
Co-authors: Johan Lim
The maximum penalized likelihood estimation for logistic regression in high dimension is studied. When dealing with high dimensional regression problem, one uses either the dimension reduction techniques in input variables or use penalized approach using ridge, LASSO, or elastic net as penalty functions. We propose a new penalty function for logistic regression that alleviates some of the issues in ridge or LASSO type penalties, but at the same time preserves the optimal convergence rate under mild conditions.

E0599: **A hierarchical model to handle missing data in high-dimensional spatio-temporal models**
Presenter: **Shyam Ranganathan**, Virginia Tech, United States
High-dimensional spatio-temporal models have important applications in health analytics. A number of variables, especially from birth records are available with increasing frequency, and a current challenge is to make efficient models that can explain the data in a manner that corresponds to causal or mechanistic explanations (as opposed to black box machine learning methods). We demonstrate how missing data can significantly impact inferences and prediction, and hence black box methods are "dangerous". We present an efficient model to handle missing data in these complex settings and present a hierarchical model that is better suited to prediction in these high-dimensional spatio-temporal problems.

E0618: **Feature selection for functional predictors**
Presenter: **Dennis Cox**, Rice University, United States
In prediction problems with functional predictors the individual values of the functions typically have little predictive value. The problem of selecting features depends very much on the particular problem at hand. Several data sets with spectroscopic measurements of patient tissue are considered. The objective is to predict whether or not there is precancerous disease at the measurement site. A variety of methods for extracting potentially useful features are considered in the context of various machine learning algorithms that are employed for the prediction. Some simple, classical methods perform quite well for these applications.

EO057 Room AT241 STATISTICAL ANALYSIS OF COMPLEX DATA Chair: Xinyuan Song

E0238: **An effective likelihood-free approximate computing method with statistical inferential guarantees**
Presenter: **Wentao Li**, Newcastle University, United Kingdom
Co-authors: Suzanne Thornton, Min-ge Xie
Approximate Bayesian computing is a powerful likelihood-free method that has grown increasingly popular. However, complications arise in the theoretical justification for Bayesian inference conducted from this method with a non-sufficient summary statistic. We seek to re-frame approximate Bayesian computing within a frequentist context and justify its performance by standards set on the frequency coverage rate. In doing so, we develop a new computational technique called approximate confidence distribution computing, yielding theoretical support for the use of non-sufficient summary statistics in likelihood-free methods. Furthermore, we demonstrate that approximate confidence distribution computing extends the scope of approximate Bayesian computing to include data-dependent priors without damaging the inferential integrity. This data-dependent prior can be viewed as an initial distribution estimate of the target parameter which is updated with the results of the approximate confidence distribution computing method. A general strategy for constructing an appropriate data-dependent prior is also discussed and is shown to often increase the computing speed while maintaining statistical inferential guarantees. We supplement the theory with simulation studies illustrating the benefits of the proposed method, namely the potential for broader applications and the increased computing speed compared to the standard approximate Bayesian computing methods.

E0241: **Direct local linear estimation for Sharpe ratio function in heteroscedastic regression models**
Presenter: **Hongmei Lin**, Shanghai University of International Business and Economics, China
The heteroscedastic regression model has been widely used in financial econometrics. It allows us to deal with nonlinearity and heteroscedasticity in financial time series. As the ratio of the mean and volatility functions, the Sharpe ratio is one of the most widely used risk or return measures

in finance. We propose a new nonparametric method to estimate the Sharpe ratio function directly using local linear regression. We establish the asymptotic normality for the proposed estimator. Monte Carlo simulation studies show the proposed estimator has excellent finite sample performance and outperform existing indirect method. We illustrate our method with a real data example.

E0237: **Batch effects correction for single-cell RNA sequencing data**
Presenter: **Fangda Song**, The Chinese University of Hong Kong, Hong Kong
Co-authors: Yingying Wei
Despite their widespread applications, single-cell RNA-sequencing (scRNA-seq) experiments are still plagued by batch effects and dropout events. Although the completely randomized experimental design has frequently been advocated to control for batch effects, it is rarely implemented in real applications due to time and budget constraints. We mathematically prove that under two designs—true biological variability can also be separated from batch effects. We develop Batch Effects correction with Unknown Subtypes for scRNA-seq data (BUSseq), which is an interpretable Bayesian hierarchical model that closely follows the data-generating mechanism of scRNA-seq experiments. BUSseq can simultaneously correct batch effects, cluster cell types, impute missing data caused by dropout events and detect differentially expressed genes without requiring a preliminary normalization step. We demonstrate that BUSseq outperforms existing methods with simulated and real data.

E0589: **Variable selection for high-dimensional features with missing data**
Presenter: **Alex Kin Yau Wong**, Hong Kong Polytechnic University, Hong Kong
In biomedical, epidemiological, or social studies, one often encounters high-dimensional data with missing data. Conventional methods for handling high-dimensional data, such as penalization methods, are not directly applicable to problems with missing data. Simple methods for handling missing data, such as complete-case analysis and single imputation, are generally inefficient and may even be invalid. We consider a regression framework with incomplete predictors and propose a latent variable model to characterize the relationships among predictors and to infer missing values from observed data. Under this framework, we propose a penalized regression parameter estimator and develop a computationally efficient Expectation-Maximization algorithm for its computation. We demonstrate the satisfactory performance of the proposed methods through simulation studies and provide an application to a motivating cancer study that contains substantial proportions of missing genomics data.

E0600: **Linear mixed effects models with flexible random effects and error distributions**
Presenter: **Rui Wang**, Harvard Pilgrim Health Care, United States
Co-authors: Tom Chen
In many biomedical investigations, parameters of interest, such as the intraclass correlation coefficient (ICC), are functions of higher order moments reflecting finer distributional characteristics. One popular method to make inference for such parameters is through postulating a parametric random effects model. We relax the standard normality assumptions on both the random effects and errors through the use of the Fleishman distribution, a flexible four-parameter distribution which accounts for the third and fourth cumulants. We propose a Fleishman bootstrap method to construct confidence intervals for correlated data and develop a normality test for the random effects and errors distributions. Recognizing that the ICC operates on a linear scale and may not be appropriate for wildly skewed or heavy-tailed distributions, we propose a modified, scale-free ICC. We evaluate our methods in simulation studies and apply these methods to the Childhood Adenotonsillectomy Trial sleep electroencephalogram data in quantifying wave-frequency agreement among different channels.

EO171 Room AT242 SURVIVAL ANALYSIS WITH COPULAS AND RANDOM EFFECTS Chair: Weijing Wang

E0287: **Semiparametric transformation models for left-truncated and interval-censored data without or with a cure fraction**
Presenter: **Chyong-Mei Chen**, National Yang-Ming University, Institute of Public Health, Taiwan
Interval censoring and truncation often arise in cohort studies, longitudinal and sociological research. We formulate the effects of covariates on left-truncated and mixed case interval-censored (LTIC) data without or with a cure fraction through a general class of semiparametric transformation models. We propose the conditional likelihood approach for statistical inference. For data without a cure fraction, we propose a computationally efficient EM algorithm, facilitated by a gamma-Poisson data augmentation, for obtaining the conditional maximum likelihood estimator (cMLE). For data with a cure fraction, we consider a semiparametric mixture cure model, which combines a logistic regression formula for the uncured probability with the class of transformation models for the failure time of uncured individuals. To overcome the computational complexity due to the presence of a cure fraction, we propose a novel expression for the conditional likelihood function and then create a new complete-data likelihood function. Based on this, we develop a computationally stable EM algorithm for obtaining the cMLE. We show that the cMLEs for the regression parameters are consistent and asymptotically normal. Based on the profile likelihood, we apply an EM-aided numerical differentiation method to compute the asymptotic variance estimates. We demonstrate the performance of our procedures through intensive simulation studies and application to the datasets from a cohort study.

E0683: **Nonparametric estimation of the cross ratio function**
Presenter: **Noel Veraverbeke**, Hasselt University, Belgium
For a pair (T_1, T_2) of absolutely continuous random variables, the cross ratio function is defined as the ratio of the conditional hazard rate functions of $T1$, given $T_2 = t_2$ and $T_2 > t_2$ respectively. Independence between T_1 and T_2 corresponds to cross ratio equal to 1 and positive association corresponds to cross ratio > 1. Nowadays the cross ratio function is a commonly used measure to describe local dependence between two correlated random variables. Being a ratio of conditional hazard functions, the cross ratio can be written in terms of the survival copula of T_1 and T_2 and its partial derivatives. Using Bernstein estimators for the survival copula and its derivatives, we obtain Bernstein based estimators for the conditional hazards and a nonparametric estimator for the cross ratio function. The reason for using a Bernstein copula-based estimator for the cross ratio function is motivated from earlier results showing good bias and variance properties. The asymptotic distributional behavior of the new estimator is established. We also consider a number of simulations to study the finite sample performance for copulas with different types of local dependence. A real data set on asthma attacks in children is used to investigate the local dependence between event times in the placebo and treated groups.

E0769: **A nested copula duration model for competing risks with multiple spells**
Presenter: **Ralf Wilke**, Copenhagen Business School, Denmark
Co-authors: Enno Mammen, Simon Lo
A copula graphic estimator for the competing risks duration model with multiple spells is presented. By adopting a nested copula structure the dependencies between risks and spells are modelled separately. This breaks up an implicit restriction of popular duration models such as multivariate mixed proportional hazards. It is shown that the dependence structure between spells is identified and can be estimated, in contrast to the dependence structure between competing risks. Thus, by allowing these two components to differ, the model is not identified. This is an important finding related to the general identifiability of competing risks models. Various features of the model are investigated by simulations and its practicality is illustrated by an application to unemployment duration data.

E0835: **Precision matrix estimation in random effect models**
Presenter: **Tso-Jung Yen**, Academia Sinica, Taiwan
Co-authors: Takeshi Emura
A method for estimating the precision matrix of random effects in mixed effect models is proposed. We adopt a stochastic approach to approximating the gradient of the log marginal likelihood function. This approach uses a Metropolis-Hastings algorithm to sample random effects from distributions conditional on observed information. In particular, to make the algorithm efficient, this approach uses a data-driven proposal for sampling random effects from target distributions. As a result of that, this approach can achieve high accuracy in approximating the log marginal likelihood. We apply the approach to estimate the precision matrix of the random effects in various models, including generalized linear mixed models and the Gaussian graphical model.

EO049 Room AT335 ADVANCES IN VARIATIONAL INFERENCE AND BAYESIAN COMPUTATION Chair: Robert Kohn

E0546: **Doubly geometry-informed variational Bayes**
Presenter: **Minh-Ngoc Tran**, University of Sydney, Australia
Increasingly complicated models in modern statistics have called for more efficient Bayesian estimation methods. A Variational Bayes algorithm is developed that exploits both the information geometry of the manifold of probability distribution functions and that of the manifold of the variational parameters. The information geometry of the manifold of probability distributions results in the natural gradient which is the steepest ascent on this manifold. Utilising the information geometry of the manifold of the variational parameters leads to an efficient non-linear optimization technique that takes into account the structure of the parameter space. The convergence of the proposed algorithm is theoretically guaranteed and its performance is tested on several challenging examples including deep neural networks.

E0667: **The block-Poisson estimator for optimally tuned signed pseudo-marginal MCMC**
Presenter: **Matias Quiroz**, University of New South Wales, Australia
Co-authors: Minh-Ngoc Tran, Mattias Villani, Robert Kohn, Doan Khue Dung Dang
A pseudo-marginal Markov Chain Monte Carlo (MCMC) method is proposed that estimates the likelihood using a block-Poisson estimator. The estimator is a product of Poisson estimators, each based on an independent set of random numbers used to construct the estimator. The construction allows us to update the random numbers in a subset of the blocks in each MCMC iteration, thereby inducing a controllable correlation between the estimates at the current and proposed draw in the Metropolis-Hastings ratio. This makes it possible to use highly variable likelihood estimators (which are computationally much faster) without adversely affecting the sampling efficiency. Poisson estimators are unbiased but not necessarily positive. We therefore follow a previous work and run the MCMC on the absolute value of the estimator, which we term signed pseudo-marginal MCMC, and use an importance sampling correction for occasionally negative likelihood estimates to estimate expectations of any function of the parameters consistently. We provide analytically derived guidelines to select the optimal tuning parameters for the block-Poisson estimator by minimizing the variance of the importance sampling corrected estimator per unit of computing time. We apply the block-Poisson estimator to doubly intractable problems, which are typically challenging to estimate efficiently.

E0679: **Deep compositional spatial models**
Presenter: **Andrew Zammit Mangion**, University of Wollongong, Australia
Nonstationary, anisotropic spatial processes are often used when modelling, analysing and predicting complex environmental phenomena. One such class of processes considers a stationary, isotropic process on a warped spatial domain. The warping function is generally difficult to fit and not constrained to be bijective, often resulting in 'space-folding.' We propose modelling a bijective warping function through a composition of multiple elemental bijective functions in a deep-learning framework. We consider two cases; first, when these functions are known up to some weights that need to be estimated, and, second, when the weights in each layer are random. Inspired by recent methodological and technological advances in deep learning and deep Gaussian processes, we employ approximate Bayesian methods to make inference with these models using graphical processing units. Through simulation studies in one and two dimensions we show that the deep compositional spatial models are quick to fit, and are able to provide better predictions and uncertainty quantification than other deep stochastic models of similar complexity. We also show their remarkable capacity to model highly nonstationary, anisotropic spatial data using radiances from the MODIS instrument aboard the Aqua satellite.

E0428: **Consistency of variational approximations in misspecified models**
Presenter: **Pierre Alquier**, CREST, ENSAE ParisTech, France
Variational Bayesian algorithms (VB) aims at approximating the posterior by a distribution in a tractable family. Thus, MCMC are replaced by an optimization algorithm which is orders of magnitude faster. VB methods have been applied in such computationally demanding applications as including collaborative filtering, image processing, NLP and text processing... However, despite nice results in practice, the theoretical properties of these approximations were not known until the past two years. We will review some of these recent results. We will emphasize that approximations of tempered posteriors are more robust to model misspecification. We will also discuss efficient optimization procedures for robust-VB.

E0603: **Mean field Ising models**
Presenter: **Sumit Mukherjee**, Columbia University, United States
Co-authors: Anirban Basak, Nabarun Deb, Promit Ghosal
The asymptotics of the log partition function of an Ising model on a sequence of finite but growing graphs/matrices are considered. We give a sufficient condition for the mean field prediction to the log partition function to be asymptotically tight, which in particular covers all graphs with average degree going to infinity. We show via several examples that our condition is "almost necessary" as well. As application of this mean field approach, we are able to do the following: a) Derive the asymptotics of the log partition function, b) Study consistency properties of the pseudo likelihood estimator, c) Study non degenerate limit distribution for the sum of spins.

EO257 Room AT337 RECENT ADVANCES IN STATISTICAL METHODOLOGY FOR SOCIAL SCIENCE RESEARCHES Chair: Yuejun Zheng

E0560: **Quantitative analysis on causal relation between pro-environmental behavior and consciousness based on survey data**
Presenter: **Yuejun Zheng**, Doshisha University, Japan
Awakening pro-environmental behaviors among general citizens has become into an important topic in order to realize the environmental conservation. Although many attempts on how to promote the pro-environmental behaviors through an economic stimulus have been proposed actively, such as an eco-car tax cut, a fee charge for garbage etc., the approach to the evocation of pro-environmental behaviors in daily life by improving ecology-oriented consciousness has not fully attracted attention yet. While identifying people's pro-environmental behavior by demographic attributes, the author tries to extract psychological and social factors which can influence peoples pro-environmental behaviors in their daily life for an effective environmental education based on the survey data collected at Tokyo in 2017. Results derived from data analysis have shown that peoples pro-environmental behaviors are closely associated with an ecology-oriented consciousness, recognition on environmental quality, and a high environmental satisfaction except demographic attributes such as female, senior, high education, and religious faith.

E0586: **Using perceptual tomography for balance clustering in network construction and validation**
Presenter: **Hsuan-wei Lee**, Academia Sinica, Taiwan

In social network analysis, when the information of socio-centric (i.e. whole) networks are difficult to get, researchers often use egocentric to understand the structure of network ties around an individual. In both methodological schemes, network sampling is widely used when self-reported ties are costly or could not be easily obtained. We combine network sampling approaches with third-party reporting: randomly chosen people are shown a random sample of photos from a group to which they belong and asked to group the photos according to whether or not the people in the photos are close to one another. Aggregated multiple 3rd-party data are analyzed to constructing the perceived presence or absence of individual properties and pairwise relationships, and a 3rd-party perceived network is thus built. This is a continuous work of the paper Mapping the structure of perceptions in helping networks of Alaska Natives. We generalize the parameters such as the number of clusters, the sampling size of the population, the number of photos that are shown to the sampled people and explore the theoretical implication of these change of settings. Lastly, we compare the perceptual networks and the classical self-reported social networks.

E0562: **Explaining court judgments on plaintiff's comparative negligence in vehicle related personal injury cases**
Presenter: **Han-Wei Ho**, Academia Sinica, Taiwan

In Taiwan, accident victims can recover compensation for their injuries even if they are 99% responsible for the accident. It remains a mystery, however, that how the court determines the degrees of negligence on a percentage basis to both parties claimed to be at fault. Based on district court decisions 2000-2016 for personal injuries sustained from motor vehicle collision and pedestrian-car accident claims in which the blame is shared, the effects of traffic violations on the assignment of fault are analyzed. Breaches of duties considered include driving under the influence, skipping a red light, failing to yield the right-of-way, breaking the speed limit, using a cell phone while driving, unaware of traffic in front, not wearing a helmet, jaywalking, crossing against the traffic signal and so on. Some adages like "The pedestrian has the right of way," "Bigger vehicles are always blamed in an accident, even if it was the smaller vehicles fault," are empirically tested as well.

E0469: **Impact analysis and spatial R^2 for spatial autoregressive models: Application to air pollution in China**
Presenter: **Hsuan-Yu Chang**, Peking University, Taiwan
Co-authors: Jihai Yu

Impact analysis and its asymptotic inference for spatial autoregressive models are investigated. A spatial version of coefficient of determination (R^2) to measure the model fit is also proposed. We first study the cross-section case, where various impacts are introduced to measure the interaction and feedback effects in a space dimension. We then study the spatial dynamic panel case with simultaneous spatial and dynamic feedback involved in the impacts. A R^2 is developed for spatial autoregressive models, which is usually not defined for a linear regression model with endogenous regressors. Monte Carlo results show that the impact analysis and spatial R^2 has satisfactory finite sample properties. Finally, we apply the impact analysis and the spatial R^2 to investigate how the meteorological factors and air pollutants affect PM2.5 in Chinese cities.

E0831: **Factors associated with employment status among people with disabilities: Results of a disability survey**
Presenter: **Chao-Yin Lin**, National Taipei University, Taiwan

People with disabilities are often under employed compared to those without disabilities. Factors related to unemployment or under-employment could be of personal, environmental or disability-related. The aim is to explore the levels of employment and factors associated with employment among people with physical or intellectual disabilities, by analyzing data drawn from a representative population-based survey carried out in New Taipei City. It is expected to provide suggestions for policy makers or service providers in assisting disabled people enter, return or stay in the labor market.

EO063 Room U302 COMPLEX FINANCIAL AND ECONOMETRIC DATA ANALYSIS Chair: Ray-Bing Chen

E0309: **A semiparametric estimation of value-at-risk and its applications**
Presenter: **Shih-Feng Huang**, National University of Kaohsiung, Taiwan

A semiparametric approach consisting of GARCH-type models and a generalized nearly-isotonic regression (GNIR) is proposed for Value-at-Risk (VaR) estimation. The GNIR is capable of depicting the up/down fluctuation of data automatically. An algorithm for the GNIR is proposed and its convergence property is derived. The proposed VaR estimator, denoted by NVaR, is shown to have fewer fluctuations than the VaR estimators obtained from GARCH-type methods. It also has better timely responses to market risks than the VaR estimators obtained from extreme value theory (EVT). We apply the NVaR to compute capital requirements and employ the daily indices of 13 global financial markets from 2003 to 2017 for our empirical investigation. The numerical results reveal that the NVaR is capable of satisfying the backtesting, reflecting market risks in time, and reducing the fluctuations of the VaR sequence and capital requirements simultaneously.

E0396: **Estimation under copula-based Markov mixture normal models for serially correlated data**
Presenter: **Li-Hsien Sun**, National Central University, Taiwan
Co-authors: Takeshi Emura

The estimation problem under a copula-based Markov model for serially correlated data is proposed. Owing to the fat tail feature in stock market, we select mixture normal distribution as the marginal distribution for the log return. Based on the mixture normal distribution as the marginal distribution and the Clayton copula, we obtain the corresponding likelihood function. In order to obtain the maximum likelihood estimators, we apply the Newton-Raphson method. In the empirical analysis, the stock price of Dow Jones Industrial Average is analyzed for illustration.

E0402: **Two-stage nonparametric estimation of multivariate densities**
Presenter: **Juan Lin**, Department of Finance & WISE, Xiamen University, China
Co-authors: Ximing Wu

In addition to slower convergence, nonparametric multivariate density estimation is plagued by the increasingly difficult specification of tuning parameters as dimension increases. We propose a two stage estimator of a d-variate density f that starts with a standard kernel density hat f_0. This pilot estimate can be refined via a multiplicative correction that estimates the density ratio $r = f/\hat{f}_0$, a challenging task due to its random denominator. Instead of direct estimation of r, we use a basis expansion to approximate $\log r$, which is achieved via minimizing the Kullback-Leibler discrepancy between f and $\hat{f}_0$ subject to moment conditions associated with the basis functions. Thanks to the consistency of the pilot estimate, $\log r$ resides in a shrinking d-sphere around origin such that all coefficients of its basis approximation tend to zero as sample size increases. Thus in our penalized spline estimation of $\log r$, it suffices to use a single penalty parameter across all dimensions, effectively alleviating the curse of dimensionality. We derive the local and global convergence of this density estimator in terms of mean squared error and Kullback-Leibler discrepancy respectively. We use Monte Carlo simulations to demonstrate its good finite sample performance and present two real data examples.

E0575: **An interpretable sparse estimator for large-scale network autoregressive models**
Presenter: **Simon Trimborn**, National University of Singapore, Singapore

The era of Big Data requires the discovery of the essential underlying structure. Sparsity estimators imposing structures have the potential to detect (describe) the dynamic dependence while bringing an interpretable system with them. We propose an interpretable sparse estimator which restricts the model for the Lag, Network and individual effects. We derive the theoretical properties and investigate the numerical performance of the estimator in extensive simulations. The sparsity operator facilitates a higher accuracy than alternative ones and simultaneously being fast in computation. We show the applicability of the estimator on a huge network of cryptocurrency pricing series.

E0577: **Demographic distribution and bond pricing: A semi-parametric affine arbitrage-free yield curve model**
Presenter: **Linlin Niu**, Wang Yanan Institute for Studies in Economics, Xiamen University, China
Co-authors: Zongwu Cai, Jiazi Chen

Life cycle consumption and savings decision in individuals determine the aggregate savings and interest rate level as the demographic structure changes along time. We build a semi-parametric affine arbitrage-free yield curve model, to incorporate a slow moving determinant of interest rate levels explained by the evolving age distribution and its expectation. Estimating the model with US data from 1950s to present, we find the extracted long term determinant captures very well the smooth trend in US yields. The business cycle component is well explained by a VAR with macro or latent variables. It produces better forecast performance than a traditional autoregressive yield curve model without time-varying means or with simple demographic ratio as explanatory variable.

EC350 Room U301 CONTRIBUTIONS IN FINANCIAL MODELLING AND QUANTITATIVE FINANCE Chair: Malika Hamadi

E0750: **Yield curve volatility and macroeconomic risks**
Presenter: **Anne Hansen**, University of Copenhagen and Danmarks Nationalbank, Denmark

Understanding the joint dynamics of the yield curve and macroeconomic variables is important for policy makers and risk managers. While these dynamics are well-studied in terms of levels, less is known about the interactions between yield curve volatility and the macroeconomy. We bridge this gap by proposing a tractable term structure model that allows macroeconomic risks to determine both bond yields and their conditional variances. To begin with, our model can match U.S. Treasury bond yield levels and realized variances when model parameters are estimated based on yield data alone. Including measurements of inflation and real activity improves the ability of the model to capture yield curve volatility. By means of regressions, we provide further evidence that macroeconomic risks predict realized variances over and above the information contained in the yield curve itself. Key to the success of our model is a generalized stochastic discount factor that allows investors to price both level and variance risks. We show how macroeconomic risks relate to these risk preferences.

E0802: **Probit or logit model for the regulatory capital and IFRS9 calculations**
Presenter: **Jiri Witzany**, University of Economics in Prague, Czech Republic

The Probit model based on the normal distribution has become a workhorse of the Basel II internal rating based (IRB) regulatory capital and IFRS9 expected credit loss (ECL) calculations. On the other hand, the Logit (scoring) model has become an industrial standard in credit risk modeling in banks and financial institutions. There are many arguments why Logit should be preferred with respect to the Probit model. We propose and analyze how to replace the Probit assumptions with Logit in the IRB formula and in IFRS9 calculations. In an empirical study, we will demonstrate substantial differences in default correlation estimations conditional on the two models and the impact on regulatory capital and ECL numbers. The Logit model yields generally more conservative estimates that would have, in our opinion, a significant impact on capital adequacy ratios of the overall banking sector.

E0764: **Does it pay to follow anomalies research: A machine learning approach with international evidence**
Presenter: **Martin Hronec**, Czech Academy of Sciences, Czech Republic
Co-authors: Ondrej Tobek

Out-of-sample returns on 153 anomalies in equities documented in academic literature are studied. We show that machine learning techniques that aggregates all the anomalies into one mispricing signal are 4 times more profitable than a strategy based on individual anomalies and survive on a liquid universe of stocks. The machine learning also leads to 2 times larger Sharpe ratios with respect to the corresponding standard finance methods. We next study value of international evidence for selection of quantitative strategies that outperform out-of-sample. Past performance of quantitative strategies in the regions other than the US does not help to pick out-of-sample winning strategies in the US. Past evidence from the US, however, captures most of the predictability within the other regions. The value of international evidence in empirical asset pricing is thus very limited.

E0522: **Modified compound binomial risk model with by-claims and randomized dividends**
Presenter: **B S Aparna**, Indian Institute of Technology Madras, Chennai., India
Co-authors: Neelesh Shankar Upadhye

Consider the discrete time compound Binomial risk model with by-claims and randomized dividends. Let V be an indicator random variable representing the issuance or non-issuance of a dividend. A dividend is issued with probability $\mathbb{E}(V)$ whenever the surplus exceeds a non negative threshold d. Main claims are assumed to induce by-claims. Also, by-claims are settled at most by the next time period. We assume that the claim and by-claim probabilities follow a Beta distribution. Thus, the end is to investigate the behaviour of the discrete time compound Beta-Binomial risk model with by-claims and randomized dividends. We derive recursive expressions for the discounted conditional expected penalty function, the probability of ruin, the distribution function of the deficit at ruin, the generating function of the deficit at ruin and the probability of the surplus prior to ruin.

E0817: **Price endogeneity of the cryptocurrency market**
Presenter: **Jan Sila**, Univerzita Karlova, Czech Republic
Co-authors: Michael Mark

A novel source of high-frequency data is explored by the means of temporal point processes. We propose a non-parametric mutually exciting Hawkes process to examine the inner and cross dependencies of transactions for four major cryptocurrencies on the BitMEX exchange. As these instruments are considered being highly volatile and sensitive to news releases, we enrich the model with a kernel matrix representing the news impact. Thus introducing an original extension of the Hawkes process by combining news and mutual excitations of the series. This allows for a realistic time-dependent measure of the degree of endogeneity in the market.

EC348 Room U501 CONTRIBUTIONS IN APPLIED ECONOMETRICS II **Chair: Zinoviy Landsman**

E0748: **Identification and estimation of a search model: A procurement auction approach**
Presenter: **Daniel Silva Junior**, City, University of London, United Kingdom
Co-authors: Fabio Sanches, Sorawoot Srisuma, Mateusz Mysliwski

A non sequential search model with a continuum of consumers and a finite number of firms is proposed. Both consumers and firms are heterogeneous. Consumers differ in search costs. Firms have private marginal costs of production. We show that an equilibrium price dispersion can arise in this model as firms employ a Bayesian Nash pricing strategy. We provide conditions to identify the model using price and another supply side data (such as market share). Our identification strategy is constructive. We derive the uniform rate of convergence of our estimator.

E0794: **The generalised hyperbolic distribution and its subclass in the analysis of a new era of cryptocurrencies: Ethereum**
Presenter: **Stephen Chan**, American University of Sharjah, United Arab Emirates

Ethereum was the first decentralised platform to support smart contracts. It has attracted significant publicity and captured the interests of a wide range of institutions, enthusiasts and even world leaders. We have analysed the market price index for all exchanges trading in Ethereum versus three global currencies, the Korean Won; Euro; US Dollar; Bitcoin, and the Global Price Index for Ethereum, through the fitting of the Generalised Hyperbolic distributions and its subclasses. Our results show that returns are clearly non-normal and the Generalised Hyperbolic and its subset of distributions fit well jointly for all of the indices. We also analyse the long term memory effect for the returns of Ether, compare the Value at Risk and Expected Shortfall based on historical Ether data and other financial instruments, and backtesting was also performed to test the extreme tails.

E0754: **Local scaling in marked point process: On the influence of a company's health on another company**
Presenter: **Ilona Berkova**, University of South Bohemia in Ceske Budejovice, Czech Republic
Co-authors: Tomas Mrkvicka

The localization of corporate activities is the most important decision for new company and is the result of many factors. Due to this fact, the location of firms has been a part of economics for many years. The aim was to estimate an econometric model to recognize the spatial relationships of the firms according to their health in inhomogeneous space. The data set consists of firms position in 3 regions in the Czech Republic whose health was determined according to the Neumeiers index IN05. The dependence of marks in a point pattern is usually expressed through the marked L-function. The function can be extended for inhomogeneous processes with second-order reweighted stationarity assumption through weighting by intensity. Since the behavior of firms in a distance r depends also on population density or density of firms, the appropriate inhomogeneous tool for the firms environment is local scaling. The local scaling assumes larger distances between firms in a less populated area and transforms the space according to a function. Therefore, the local scaled marked L-function that expresses dependence between the firms' health in transformed distance r is defined. There was applied a global envelope test with the null hypothesis of no dependence between the firm health. It allows deciding if the near firm with good health influences the health of the firm and if the near firm with whatever health influences the health of the firm.

E0810: **The adaptive market hypothesis in the high frequency cryptocurrency market**
Presenter: **Jeffrey Chu**, Universidad Carlos III de Madrid, Spain

The aim is to provide evidence of the adaptive market hypothesis (time-varying predictability) in the markets for the two largest cryptocurrencies: Bitcoin and Ethereum. High frequency hourly prices and returns of the two cryptocurrencies versus the Euro and US Dollar were analysed, over a nine-month period from December 2017 to August 2018, and the martingale difference hypothesis was tested to determine market efficiency/inefficiency. We conducted the consistent and integrated test proposed previosly, and a rolling window approach using the previous 168 hours (7 days) of data was implemented to investigate how the market efficiency of the cryptocurrency markets varied over time. Our results suggest that the efficiency of the markets for both Bitcoin and Ethereum are not constant but vary over time, and we also show that the level of efficiency may correspond to and be influenced by positive and negative news or events. This is the first time that the adaptive market hypothesis has been investigated in a high frequency setting for the cryptocurrency market, and we believe that the results would be beneficial as further inputs for modelling cryptocurrency prices and returns but also for investors seeking suitable indicators of when to buy or sell cryptocurrencies.

E0249: **The impact of health insurance reforms on childrens educational attainment: An evidence from Vietnam**
Presenter: **Khiem Phuong Huu**, Feng Chia University, Vietnam
Co-authors: Yu-Chen Kuo

Research has shown that parental health shocks and child health status each exert measurable effects on child educational attainment, particularly in low-middle income countries. In 2005, the Vietnamese government enacted a new health insurance policy increasing the proportion of population covered by health insurance from 22% of total population 43%. Using a quasi-experimental setup and difference-in-differences (DID) approach, the effects of health insurance reforms on child educational outcome are examined. Because households in the state sector were almost unaffected before and after the reform, children in that group served as a natural control group, children growing up in non-state employed households formed a treatment group. Educational outcomes were measured for three levels of general education: primary, secondary and high school. Results showed that the NHI reform improved educational outcomes for children in high school, both in terms of enrolment and school completion likelihood. Furthermore, it was shown that children from minority groups, females, those in rural areas, and those from poorer families were less likely to derive the same educational outcomes compared to their counterparts. These findings are the first of their kind using the VHLSS survey data and would be of value to policy makers in countries that plan to adopt a similar health policy.

EC340 Room U502 CONTRIBUTIONS IN STATISTICAL MODELLING **Chair: Hong-Dar Wu**

E0709: **mpcmp: Mean-parametrized Conway-Maxwell-Poisson regression**
Presenter: **Thomas Fung**, Macquarie University, Australia
Co-authors: Justin Wishart, Alan Huang, Aya Alwan

Conway-Maxwell-Poisson (CMP) distributions are flexible generalizations of the Poisson distribution for modelling overdispersed or underdispersed counts. The main hindrance to their wider use in practice seems to be the inability to directly model the mean of counts, making them not compatible with nor comparable to competing count regression models, such as the log-linear Poisson, negative-binomial or generalized Poisson regression models. We review how CMP distributions can be parametrized via the mean, so that simpler and more easily interpretable mean-models can be used, such as a log-linear model. Moreover, we introduce the R package: mpcmp which provides a collection of functions for estimation, testing and diagnostic checking for the proposed model. The performance of the R routine against the earlier proposed MATLAB routine will also be discussed.

E0739: **A control chart using a copula-based Markov chain for attribute data**
Presenter: **Xin-Wei Huang**, National Central University, Taiwan
Co-authors: Takeshi Emura

Statistical process control (SPC) is a fundamental tool in industrial manufacturers, financial engineers, medical researchers, and others. The application of copula-based Markov chain to SPC is a recent approach, where a copula can capture serial dependence between observations. So far,

only continuous data are considered to perform SPC under the copula-based Markov chain model. We propose a SPC method under the copula-based Markov chain model for attribute data that follow the binomial distribution. We develop methods to compute the maximum likelihood estimator and control limits that are necessary to draw an attribute control chart. We also develop simulation algorithms to generate dependent attribute data that can be used to compute the average run length of the proposed control chart. Furthermore, we propose a goodness-of-fit method and a copula selection method. We conduct simulation studies to check the accuracy of the proposed estimator and to compare our method with other methods. We demonstrate the proposed method by analyzing the Korean stock market data. We implement the proposed methods in the R Copula.Markov package so that users can easily apply the proposed methods.

E0183: **Ordinal response and ordinal predictors: A regression model and monotonicity direction classification of effects**
Presenter: **Javier Espinosa**, University College London, United Kingdom
Co-authors: Christian Hennig
A regression model is proposed for the analysis of an ordinal response variable depending on a set of multiple covariates containing ordinal and potentially other variables. The proportional odds cumulative logit model is used for the ordinal response, and constrained maximum likelihood estimation is used to account for the ordinality of covariates. Ordinal predictors are coded by dummy variables. The parameters associated with the categories of the ordinal predictor(s) are constrained, enforcing them to be monotonic (isotonic or antitonic). A decision rule is introduced for classifying the ordinal predictors' monotonicity directions, also providing information whether observations are compatible with both or no monotonicity direction. In addition, a monotonicity test for the parameters of any ordinal predictor is proposed. The monotonicity constrained model is proposed together with five estimation methods and compared to the unconstrained one based on simulations. The model is applied to real data explaining a 10-Points Likert scale quality of life self-assessment variable by ordinal and other predictors.

E0784: **Quantile regression approach to conditional mode estimation**
Presenter: **Hirofumi Ohta**, The University of Tokyo, Japan
Co-authors: Kengo Kato, Satoshi Hara
The estimation of the conditional mode of an outcome variable given regressors is considered. To this end, we propose and analyze a computationally scalable estimator derived from a linear quantile regression model and develop non-standard asymptotic distributional theory for the estimator. Specifically, we find that the limiting distribution is a scale transformation of Chernoff's distribution, i.e., the distribution of a maximizer of a two-sided Brownian motion with a negative quadratic drift, despite the presence of regressors. In addition, we consider analytical and subsampling-based confidence intervals for the proposed estimator. We also conduct Monte Carlo simulations to assess the finite sample performance of the proposed estimator together with the analytical and subsampling confidence intervals.

E0732: **Modelling over-dispersion in price jumps arrivals: A comparison between Poisson mixtures and linear Hawkes model**
Presenter: **Ping Chen Tsai**, Southern Taiwan University of Science and Technology, Taiwan
Price jumps may display some degree of clustering and this feature is typically manifested through a phenomenon known as over-dispersion in count data. The time series of counts will show a variance larger than its mean, and hence rejects a Poisson null hypothesis. Two competing models, namely, mixtures of Poissons and linear Hawkes process, are shown to be able to reproduce the over-dispersion feature, but the former by definition has a clustering parameter equal to zero, whereas the latter has a strictly positive clustering parameter. Different versions of the two models are fitted to price jumps data and the estimation results compared. Special attention is paid to the base intensity of linear Hawkes process, and whether one of the components in mixtures of Poisson corresponds to this base intensity. The issue of many zeros in the count data is also addressed using a zero-inflated model.

EC341 Room U517 CONTRIBUTIONS IN TIME SERIES Chair: Roderick McCrorie

E0743: **Refined RBFN test for martingale difference hypothesis: Application to predictability of exchange rate returns**
Presenter: **Jinu Lee**, King's College London, United Kingdom
A regression-typed test for a martingale difference hypothesis (MDH) based on a radial basis function network (RBFN) is revisited. New testing procedures are discussed to construct more rigorous RBFN specifications. A Monte Carlo experiment is conducted to show that the new test has improved finite sample properties in terms of size and power. Further, the proposed method is applied to examine time-varying predictability of major foreign exchange rates from January 1999 to April 2015 with a moving time window of nearly two years at a daily frequency. The empirical results confirm that the exchange rate returns are not consistently unpredictable or weak-form efficient with continued fluctuations. There are statistically significant evidences to more often deviate from the martingale behaviour since the recent global financial crisis. The findings may be in line with an implication of the adaptive market hypothesis due to market condition changes.

E0814: **Estimation for ARMA models with t-distributed innovations**
Presenter: **Haruhisa Nishino**, Hiroshima University, Japan
The standard ARMA model assumes that its innovations are white noise processes with 0 mean and a constant variance. That is, the ARMA model is a second-order stationary process characterised by its autocovariance function. The white noise process has no information about its fourth-order moment. To estimate the ARMA model, we assume a Gaussian process and a Gaussian likelihood. On the other hand, the literature of financial time series tells that the financial returns have fatter tails than Gaussian ones. The Student t-distribution is a typical fat-tailed distribution. The fat-tailed property is related to the fourth order moment. We thus consider that the estimation for ARMA models with t-distributed innovations is useful for analysing financial time series. If we know the degrees of freedom of the t-distribution of the model, can we estimate the parameters of the ARMA model more efficiently than the Gaussian likelihood? Besides, the talk proposes a preliminary estimate for the degrees of freedom based on the method of moments, since the application of MLE for the degrees of freedom has a severe problem.

E0786: **Semi-parametric estimation of multivariate possibly non-causal and possibly non-invertible time series models**
Presenter: **Bernd Funovits**, University of Helsinki, Finland
A semi-parametric two-step estimation strategy is proposed for possibly non-causal vector autoregressive (VAR) systems and possibly non-invertible vector autoregressive moving average (VARMA) systems driven by non-Gaussian i.i.d. shocks. First, we obtain an initial estimate based on second moment information (e.g. a Gaussian quasi maximum likelihood estimator). Subsequently, multivariate all-pass filters are used to generate the finite set of systems with the same second moment properties, i.e. the same spectral density, but with different determinantal AR or MA roots. Last, we use an objective function based on the Fourier transform of the third and fourth order cumulants, i.e. the bi- and trispectral density, of the residuals to identify the true root and pole locations among the set of models with identical second moment properties. The obtained estimator can serve as an initial estimate for maximum likelihood estimation.

E0797: **Generalized hypergeometric series allied to moments in the four-step uniform random walk problem**
Presenter: **Roderick McCrorie**, University of St Andrews, United Kingdom
The focus is on the representation of odd moments of the distribution of a four-step random walk in even dimensions, which are based on a mathematical constant representable in terms of the integral over the unit interval of the square of the complete elliptic integral of the first kind. New symmetries in critical values of the L-series of two underlying cusp forms are established, unblocking the problem of representing the constant

in terms of very well-poised generalized hypergeometric series and revealing it has a formal counterpart. Both the constant and its counterpart are seen already to play significant roles in multidisciplinary contexts. A connection is made to the open econometric problem of fully characterizing the bias in the canonical autoregressive model under the unit root hypothesis, which shares similar features.

E0790: **Some properties of the quantile regression version of Hodrick-Prescott filtering**
Presenter: **Hiroshi Yamada**, Hiroshima University, Japan
The quantile regression version of Hodrick-Prescott filtering enables us to estimate not only the median trend, which is more robust to outliers than HP filtering, but also other quantile trends, which may provide a deeper understanding of time series properties. Some properties of the filtering are reported.

EP001 Room Hall POSTER SESSION Chair: Elena Fernandez Iglesias

E0217: **A Bayesian smoothing spline ANOVA model to re-examine the effects of an intervention on infant massage**
Presenter: **Chin-I Cheng**, Central Michigan University, United States
An intervention on infant massage called the M technique, which is a structured manual massage technique to reduce stress, was administered to nine hospitalized, very preterm infants. The outcome variables are repeated measures on physiological (heart rate, RR -respiratory rate-, and SaO2 -oxygen saturation) and behavioral (ABSS -Anderson Behavioral State Scale) status during a period of time. The outcome measurements from the intervention were published based on results from repeated measure ANOVA. The outcome measurements are re-examined with the Bayesian smoothing spline ANOVA model. Smoothing spline ANOVA is an approach to semiparametric function estimation based on an ANOVA type of decomposition. This model allowed testing for effects which can be linear or nonparametric (i.e., smooth or interactions between selected linear and smooth effects). The ability to test for these effects provides insights of the cumulative impact of the M technique intervention. The Bayesian approach was considered, and the Bayes factor was used for variable selection and hypothesis testing. Findings showed that the intervention had a cumulative impact on heart rate, SaO2 and that for ABSS effect is curvature (nonlinear).

E0714: **A new test statistic to assess the goodness of fit of exponential distribution under multiply progressive censoring**
Presenter: **Kyeongjun Lee**, Daegu University, Korea, South
Co-authors: Subin Cho, YeongEun Hwang, Seonghee Park
The exponential distribution is the probability distribution that describes the time between events in a Poisson point process, i.e., a process in which events occur continuously and independently at a constant average rate. The goodness of fit test for the exponential distribution is very important in lifetime data analysis. Therefore, we propose the two test statistics to test goodness of fit for the exponential distribution under multiply progressive type II censoring scheme. Also, we propose a new graphic method for the goodness-of-fit test for the exponential distribution under multiply progressive type II censoring scheme. We assess the new test statistic in terms of the power of the test through by Monte Carlo method. And we check the new plot and test statistic by using real data.

E0716: **Maximum product of spacings estimation based on inverse Weibull distribution under progressive censoring**
Presenter: **Minyeong Han**, Daegu University, Korea, South
Co-authors: Kyeongjun Lee
Progressively type II censoring schemes have become quite popular in reliability and lifetime-testing studies. We estimate the maximum product of spacings estimators (MPSE) of Inverse Weibull Distribution (IW) based on progressively type II censored data. Asymptotic confidence intervals are also proposed. Bayes estimates and credible intervals of the unknown parameters are obtained under the assumption of gamma priors on the unknown parameters. Different methods are compared using Monte Carlo simulations. One real data set is analyzed for illustrative purposes.

E0727: **On high-dimensional cross-validation and accumulated prediction error**
Presenter: **Wei-Cheng Hsiao**, Soochow University, Taiwan
Co-authors: Ching-Kang Ing, Wei-Ying Wu
Cross validation (CV) has been one of the most popular methods for model selection. By splitting n data points into a training sample of size nc and a validation sample of size nv in which nv/n approaches 1 and nc tends to infinity, it has been shown that subset selection based on CV is consistent in a regression model of p candidate variables with $p << n$. However, in the case of $p >> n$, not only does CV's consistency remain undeveloped, but subset selection is also practically infeasible. Instead of subset selection, we suggest using CV as a backward elimination tool for excluding redundant variables that enter regression models through high-dimensional variable screening methods such as LASSO, LARS, ISIS, and OGA. By choosing a nv such that nv/n converges to 1 at a rate faster than the one suggested previously, we establish the selection consistency of the proposed method. Accumulated prediction error (APE), on the other hand, can be viewed as a counterpart of CV in situations where a random split of data is pointless (e.g., when data are serially correlated). While APE's behavior in the case of $p << n$ has been well understood, no results have been reported regarding its performance in the high-dimensional case. To fill this gap, we provide a high-dimensional amendment of APE and justify its asymptotic validity. Simulation evidence will also be presented.

E0735: **Variational Bayesian approaches for Structure Selection**
Presenter: **Lai Wei-Ting**, National Cheng Kung University, Taiwan
The problems of group structure selection in high-dimensional regression models are considered. The variables or regressors are partitioned into different group and only a few groups are active or important. Therefore, it is an interesting issue finding important groups. The Bayesian approach based on the spike and slab priors for the regression coefficients is considered, and each candidate group is associated with a binary variable indicating whether the group is active or not. Instead of Markov Chain Monte Carlo (MCMC) methods, a fast and scalable vibrational Bayesian approach is introduced for the posterior inference. Furthermore, we extend the proposed method for the multi-task learning and the structure selection problem in vector autoregression (VAR) model. Simulation studies and real examples are used to demonstrate the performances of the proposed Bayesian approaches.

E0744: **Multi-level mixed-effects model analysis for reproductive performance of high prolific but low longevity sows**
Presenter: **Yuzo Koketsu**, Meiji University, Japan
Prolificacy and longevity are critical on commercial pig farms. Our objective was to examine the impact of piglets born alive (PBA) in parity 1 on early removal of sows from farms. The PBA groups were categorized based on the upper and lower 10th percentiles of PBA in parity 1: High-prolific, ordinary, and Low-prolific sows. Also, early removal of sows were defined as culling and death of sows in parities 1-3. Data comprised 411,360 parity records of 153,523 sows farrowed between 2011 and 2017 on 155 Spanish farms. Three-level mixed-effects models were applied to the data by using a farm at level 3, a sow at level 2 and a parity record at level 3. Removal rates by parity 3 were 27.9, 28, and 39.2% for High-prolific, ordinary, and Low-prolific sows respectively. There were 2-way interactions between PBA groups and removal groups in farrowed sows for PBA and piglets weaned in parity 1 ($P < 0.01$). Low-prolific sows that were removed had 0.8-1.6 fewer PBA and 0.3-0.8 fewer piglets weaned than survived sows ($P < 0.05$), but there were no such differences for High-prolific sows ($P > 0.05$). In conclusion, PBA in parity 1 affects early removal of farrowed sows.

E0747: **Efficient Bayesian estimation for the space-time stationary condition with blocked sampling approach**
Presenter: **Yoshihiro Ohtsuka**, Tohoku Gakuin University, Japan

An efficient posterior sampling algorithm is proposed for the spatial dynamic panel data model from the viewpoint of the Bayesian inference. The stationary condition of this model mutually depends on three parameters such as simultaneous and lagged spatial dependencies, and serial correlation. Thus, the interdependence between these parameters yield law convergence to their target marginal distribution. To accelerate sampling efficiency, Bayesian estimation algorithm for these parameters is developed by using a Tailored approximation and blocked Metropolis-Hastings (TaB-MH) algorithm. With the respect to an inefficient factor, the TaB-MH algorithm is superior to the random walk MH in the experimental studies and empirical analysis on regional data.

E0755: **On efficient designing of nonlinear experiments for 3D Stacked Dies Packages**
Presenter: **Tzu-Lun Yuan**, National Sun Yat-sen University, Taiwan
Co-authors: Mong-Na Lo Huang, Yu-Jung Huang

3D stacked die technology can be employed to achieve an ultrahigh mobile data rate on fifth generation (5G) wireless systems. AC Coupled Interconnect (ACCI) can enable low-cost, high-density, high pin count, and achieve a high speed chip to chip communications scheme. The placement of corner differential signal pad has a great impact on the system performance. The corner differential signaling pad placement arrangements under different pad sizes are analyzed to examine the behaviors of the SNR values versus different frequencies for pad design with given pitch distance and overlapping percentage. We provide nonlinear parametric models under different pad sizes to exhibit the patterns of the SNR values versus the frequencies for various die placement designs with given pitch distance and overlapping percentage. In the end, we use a sequential approach to find the optimal design configuration for each pad size with an optimal response.

E0757: **Approximate maximum product spacing estimation of the half logistic distribution based on multiply progressive censoring**
Presenter: **Dayoung Kang**, Daegu University, Korea, South
Co-authors: Kyeongjun Lee, Gaeun Lee

The problem of estimating unknown parameter of a half logistic distribution is considered on the basis of multiply progressive censoring. The unknown parameter of half logistic distribution is estimated by the maximum likelihood estimation (MLE) and maximum product spacing estimation (MPSE) using Newton-Raphson method. Also, we obtain the approximate maximum product spacing estimation (AMPSE) of the unknown parameter for half logistic distribution under multiply progressive censoring. We compare the estimators in the sense of the mean square error and bias. Finally, real data sets are analyzed for the purpose of illustration.

E0761: **Copula for spatial dependence of rice production in Korean peninsula**
Presenter: **Sanghoo Yoon**, Daegu University, Korea, South
Co-authors: Hunseok Park, Huijae Lee

The purpose is to estimate the spatial dependence of rice production yield among eight provinces in South Korea. Because the large scale of the disaster, such as typhoon, drought, and flood, can affect many rice yields. The optimal copula function for a spatial dependence of rice yields in Korean peninsula has not been studied. The data was collected between 1965 and 2018 by Statistics Korea. The trend of rice production has been eliminated by a linear model. The marginal distribution of residual was estimated by maximum likelihood estimation. A pseudo data was generated by inverse transformation method. Elliptical copula, Archimedean copula, and extreme copula family were considered. The goodness of fit test and cross-validation were performed to find the best copula. As a result, we found that rice production in most provinces has a strong correlation with each other.

E0762: **Control charts for monitoring the Weibull parameters under progressive hybrid censored data**
Presenter: **Jaeyoung Gwag**, Daegu University, Korea, South
Co-authors: Kyeongjun Lee, Nanhee Yun

The control chart is an important tool of statistical process control to continuously improve the production quality in order to maintain the repute of the produced items in the competitive markets. A new reliability testing control chart for monitoring the Weibull parameters under progressive hybrid censored data has been developed. Also, we provide a comparison between these control charts in terms of the out-of-control average run length obtained by simulation for the unknown parameter. A real data from breaking stress of carbon fibres is presented for illustration and comparison of the proposed control charts.

E0778: **Fay-Herriot model with functional measurement error of outcome variable and covariate**
Presenter: **Jinseub Hwang**, Daegu University, Korea, South
Co-authors: Chaeyeong Kang, Do Hyang Kim, Soo Rack Ryu

Various extension versions for Fay-Herriot model have been proposed. However, most studies have considered only covariate measurement errors, and the outcome variable may also have measurement errors. We propose an extended Fay-Herriot model that can reflect the measurement error of dependent variable and covariate. We consider a functional measurement error model that assumes a non-stochastic true value of the outcome variable and covariate. To fit the model and estimate parameters, we consider hierarchical Bayesian model approach based on Markov chain Monte Carlo method. To check the superiority of the proposed model, we two linear functions in simulation studies, and we use the seventh KNHANES (Korean national health and nutrition examination survey) data and 2010 Italian household budget survey in the application. As a result for simulation studies and application, the performance of the proposed model is better based.

E0781: **Extending mixtures of t linear mixed-effects models with logistic function of covariates**
Presenter: **Yu-Chen Yang**, National Chung Hsing University, Taiwan
Co-authors: Tsung-I Lin, Luis Mauricio Castro, Wan-Lun Wang

The issue of model-based clustering of longitudinal data has attracted increasing attention in past two decades. Finite mixtures of Student's-t linear mixed-effects (FM-tLME) model have been considered for implementing this task especially when data contain extreme observations. An extended finite mixtures of Student's-t linear mixed-effects (EFM-tLME) model is presented, where the categorical component labels are assumed to be influenced by the observed covariates. As compared with the naive methods, assuming the mixing proportions to be fixed but unknown, our proposed EFM-tLME model exploits a logistic function to link the relationship between the prior classification probabilities and the covariates of interest. To carry out maximum likelihood estimation, an alternating expectation conditional maximization (AECM) algorithm is developed under several model reduction schemes. The technique for extracting the information-based standard errors of parameter estimates is also investigated. The proposed method is illustrated using simulation experiments and real data from an AIDS clinical study.

E0804: **On goodness-of-fit test for beta-binomial regression models**
Presenter: **Junghwan Kim**, Institute of Water Resources System, Inha University, Korea, South
Co-authors: Woojoo Lee

Beta-binomial regression models have been widely used for analyzing over-dispersed binomial data. However, they are not always flawless models that can capturing all types of extra-binomial variation. Like other regression models, statistical inferential procedures from beta-binomial

regression models can be problematic if an important co-variate is not included or some heterogeneity is shown in individuals. To test the existence of the hidden heterogeneity in beta-binomial regression models, we provide score test statistics and investigate their performance in terms of type I error and power.

E0805: **On the relationship among different statistical methods for dynamic treatment regimes**
Presenter: **Seung Jae Lee**, Inha University, Korea, South
Co-authors: Woojoo Lee
For each patient, medical doctors need to determine the best treatment among available options based on patient's information. Especially, for chronic diseases such as hypertension and diabetes, they need to determine the best sequence of treatments using the updated information including the past outcomes affected by the past treatment. This topic has been studied under the name of optimal dynamic treatment regime. So far, various statistical methods haven been proposed for finding optimal dynamic treatment regime. Q-learning, A-learning and O-learning are those examples. In addition, in epidemiology literature, g-computation and inverse probability treatment weighting methods for time-varying treatment have been considered for a similar purpose. However, their relationship is not clearly recognized in literature. We examine the relationship among different statistical methods for discovering optimal dynamic treatment regime and discuss their advantages and disadvantages in terms of computation.

E0816: **Topic analysis of statistical journals: A vector space model approach**
Presenter: **Charlene Mae Celoso**, University of the Philippines Diliman, Philippines
From its early stages up to today, the field of statistics has gone through several developments. With increased availability of computational tools, the problems that statisticians are able to tackle have changed through the years. Abstracts of articles in several prominent statistics journals are analyzed via a vector space model approach to uncover underlying concepts that have emerged in different time periods. From such models, visualizations can be created using semantic maps to discover topics that are grouped similarly.

E0780: **The socioeconomics status factors of framingham risk scores based on national survey data in Korea**
Presenter: **Donghee Kim**, Daegu University, Korea, South
Co-authors: Jinseub Hwang, Hyerim Han
The aim is to identify the socioeconomics status factors of framingham risk scores to be used for prediction the risk of future cardiovascular disease. We use the 7th National Health and Nutrition Survey data in South Korea. Of the total 16,277 subjects, we select 583 subjects who aged between 20 and 59 years except the subjects who variables considered in the analysis is missing. Considering the strata, cluster and weight according to the complex survey design, we conduct a descriptive analysis and multiple linear regression for framingham risk scores. As a result, education level and marital status are significant factors for framingham risk scores. The lower the education level and unmarried status tend to have the higher scores. In order to reduce the prevalence of cardiovascular disease, the management of factors included in the calculation of framing risk scores is important, but interest in subjects with low socioeconomic status that may affect the framingham risk score will also be needed.

E0840: **Identification of pairs trading opportunities using copula-based conditional probabilities and machine learning models**
Presenter: **Ting Hin Cheng**, The Open University of Hong Kong, Hong Kong
Co-authors: Ka Lok Li, Chun Lam Wong, Suet Ying Lau, Chak Long Ng, Carlin Chu
Copula is a multivariate probability distribution function that allows a separate estimation of marginal and joint distributions for modeling relationships between variables. It is flexible to capture non-symmetric relationships between the tails and this makes it a superior candidate for financial applications. In this study, the conditional probabilities of copula are modeled as input features for the identification of pairs trading investment opportunities. Apart from the usage of traditional logistic regression model, 3 machine learning models (Neural Network, AdaBoost and Random Forest) are employed to explore the corresponding contributions under various settings. High frequency (1-minute interval) trading prices during the period of October 2018 to February 2019 are collected from Bloomberg terminal for carrying out the empirical analysis of this study. To come up with a more comprehensive picture, several distinct settings (i.e. under-sampling, cross-validation, early stopping, dropout and grid search of hyper parameters) are investigated in this study. Our empirical results show that logistic regression of copula-based features does not work well. However, the generated features help to produce favorable performances when they are used together with sophisticated machine learning models. The AdaBoost model produces the highest ROC index while the 5-layer Neural Network model delivers the largest profit.

E0843: **The role of littoral processes on the sand movement in the Northern Spanish coast by fitting regression models**
Presenter: **Elena Fernandez Iglesias**, University of Oviedo, Spain
Co-authors: Ana Belen Ramos-Guajardo, Gil Gonzalez-Rodriguez, Jorge Marquinez
Coastal sand dune ecosystems are facing threats and pressures across Europe, but the quantification of the exact impacts is not easy, particularly when natural and anthropic effects are overlapping. In order to tackle this question, during two years sand movement has been measured in pilot dune plots located in the Northern Spanish coast. We aimed at correlating the waves height, wind velocity and tidal range with the temporal changes measured in the beach-dune systems. The role of littoral processes on the sand movement is quantified by fitting different regression models using R, in order to discuss the influence of marine variables, climate change and human activities in the changes identified in coastal sand systems, which also affects Spanish coasts.

Thursday 27.06.2019 14:00 - 15:40 Parallel Session L – EcoSta2019

EO316 Room UB99(B1) CHALLENGES AND ADVANCES FOR STATISTICAL MODELLING IN DATA SCIENCE Chair: Shu-Kay Ng

E0411: **Comparing classical criteria for selecting intra-class correlated features for mixtures with three-mode three-way data**
Presenter: **Lynette Hunt**, University of Waikato, New Zealand
Co-authors: Kaye Basford, Lynette Hunt
Many unsupervised learning tasks involve data sets with both continuous and categorical attributes. One possible approach to clustering such data is to assume that the data to be clustered come from a finite mixture of populations. There has been extensive use of mixtures where the component distributions are multivariate normal and where the data would be described as two mode two way data. The finite mixture model can also be used to cluster three way data. The mixture model approach requires the specification of the number of components to be fitted to the model and the form of the density functions of the underlying components. The performance of several commonly used model selection criteria is illustrated in selecting both the number of components and the form of the correlation structure amongst the attributes when fitting a finite mixture model to three way data containing mixed categorical and continuous attributes.

E0566: **Predicting pregnancy complications: Models and challenges**
Presenter: **Shalem Leemaqz**, Robinson Research Institute, University of Adelaide, Australia
Co-authors: Gus Dekker, Claire Roberts
An estimated 25% of first pregnancies are affected by Preeclampsia (PE), preterm birth (PTB), intrauterine growth restriction (IUGR) and/or gestational diabetes mellitus (GDM). However, providing interventions to prevent these pregnancy complications is difficult because we are not able to identify which nulliparous women are at risk. Modelling of these complications is challenging due to the complex underlying relationships with a vast number of potential risk factors and co-existing complications. We propose a tiered prediction strategy using penalised regression and integrating multiple systems to estimate the occurrence and co-occurrence probability of each complication for simultaneous prediction of multiple pregnancy complications. The first tier, aimed at a high sensitivity, classifies women who are deemed to be at risk or at low risk. The second tier, with a higher positive predictive value, further stratifies at risk women into high or moderate risk. We utilised data from the Screening for Pregnancy Endpoints (SCOPE) study which consists of detailed clinical and lifestyle variables of pregnant women in Adelaide, Australia. We have also explored visualisation of the concurrent prediction models and development of a learning system through backpropagation feedback. This may assist in providing tailored antenatal care and early interventions that could benefit both the mother and child.

E0653: **A survival ensemble of extreme learning machine**
Presenter: **Hong Wang**, Central South University, China
Due to the fast learning speed, simplicity of implementation and minimal human intervention, extreme learning machine has received considerable attentions recently, mostly from the machine learning community. Generally, extreme learning machine and its various variants focus on classification and regression problems. Its potential application in analyzing censored time-to-event data is yet to be verified. We present an extreme learning machine ensemble to model right-censored survival data by combining the Buckley-James transformation and the random forest framework. According to experimental and statistical analysis results, we show that the proposed model outperforms popular survival models such as random survival forest, Cox proportional hazard models on well-known low-dimensional and high-dimensional benchmark datasets in terms of both prediction accuracy and time efficiency.

E0674: **Mixtures of local logistic regressions for nonlinear classification when data are heterogeneous**
Presenter: **Hien Nguyen**, La Trobe University, Australia
Logistic regression has long been a staple method for the conduct of discrimination. Unfortunately, the standard logistic regression model only permits linear decision boundaries with respect to the input features of the model. We utilise the cluster-weighted modelling approach to construct logistic regression models that permit the expression of local rules and thus allow for nonlinear classification boundaries. These models are particularly suitable for the analysis of heterogeneous data, due the mixture construction.

EO219 Room S101 COMPUTATIONAL CHALLENGES IN STATISTICAL LEARNING Chair: Teng Zhang

E0294: **CESME: Cluster analysis with latent semiparametric mixture models**
Presenter: **Wen Zhou**, Colorado State University, United States
Co-authors: Lyuou Zhang, Hui Zou, Lulu Wang
Model-based clustering is one of the most popular statistical approaches for cluster analysis and has been widely applied in traditional exploratory analyses. However, the Gaussian assumption plays a critical role for model-based clustering, which is not true in general and prevents the model-based clustering to be used for data with complex distributions, such as those from omics study or climate experiments. We propose a semiparametric latent model for clustering multivariate data with complex distributions, particular those are far different from Gaussian. The model assumes that the observed random variables are obtained from unknown monotone transformations of latent variables that satisfy the Gaussian mixture distribution. The identifiability of the proposed model is carefully studied. An alternating maximization procedure is developed to estimate the proposed model, whose convergence property is investigated. Beside the theoretical exploration, the proposed method is also numerically assessed through extensive simulations and has demonstrated superior performance compared to most of the contemporary competitors. Real data analysis has also been studied to demonstrate the usage of the proposed method in practice.

E0357: **Power k-means clustering**
Presenter: **Jason Xu**, Duke University, United States
An alternative to Lloyd's classic algorithm for k-means clustering is proposed that retains its simplicity but mitigates its tendency to get trapped by local minima. Called power k-means, the method embeds the k-means problem in a continuum of similar, better behaved problems with fewer local minima. The previously discovered k-harmonic means algorithm coincides with one point along this continuum. Power k-means anneals its way toward the solution of ordinary k-means by way of majorization-minimization (MM), sharing the appealing descent property and low complexity of Lloyd's algorithm. Further, the method complements widely used seeding strategies, reaping marked improvements when used together. We demonstrate its advantages over simulated and real data examples.

E0565: **On sufficient dimension reduction with mixture normally distributed predictors**
Presenter: **Wei Luo**, Zhejiang University, China
Co-authors: Yan Guo
A family of sufficient dimension reduction methods, called inverse regression, commonly require the linearity condition that $E(X|\beta'X)$ must be a linear function of $\beta'X$ and the constant variance condition that $var(X|\beta'X)$ must be degenerate for certain β. We relax these conditions by allowing more flexibility on the functional forms of $E(X|\beta'X)$ and $var(X|\beta'X)$, under the assumption on the existence of a latent variable that generates X. The generalized conditions are satisfied when X has a mixture elliptical or mixture normal distribution, which is fairly general in practice. Under

the relaxed conditions, we generalize the existing inverse regression methods, with additional adjustments that enhance the exhaustiveness of the methods. Efficient iterative algorithms are proposed for implementation, and simulation models and a real data example are studied to illustrate the effectiveness of the proposed methods.

E0601: **Optimal threshold selection for covariance estimation**
Presenter: **Yumou Qiu**, Iowa State University, United States
Thresholding is a regularization method commonly used for covariance estimation, which provides consistent estimators if the population covariance satisfies certain sparsity condition. However, the performance of the thresholding estimators heavily depends on the threshold level. By minimizing the Frobenius risk of the adaptive thresholding estimator for covariances, we conduct a theoretical study for the optimal threshold level, and obtain its analytical expression. A consistent estimator based on this expression is proposed for the optimal threshold level. Comparing to the state-of-art cross validation method, the proposed method is easy to implement and much more efficient in computation. Numerical simulations and a case study on gene expression data are conducted to illustrate the proposed method.

EO360 Room S102 STATISTICAL METHODS IN BIOINFORMATICS AND BIOSTATISTICS Chair: Shu-Chuan Chen

E0302: **Boolean function networks**
Presenter: **Maria Simak**, Academia Sinica and National Chiao Tung University, Taiwan
Co-authors: Chen-Hsiang Yeang, Jinn-Moon Yang, Henry Horng-Shing Lu
A Boolean Function Network (BFN) model is developed by the integrated approach with the hidden Markov model (HMM), likelihood ratio tests and Boolean logic functions. We evaluate the performance of BFN through the applications to S. cerevisiae and the other biologic time course data. The BFN can produce regulatory relations that are mostly consistency with literature. In addition, it improves sensitivity and specificity with low computational complexity. Moreover, the Boolean functions discovered by BFN can provide useful biological insights for the control mechanisms of regulatory processes.

E0449: **Reliability of meta-analysis studies**
Presenter: **Stan Young**, CGStat, United States
Claims coming from scientific studies, observational and experimental, are reliably estimated to fail to replicate over 50% of the time. All these studies contain statistical analysis in support of claims made. There is a need for an analysis strategy to evaluate claims made in studies to ascertain their reliability. A meta-analysis uses multiple studies to address a common question. The idea is to use these multiple studies to cross-check the claim rather than to attempt to evaluate an individual study. A combination of simple techniques is used: multiple testing and multiple modeling are examined in individual studies; heterogeneity of effects across studies is evaluated using graphical methods; publishing pressures and biases are discussed. Examples from environmental epidemiology are presented. It is found that many claims made in the literature lack statistical support. The benefit is that any claim made in a meta-analysis can be examined for reliability.

E0718: **Developing the association between air pollution and incident of asthma**
Presenter: **Jin-Hua Chen**, Taipei Medical University, Taiwan
Co-authors: Shu-Chuan Chen, Yi-Chun Lin, Tzu-Min Lin
Respiratory allergic patients with poor immunity are susceptible to inhalation allergens to cause disease relapse. Air pollutants are major source of inhalation allergens, such as PM 2.5, O3, NO2, etc. The association between air pollutants and emergency events in asthma patients is investigated. In order to highlight the influence factor of air pollution, local control method and case crossover design were adopted to control the influence from differences of individual characteristics, severity of asthma and weather characteristics between asthma patients with emergency events and without no emergency event. Since too many influence factors we have controlled, we use non-negative matrix factorization (NMF) to cluster these factors, and to reduce number of influence factors we need to control. In addition, NMF will not lose intra information of these factors in the process of reduction of dimensionality. Compared to prior research, we could find the asthma patient group with emergency events which is more susceptible to air pollutants. We use the non-parametric analysis and conditional logistic regression to find the evidence that air pollutants could affect the behavior of emergency visit of asthma patients. Furthermore, we investigate the characteristics of this group and try to find the high risk group of asthma patients with emergency events, and to provide the potential risk factor.

E0530: **Ancestral mixture models for phylogeny reconstruction**
Presenter: **Shu-Chuan Chen**, Idaho State University, United States
Phylogeny, the evolutionary tree of life, can be reconstructed from biological data such as DNA sequences. The phylogenetic tree reconstructed from the biological data gives the conjecture of the relationship between the genetic sequences. Many statistical methods have been proposed to reconstruct the phylogeny from the biological sequences. A new method, the ancestral mixture models, has been proposed by Chen and Lindsay. In the talk, we will extend the ancestral mixture model to the one that can infer phylogeny from the DNA sequences with multiple states on each site. The extended ancestral mixture model is in fact a HKY85 based mixture model assuming transitions and transversions not equally likely, and the frequencies of state A, G, C, T not equal as well. In addition, generalized ancestral mixture models will be presented. Examples and future work will be discussed.

EO089 Room S106 RECENT ADVANCES IN COMPLEX DATA MODELING Chair: Mauricio Castro

E0169: **Bayesian inference on multivariate-t nonlinear mixed-effects models for multiple longitudinal data**
Presenter: **Wan-Lun Wang**, Feng Chia University, Taiwan
Co-authors: Luis Mauricio Castro
The multivariate-t nonlinear mixed-effects model (MtNLMM) has been shown a promising robust tool for analyzing multiple longitudinal trajectories following arbitrary growth patterns in the presence of outliers and possible missing responses. Owing to intractable likelihood function of the model, a fully Bayesian estimating procedure is presented to account for the uncertainties of model parameters, random effects, and missing responses via the Markov chain Monte Carlo method. Posterior predictive inferences for the future values are also investigated. A simulation study is conducted to demonstrate the feasibility of our Bayesian sampling schemes. The proposed techniques are illustrated through applications to case studies.

E0301: **A method for improving the efficiency of phase II trials**
Presenter: **Chien-Ju Lin**, MRC Biostatistics Unit University of Cambridge, United Kingdom
Co-authors: James Wason
In many trials, duration between patient enrollment and an event occurring is used as the efficacy endpoint. Common endpoints of this type include the time until relapse, progression to the next stage of a disease, or time until remission. The criteria of an event may be defined by multiple components, some of which are a continuous measurement being above or below a threshold. Typical analyses consider all components as binary variables and label patients as having events or non-events at multiple time points. This is analysed through constructing and testing survival

functions using Kaplan-Meier, parametric models or Cox models. This approach ignores information on the continuous form of measurements in defining events. We propose a method to make use of this information to improve the precision of analyses using these types of endpoints. We use joint modelling of the continuous and binary components to construct survival curves. We show how to compute confidence intervals for quantities of interest. We assess the properties of the proposed method using simulations and data from a phase II cancer trial and an observational study of in renal disease. Finally, we showcase a R-Shiny app that implements the proposed method.

E0498: **Linear models for multivariate repeated measures data with pattern covariance structures**
Presenter: **Anuradha Roy**, The University of Texas at San Antonio, United States
Co-authors: Timothy Opheim
The popularity of the classical general linear model (CGLM) is mostly due to the ease of modeling and authentication of the appropriateness of the model. However, CGLM is not appropriate and thus not applicable for correlated two-level or three-level multivariate repeated measures data. We propose an extension of linear model with exchangeably distributed errors for multivariate repeated measures data for multiple observations. Maximum likelihood estimates of the matrix parameters of the intercept, slope and the eigenblocks of the exchangeable error matrix are derived for two-level data. The practical implications of the methodological aspects of the proposed extended model for two-level data are demonstrated using two medical datasets.

EO173 Room S1A01 ADVANCEMENTS IN COMPLEX SPATIAL DATA ANALYSIS Chair: Huiyan Sang

E0313: **Multivariate functional data visualization and outlier detection**
Presenter: **Wenlin Dai**, Renmin University of China, China
A new graphical tool, the magnitude-shape (MS) plot, is proposed for visualizing both the magnitude and shape outlyingness of multivariate functional data. The proposed tool builds on the notion of functional directional outlyingness, which measures the centrality of functional data by simultaneously considering the level and the direction of their deviation from the central region. The MS-plot intuitively presents not only levels but also directions of magnitude outlyingness on the horizontal axis or plane, and demonstrates shape outlyingness on the vertical axis. A dividing curve or surface is provided to separate non-outlying data from the outliers. Both the simulated data and the practical examples confirm that the MS-plot is superior to existing tools for visualizing centrality and detecting outliers for functional data.

E0333: **A Bayesian spatially clustered coefficient regression model**
Presenter: **Huiyan Sang**, Texas A&M University, United States
A new Bayesian spatially clustered coefficient (BSCC) regression model is proposed to detect spatial clustering patterns in the associations between response variables and covariates. In BSCC, regression coefficients are assumed to be constants within each spatially contiguous cluster. To model the clustering patterns, we develop a novel and flexible space partitioning prior based on Euclidean spanning trees, which is capable of capturing irregularly shaped clusters. An efficient Reversible Jump Markov chain Monte Carlo (MCMC) algorithm is designed to estimate the clustered coefficient values and their uncertainty measures. Finally, we illustrate the performance of the model with simulation studies and a real data analysis of temperature-salinity relationship in the Atlantic Ocean.

E0372: **Bayesian spatio-temporal modeling of Arctic sea ice extent**
Presenter: **Bohai Zhang**, Nankai University, China, China
Arctic sea ice extent has drawn considerable interest from geoscientists for the last two decades owing to its rapid decline. We propose a Bayesian spatio-temporal hierarchical model for Arctic sea ice extent data, where a latent spatio-temporal Gaussian process is used to model the data dependence and linked to the observations, which here are binary. Through a simulation study, we investigate how parameter uncertainty in a complex hierarchical model can influence spatio-temporal prediction. These results inform how inference will proceed on Arctic sea ice extent over a period of more than twenty years. Covariates that are physically motivated are chosen through autologistic diagnostics. Finally, new summary statistics are proposed to detect the changing patterns of Arctic sea ice between successive time periods.

E0658: **Estimation and inference for generalized geoadditive models**
Presenter: **Lily Wang**, Iowa State University, United States
Co-authors: Shan Yu, Lijian Yang, Chenhui Liu, Guannan Wang
In many application areas, data are collected on a count or binary response with spatial covariate information. We introduce a new class of generalized geoadditive models (GGAMs) for spatial data distributed over complex domains. Through a link function, the proposed GGAM assumes that the mean of the discrete response variable depends on additive univariate functions of explanatory variables and a bivariate function to adjust for the spatial effect. We propose a two-stage approach for estimating and making inferences of the components in the GGAM. In the first stage, the univariate components and the geographical component in the model are approximated via univariate polynomial splines and bivariate penalized splines over triangulation, respectively. In the second stage, local polynomial smoothing is applied to the cleaned univariate data to average out the variation of the first-stage estimators. We investigate the consistency of the proposed estimators and the asymptotic normality of the univariate components. We also establish the simultaneous confidence band for each of the univariate components. The performance of the proposed method is evaluated by two simulation studies. We apply the proposed method to analyze the crash counts data in the Tampa-St. Petersburg urbanized area in Florida.

EO041 Room AT241 DESIGN AND ANALYSIS OF COMPLEX EXPERIMENTS: THEORY AND APPLICATIONS Chair: MingHung Kao

E0392: **Maximin optimal designs for cluster randomized trials with binary outcomes and related applications**
Presenter: **Weng Kee Wong**, UCLA, United States
A nature-inspired metaheuristic algorithm is developed to find extended two-stage adaptive optimal designs for phase II trials with many parameters which is called discrete particle swarm optimization (DPSO). These designs include previous ones as special cases. We show that DPSO not only frequently outperforms greedy algorithms, which are currently used to find such designs when there are only a few parameters; it is also capable of effectively solving adaptive design problems with many parameters that greedy algorithms cannot. In particular, we consider situations where a treatment seems promising in stage 1 but there is great uncertainty in its efficacy rate, and both drug development cost and ethics dictate that there be three pre-determined user-specified efficiency rates for possible testing at stage 2 given testing error rate constraints. We provide a real application and demonstrate benefits of our proposed design strategy for a Phase II trial.

E0391: **Optimal designs for mixed responses with qualitative and quantitative explanatory variables**
Presenter: **MingHung Kao**, Arizona State University, United States
The focus is on optimal designs for experiments where possibly correlated bivariate responses of mixed data types are collected from each experimental subject. Mixed response regression models that involve both qualitative and quantitative explanatory variables are considered for jointly modeling the bivariate responses, and some results on locally optimal designs for such a model are derived. An efficient computer algorithm is also proposed to obtain optimal designs for such experiments.

E0399: **De-aliasing in two-level factorial designs: A Bayesian approach**
Presenter: **Ming-Chung Chang**, Graduate Institute of Statistics, National Central University, Taiwan
Under limited resources on conducting follow-up trials, the inability to disentangle aliased factorial effects hinders the ubiquitous practicality of regular fractional factorial designs in the analysis of experiments. Some frequentist remedies for de-aliasing could misunderstand the underlying system behind the data. Such misinterpretation can be serious if the purpose of experimentation is to find out the mechanism in a process rather than making predictions. A Bayesian approach is proposed for de-aliasing in two-level regular factorial designs. As shown in numerical studies, our method results in a desirable model fitting and a more reliable interpretation of data than the frequentist remedies.

E0572: **Cost considerations for group testing studies with an imperfect assay and a gold standard**
Presenter: **Shih-Hao Huang**, National Central University, Taiwan
Co-authors: Mong-Na Lo Huang, Kerby Shedden
The focus is on optimal group testing design problems when a cheap imperfect assay and an expensive perfect assay (gold standard) for a target trait are both available. The primary goal is to accurately estimate the prevalence of the trait in a given population, where the testing error rates of the imperfect assay are treated as nuisance parameters. Budget constraints are used to reflect the relative costs of performing the two assays and of collecting the individual samples. A mixed design strategy can be adopted, where individual samples are tested as a group sample by one of the three procedures: only the imperfect assay, only the perfect assay, and both assays. We characterize the optimal designs within the class with a tight upper bound on the number of distinct group sizes for each testing procedure. Based on this information, we provide an efficient algorithm to obtain an optimal budgeted design.

EO087 Room AT242 RECENT APPLICATIONS OF LATENT VARIABLE MODELS Chair: Gongjun Xu

E0234: **Modeling non-ignorable missing for not-reached and omitted items using item response times**
Presenter: **Chun Wang**, University of Washington, United States
Item nonresponses are prevalent in standardized testing. They happen either when students fail to reach the end of a test due to a time limit, or when the students choose to omit some items strategically. Oftentimes item nonresponses are non-random and hence the missing data mechanism needs to be properly modeled. We propose to use innovative item response time model as a cohesive missing data model to account for two most common item nonresponses: not-reached items and omitted items. Simulation studies show that the proposed approaches improve estimation precision of item parameters compared with the method based solely on observed responses (i.e., ignoring missing data). Moreover, for persons with missing data, their latent trait estimates are also less biased and more precise (i.e., lower standard error). The 2015 PISA computer-based mathematics data is analyzed to illustrate the application of the proposed method.

E0555: **A modified continuous a-stratification index for item exposure control in computerized adaptive testing**
Presenter: **Ya-Hui Su**, National Chung Cheng University, Taiwan
Computerized adaptive testing (CAT) have been increasingly applied to many fields, including educational assessment, psychological testing, personnel recruitment, and clinical diagnosis. To ensure the validity of test scores, preventing items from being overexposed is very important to practical consideration because these items might be shared with current and future examinees. One popular method for item exposure control is a-stratification. When old or overexposed items would need to be removed from the item bank or new items would need be added into the item bank, the simulation studies would need to be conducted to determinate the optimal strata for a-stratification. It is undesirable for practitioners whenever the item bank is replenished frequently, such as high-stakes testing. Recently, a continuous a-stratification index (CAI) was proposed to incorporate item exposure control into the item selection index itself, and thus no need to partition the item bank into the fixed and discrete strata. However, the CAI still produced a high percentage of overexposed items that many practitioners couldn't view as negligible. Therefore, the study was to modify the CAI for limiting the maximum items exposure rate, and to investigate its efficiency with some current item exposure control methods for item selection in CAT.

E0643: **Sparse latent class regression for multivariate binary data: A Bayesian approach**
Presenter: **Zhenke Wu**, University of Michigan, United States
In scientific problems where a vector of discrete responses is observed along with covariates, classical latent class regression models (LCRM) play a key role in studying their associations. We propose a flexible Bayesian LCRM to estimate for each individual the posterior probabilities of assignment to a small number of classes. Given covariates, our approach first decomposes a probability contingency table into classes, each characterized by a vector of response probabilities. We extend the decomposition to infinitely many classes and introduce a novel stick-breaking type prior on the class weights. We then mix these classes with weights that can depend on the covariates. We specify priors to encourage class-weight regression functions to be constant over covariate values. The posterior distribution for the number of classes in the population is estimated from data and by design tends to concentrate on the smaller values. We demonstrate the method on both simulated data and real data from a multi-country childhood pneumonia etiology study.

E0782: **Marginalized maximum a posteriori estimation for the 4-parameter logistic model under a mixture modeling framework**
Presenter: **Xiangbin Meng**, Northeast Normal University, China
Co-authors: Gongjun Xu
The 4-parameter logistic model (4PLM) has recently gained great interests in various applications. The 4PLM is reexpresed to be a mixture model with two levels of latent variables and further develop a marginalized maximum a posteriori (MMAP) estimation with an Expectation-Maximization (EM) algorithm. The mixture modeling framework of the 4PLM not only makes the proposed EM algorithm more easily to be implemented in practice, but also provides a natural connection with the popular cognitive diagnosis models. Simulation studies were constructed to show the good performance of the proposed estimation method and to investigate the impact of the additional upper asymptote parameter on the estimation of other parameters. Moreover, a real data set was analyzed by the 4PLM to show its outperformance over the 3-parameter logistic model (3PLM).

EO326 Room AT335 STATISTICS FOR ENVIRONMENTAL RISK ASSESSMENT AND MANAGEMENT Chair: Kunihiko Takahashi

E0464: **Development of chemical risk assessment and management system in environmental accidents**
Presenter: **Takahiro Otani**, Nagoya University Graduate School of Medicine, Japan
Co-authors: Kunihiko Takahashi
Disasters and accidents involving the proliferation of chemical pollutants cause serious problems for the environment and human health. For a disaster and emergency response, Ministry of the Environment of Japan has launched a study project on chemical risk assessment and management system. The aim of this project is to systematically investigate comprehensive monitoring methods of chemicals, dispersion prediction models, and risk assessment and management system under nonstationary circumstances and to develop information infrastructures involving these achievements. Several statistical issues on chemical analysis, anomaly detection, spatiotemporal modeling, exposure assessments, and health risk evaluation are arising for the realization of this system. We briefly introduce the overview of the study project, and as an example of achievements,

demonstrate a software tool for exposure assessments of air pollutants via spatial interpolation methods using limited available data of pollutant concentrations immediately after accidents.

E0548: **Preliminary application of multidimensional non-target analysis data by GCxGC/HRToFMS for environmental monitoring**
Presenter: **Shunji Hashimoto**, National Institute for Environmental Studies, Japan

A non-target analytical method to measure as many chemical pollutants as possible to avoid missing of risk is being developed. One of the most important issues of non-target analysis is to detect anomaly condition of the environment. Stability and anomaly detection power of our method were statistically evaluated. Six samples collected from small river on six different days were used for examination. Each sample was divided into 5 sub-samples and those sub-samples were separately analyzed by the method. The data matrix size of one sub-sample measured by comprehensive two-dimensional gas-chromatography/ high-resolution time-of -flight mass spectrometry (GCxGC/HRToFMS) was 857(1st GC time) * 231(2nd GC time) * 16,800(mass) containing double-precision floating-point values (signal intensities). Although around several thousand components were detected from one sample, only 80 components in each sub-sample were tested. According to the results of Kruskal-Wallis test, it was possible to detect some difference inter samples even though within-group variation was relatively large. However, to enhance the sensitivity and reliability of difference detection inter samples, reproducibility and stability of non-target analytical method has to be improved.

E0594: **Anomaly detection by analysis of water quality data of Osaka City waterworks bureau in Japan**
Presenter: **Manabu Ichikawa**, Shibaura Institute of Technology, Japan
Co-authors: Mari Asami, Takehiro Arai, Yutaka Nakai, Koya Tanaka

Water quality data which was recorded every minute for five years was analyzed for detecting anomaly. In Waterworks Bureau, to find anomaly is very important for controlling the system of purifying water. The data analysis result revealed the relationship between weather and water quality abnormality, and it also clarified how the abnormality detection affects the control of the purification systems. Through this research we will try to find useful information for optimum water purification control.

E0563: **Cohort study and environmental exposure, recent progress in Japan**
Presenter: **Ayano Takeuchi**, Keio University, Japan

Two topics will be considered. One is to assess the effect of environmental chemical exposure on fetuses or development of children using birth cohort data. Another is to assess the health effect of air pollution on cardiovascular and respiratory diseases. Ministry of the Environment, Japan have launched birth cohort study targeting 100,000 pregnant from 2011 to evaluate health effect of environmental chemical on fetuses or children. In exposure assessment, there is a problem of detection limits (DLs) and quantitation limits (QLs) of exposure substances. We can convert these problems to incomplete data problem or left-censored data problem. DLs and QLs may differ depending not only on substances, but also measurement batch, time and equipment within same substance. The other problem is that exposure and outcomes are affecting each other over time. Evaluating the relationship between baseline risks and outcomes used in conventional epidemiological studies is not enough. Statistical models dealing with birth cohort data have not yet been established. We may use causal inference approach such as drawing DAGs or separating direct and indirect effects using conditional models etc. Health effects of air pollution are often evaluated by merging existing cohort data or existing official statistics with air pollution data. Time series data analysis, case cross over analysis or generalized additive models are used when evaluating short term health effects.

EO318 Room AT337 RECENT ADVANCES IN HIGH DIMENSIONAL GENOMIC DATA ANALYSIS Chair: Yuehua Cui

E0433: **Data-adaptive omnibus tests by combining high-dimensional statistical inference for testing SNP-set effects**
Presenter: **Haitao Yang**, Hebei Medical University, China

Genome-wide association studies have identified numerous genetic variants associated with complex disease. However, these variants can only explain a small portion of heritability in many diseases. SNP-set based association methods have been proved to be alternative strategies to capture some missing heritability, but the majority of these methods cannot be generalized to high-dimensional data. Recent advances in high-dimensional model development have been shifted from high-dimensional variable selection to high-dimensional statistical inference. While these models have limited power to detect weak genetic signals. Theoretically, the maximum statistic distribution method and the p-value combination method can respectively improve the power of two genetic effect hypotheses which are main loci determinism and loci micro effect accumulation theory. Motivated by this, we proposed to construct a SNP-set based high-dimensional statistical inference procedure that can be adaptive to the two genetic effect hypotheses by combing the Omnibus Test, to improve the power of detecting genetic variants, and to provide a novel statistical framework for the study of genetic mechanism underlying complex human diseases.

E0533: **Identification of trans-eQTLs using mediation analysis with multiple mediators**
Presenter: **Zuoheng Wang**, Yale University, United States
Co-authors: Nayang Shan, Lin Hou

Identification of expression quantitative trait loci (eQTLs) advances understanding of genetics and regulatory mechanisms of gene expression in various organisms. Previous studies suggest that trans-eQTLs may regulate expression of remote genes by altering the expression of nearby genes. Trans-association has been studied in the mediation analysis with a single mediator. However, prior applications with one mediator are prone to model misspecification due to correlations between genes. Motivated from the observation that trans-eQTLs are more likely to associate with more than one cis-gene than randomly selected SNPs in the GTEx dataset, we developed a computational method to identify trans-eQTLs that are mediated by multiple mediators. In simulation studies, multiple mediator analysis had increased power to detect mediated trans-eQTLs, especially in large samples. In the HapMap3 data, we identified 11 mediated trans-eQTLs that were not detected by the single mediator analysis in the combined samples of African populations. Moreover, the mediated trans-eQTLs in the HapMap3 samples are more likely to be trait-associated SNPs. Our approach has improved the power of detecting mediated trans-eQTLs and advanced knowledge of gene regulation.

E0554: **Large-scale mediation effect signal detection in genome-wide epigenetic studies**
Presenter: **Zhonghua Liu**, The University of Hong Kong, Hong Kong

In genome-wide epigenetic studies, it is often of scientific interest assessing the mediator role of DNA methylation in the causal pathway from an exposure to a clinical outcome. Mediation analysis is commonly used to answer this question. This is often done via fitting two regression models: the mediator model and the outcome model, and then the product of coefficient method to integrate information from these two models and performing hypothesis testing using Sobel's test. We propose a novel divide-aggregate test (DAT) for the composite null hypothesis for the detection of mediation effects in genome-wide epigenetic studies. We further show that the DAT can outperform the Sobel's test and the joint significance test for the detection of mediation effects in genome-wide epigenetic studies. A fast Monte Carlo correction method is also proposed for computing the p-value of the DAT method. We show via simulation studies that the DAT method controls type I error rates and outperforms the Sobel's and the joint significance test. We applied the DAT method to the Normative Aging Study to identify putative DNA methylation sites that mediate the effect of smoking on lung function.

E0588: **Improved classification accuracy through inclusion of latent variables**
Presenter: **Johann Gagnon-Bartsch**, University of Michigan, United States
Co-authors: Yujia Pan
Classification of high-throughput genomic data is challenging because the signal is often weak and sparse. Incorporating side information or additional covariates (e.g., gender, age) can lead to better predictive accuracy, but it is often the case that such information is unknown. To this end, we introduce a classifier which adaptively leverages both observed variables as well as inferred latent ones. Including these latent variables tends to improve accuracy, sometimes substantially, as illustrated on several simulated and genomic datasets. A diverse collection of genomic datasets are considered (gene expression, methylation, and SNP data), as well as a wide range of disease phenotypes (asthma, Alzheimer's disease, tuberculosis, and schizophrenia), illustrating broad applicability.

EO243 Room U301 RECENT DEVELOPMENTS IN QUANTILE ESTIMATION AND INFERENCE Chair: Liang Chen

E0247: **Globally adaptive longitudinal quantile regression with high-dimensional compositional covariates**
Presenter: **Qi Zheng**, University of Louisville, United States
Co-authors: Limin Peng
The human microbiome is associated with many diseases and is often characterized by a high dimensional compositional structure. In many microbiome studies, measurements are taken longitudinally and the outcome of interest is subject to left censoring due to the detection limit or other reasons. We propose a cross-sectional longitudinal quantile regression framework that investigates the association between the continuous outcome of interest and human microbiome composition along with other usual covariates. Log-contrast of compositions is used and then be reformulated as a symmetric form with zero-sum coefficients. To minimize the objective function with non-differentiable terms and linear constraint, we smooth quantile loss function to use the maximization-minimization trick that yield closed form updates in each step of coordinate descent algorithm. The oracle properties of the regularized globally concerned quantile estimator are obtained. No matter which active variable is chosen as a reference, the asymptotic behaviors are equivalent. We conduct several setups of simulation studies to assess the finite sample performance of the proposed estimator. The human microbiome data is used to illustrate the practicality of the proposed method.

E0492: **Estimating quantile panel data models with interactive fixed effects**
Presenter: **Liang Chen**, Shanghai University of Finance and Economics, China
The estimation of panel data models with interactive fixed effects is considered, where the idiosyncratic errors are subject to conditional quantile restrictions. We propose a two-step estimator that is easy to implement in practice. In the first step, a principal component analysis is applied to the dependent variables to estimate the time effects, and in the second step, a smoothed quantile regression is used to estimate the slope parameter and the individual effects jointly. The consistency and asymptotic normality of the slope estimator are established under large N and large T asymptotics. It is found that the asymptotic distribution of the slope estimator suffers from asymptotic biases, and we show how to correct the biases using split-panel jackknife. Simulation studies confirm that the bias-corrected estimator performs well with moderate sample sizes.

E0293: **Blessing and curse of cross-sectional length on robust estimation for panel data**
Presenter: **Lingyu He**, The Australian National University, Australia
Co-authors: Yanrong Yang
Cross-sectional dimension (CSD) increases the rate of convergence for common information estimation in panel data models. However, Cross-sectional dependence automatically appears as more cross-sections are involved, which decreases the efficiency of estimation for homogeneous characteristics to some extent. This is the trade-off between the blessing and curse brought by cross-sectional dimensionality. The focus is on this point for robust M-estimation on panel data models. On the one hand, the rate of convergence of a high dimensional coefficient vector common for all cross-sections is provided, which shows that, using cross-sectional data as much as possible is a blessing for extracting homogeneous information. On the other hand, the asymptotic distribution of M-estimator for this parameter vector is established. It can show that the asymptotic variance heavily depends on cross-sectional dependence incurred by cross-sectional length. This interrupts the estimation efficiency. Under different settings for cross-sectional dimensions and dependence, simulations illustrates some common used M-estimation and compare them with the least-squares method. Empirical application on stock returns data from CRSP is provided, which show robust M-estimation is necessary and different cross-sections bring different results.

EO197 Room U302 ADVANCES IN FINANCIAL ECONOMETRICS Chair: Andreas Heinen

E0587: **On investing in hedge funds: Optimal portfolios with regime-switching**
Presenter: **Alfonso Valdesogo**, THEMA, Universite de Cergy-Pontoise, France
Co-authors: Andreas Heinen
The joint effect of non-linearity and time variation in hedge fund returns on the benefit for an investor to include hedge funds into his optimal portfolio is analyzed. We model time variation with regime-switching and allow for an additional source of non-linearity with the use of copulas. We estimate a multivariate regime-switching copula model with one symmetric Gaussian dependence regime and with a possibly asymmetric canonical vine regime that allows for tail dependence. We compute the gains for an active investor with CRRA utility, who considers different asset classes, from investing in any one of a number of hedge fund strategies. The asset classes we consider are U.S. stocks, a global bond index, and commodities. We observe that the gains are not homogeneous amongst the difference hedge funds strategies.

E0597: **Indirect inference estimation of nonlinear dynamic panel data models**
Presenter: **Antonio Cosma**, University of Luxembourg, Luxembourg
Co-authors: Fausto Galli
Nonlinear dynamic panel models are estimated by indirect inference, taking into account unobserved heterogeneity. We focus in particular on short panels. Monte Carlo simulations show that the method performs well, especially in short panels, both in terms of point estimates of the parameters and coverage of confidence intervals.

E0623: **Geographic dependence and diversification in house price returns: The role of leverage**
Presenter: **Andreas Heinen**, Universite de Cergy Pontoise, France
Co-authors: Mi Lim Kim
The aim is to analyze the time variation in the average dependence within a set of regional monthly house price index returns in a regime switching multivariate copula model with a high and a low dependence regime. Using equidependent Gaussian copulas, we show that the dependence of house price returns varies across time with changes in credit market conditions, which reduces the gains from the geographic diversification of real estate and mortgage portfolios. More specifically, we show that a decrease in leverage, and to a lesser extent an increase in mortgage rates, are associated with a higher probability of moving to and staying in the high dependence regime.

E0615: **Competition, fast growth and commercialization: Systemic credit risk in microcredit markets**
Presenter: **Malika Hamadi**, Birmingham Business School, United Kingdom
Co-authors: Andreas Heinen
A new copula-based measure of systemic credit risk in microcredit markets is introduced and its determinants for 37 countries from 2000 to 2014 are explored. We model the joint distribution of portfolio quality of all microfinance institutions (MFIs) in a given country using an equidependent copula and an empirical density for the marginal. The methodology is based on the idea that a higher level of dependence among MFIs in a given country is an indicator of potential fragility of the sector. We show that, after controlling for country effects, measures of competition, market penetration, commercialization and excessive growth of the sector, as well as the level of interest rates charged by MFIs in a country all increase the fragility of the microfinance sector, while lending to more women tends to reduce systemic risk. With this measure of systemic risk, we further compute the probability that a proportion of at least 20% of MFIs in a country are in serious financial distress and our simulations captures a progressive increase in risk in many countries that experienced a repayment crisis.

EO263 Room U414 ADVANCES IN STATE SPACE MODELS AND BAYESIAN COMPUTATION Chair: Minh-Ngoc Tran

E0380: **Flexible density tempering approaches for state space models with an application to factor stochastic volatility models**
Presenter: **David Gunawan**, University of New South Wales, Australia
Co-authors: Robert Kohn, Christopher K Carter, Minh-Ngoc Tran
A tempering or annealing approach to Bayesian inference for time series state space models is proposed. In such models the likelihood is often analytically and computationally intractable. Their approach generalizes the annealed importance sampling (AIS) approac when the likelihood can be computed analytically. A critical component of the annealing or density tempering method is the Markov move component that is implemented at every stage of the annealing process. The Markov move component effectively runs a small number of Markov chain Monte Carlo iterations for each combination of parameters and latent variables so that they are better approximations to that level of the tempered target density. Previously a pseudo marginal Metropolis-Hastings (PMMH) approach with the likelihood estimated unbiasedly in the Markov move component has been used. One of the drawbacks of this approach, however, is that it is difficult to obtain good proposals when the parameter space is high dimensional, such as for a high dimensional factor stochastic volatility models. We propose using instead more flexible Markov move steps that are based on particle Gibbs and Hamiltonian Monte Carlo and demonstrate the proposed methods using a high dimensional stochastic volatility factor model. An estimate of the marginal likelihood is obtained as a byproduct of the estimation procedure.

E0508: **Subsampling sequential Monte Carlo for static Bayesian models**
Presenter: **Doan Khue Dung Dang**, University of New South Wales, Australia
Co-authors: David Gunawan, Robert Kohn, Matias Quiroz, Minh-Ngoc Tran
The aim is to show how to carry out Bayesian inference by combining data subsampling with Sequential Monte Carlo (SMC). This takes advantage of properties of SMC for Bayesian computations with the ability of subsampling to tackle big data problems. SMC sequentially updates a cloud of particles through a sequence of densities, beginning with a density that is easy to sample from such as the prior and ending with the posterior density. Each update of the particle cloud consists of three steps: reweighting, resampling, and moving. In the move step, each particle is moved using a Markov kernel and this is typically the most computationally expensive part, especially when the dataset is large. It is crucial to have an efficient move step to ensure particle diversity. Our article makes two important contributions. First, in order to speed up the computation, we use an approximately unbiased and efficient annealed likelihood estimator based on data subsampling. The subsampling approach is more memory efficient than the corresponding full data SMC, which is a great advantage for parallel computation. Second, we use a Metropolis within Gibbs kernel with two conditional updates. First, a Hamiltonian Monte Carlo update makes distant moves for the model parameters. Second, a block pseudo-marginal proposal is used for the particles corresponding to the auxiliary variables for the data subsampling. We demonstrate the usefulness of the methodology through a series of examples.

E0513: **A long short-term memory stochastic volatility model**
Presenter: **Nghia Nguyen Trong**, University of Sydney, Australia
Co-authors: Minh-Ngoc Tran, Robert Kohn, David Gunawan
Stochastic Volatility (SV) models are widely used in the financial sector and LongShort-Term Memory (LSTM) models have been successfully used in many large-scale industrial applications of Deep Learning. This present work combines these two techniques in a non-trivial way and proposes a model, called LSTM-SV, for capturing efficiently the dynamics in financial volatility processes. The proposed model overcomes the short-term memory problem in conventional SV models, is able to capture non-linear dependences in the latent volatility process, and often has a better out-of-sample forecast ability. These are illustrated through three financial time series datasets: US stock market index S&P500, Australian stock index ASX200 and Australian-US dollar exchange rates. We argue that there are significant differences in the underlying dynamics between the volatility process of S&P500 and ASX200 datasets and that of the exchange rate dataset. For the stock index data, there is strong evidence of long-term memory and non-linear dependences in the volatility process, while this is not the case for the exchange rates. An user-friendly software is publicly available.

E0677: **Efficiently combining pseudo marginal and particle Gibbs sampling**
Presenter: **Robert Kohn**, University of New South Wales, Australia
Particle Markov Chain Monte Carlo (PMCMC) is a general approach to carry out Bayesian inference in non-linear and non-Gaussian state space models. We show how to scale up PMCMC in terms of the number of parameters and number of time points by generating parameters that are highly correlated with the states with the states integrated out using a pseudo marginal step while the rest of the parameters are generated conditional on the states using particle Gibbs. We make the PMCMC scalable in the number of observations by using the same random numbers in the Metropolis-Hastings ratio of the pseudo marginal step. We do so by expressing the target density of the PMCMC in terms of the basic uniform or standard normal random numbers rather than in terms of the particles, as has been done till now, and develop a constrained version of conditional sequential Monte Carlo algorithm. We illustrate the methods using a high dimensional factor stochastic volatility having both a large number of parameters and a large number of latent states and show that our proposed method makes the computation much more efficient.

EO085 Room U501 DEPENDENT DATA ANALYSIS Chair: Tao Zou

E0712: **Time-varying coefficient spatial autoregressive panel data model with fixed effects**
Presenter: **Xuan Liang**, The Australian National University, Australia
A time-varying coefficient spatial autoregressive panel data model with the individual fixed effects is developed in order to capture the nonlinear effects of the regressors, which varies over time. To effectively estimate the model, we propose a method cooperating the nonparametric local linear method and the concentrated quasi-maximum likelihood estimation method to obtain the consistent estimators for the spatial coefficient and the time-varying coefficient functions. The asymptotic properties of these estimators are derived as well, showing the regular $\sqrt{NT}$-rate of convergence for the parametric parameters and the common $\sqrt{NTh}$-rate of convergence for the nonparametric component. Monte Carlo simulations are conducted to illustrate the finite sample performances of our proposed method. Meanwhile, we apply our method to study Chinese labor productivity to identify the spatial influences and the time-varying spillover effects among 185 Chinese cities.

E0713: **Some extensions to covariance regression**
Presenter: **Tao Zou**, The Australian National University, Australia
Recently, a novel covariance regression model is proposed to study the relationship between the covariance matrix of responses and their associated similarity matrices induced by auxiliary information. To broaden the usefulness of the covariance regression model, the normality assumption is relaxed and effective and efficient methodology is developed in order to obtain inferences for covariance regression models. Moreover, sophisticated statistics are developed for some extensions of covariance regression. Both the theory and practice of statistics, and substantive fields of applications with covariance matrices of data involved, such as geostatistics, sociology and economics will be influenced.

E0721: **Asymptotic properties of ML and REML for nested error regression models**
Presenter: **Ziyang Lyu**, Australian National University, Australia
Asymptotic results are considered for the maximum likelihood and restricted maximum likelihood (REML) estimators of the parameters in the nested error regression model when both of the number of independent clusters and the cluster sizes (the number of observations in each cluster) go to infinity. A set of conditions is given under which the estimators are shown to be asymptotically normal. There are no restrictions on the rate at which the cluster size tends to infinity.

E0841: **Sieve bootstrap for high-dimensional time series**
Presenter: **Daning Bi**, The Australian National University, Australia
Co-authors: Yanrong Yang
A bootstrap procedure for high dimensional time series based on the method of sieves is proposed which applies a low dimensional factor process to the high dimensional time series and then generates a residual-based bootstrap replicates of the low dimensional factor process. In particular, we develop a bootstrap approach which works properly when the dimension of time series N is as large as, or even larger than, the length of observed time series T. The first step of our method is applying a dimension reduction method, where we used a lower-dimensional factor process, to find a low dimension representation of the original time series. Then we can find a bootstrap replicates of the factor process based on a sieve method. And finally we can use the bootstrapped samples to create the prediction intervals of the process and study the statistical inference of a statistics of the original high dimensional time series. We show its consistency for bootstrapped mean vectors. A simulation study is performed to illustrate the coverage probabilities of the confidence band for mean vectors and the prediction intervals for the forecasts.

EO129 Room U502 COMPUTATIONAL STATISTICAL INFERENCE FOR STOCHASTIC PROCESSES Chair: Kengo Kamatani

E0501: **A bootstrap test for sphericity of time series**
Presenter: **Yan Liu**, Kyoto University; RIKEN AIP, Japan
The problem of testing the sphericity hypothesis for the covariance matrix of high-dimensional time series is considered. It has been shown that the test statistic for sphericity under the null hypothesis is asymptotically normal for high-dimensional time series models. However, the asymptotic variances of the test statistic are different from each other if the time series models are different. To alleviate the computational complexity and model dependence, we propose a bootstrap procedure for the test statistic. The validity and the consistency of the proposed method are justified from the asymptotic theory. The performance is illustrated by simulation studies.

E0458: **Maximum composite likelihood estimation for determinantal point processes**
Presenter: **Yasutaka Shimizu**, Waseda University, Japan
The maximum composite likelihood estimators for stationary and isotropic parametric models of determinantal point processes (DPP) are discussed. Since the conditional marginal distribution of the point processes are given by determinant of positive definite kernels, we have the explicit form of the composite likelihoods for every order. This fact enables us to consider the generalized maximum composite likelihood estimator for every order. We will discuss the asymptotic properties using the limit theorems of DPPs.

E0460: **The finite sample properties of sparse M-estimators with pseudo-observations**
Presenter: **Benjamin Poignard**, Osaka University, Japan
Co-authors: Jean-David Fermanian
Finite sample properties of general regularized statistical criteria in the presence of pseudo-observations are provided. Under the restricted strong convexity assumption of the unpenalized loss function and regularity conditions on the penalty, we derive non-asymptotic error bounds on the regularized M-estimator that hold with high probability. This penalized framework with pseudo-observations is then applied to the M-estimation of some usual copula-based models. These theoretical results are supported by an empirical study.

E0467: **Consistent model selection for ergodic diffusions in data driven time scale**
Presenter: **Shoichi Eguchi**, Osaka University, Japan
Co-authors: Hiroki Masuda
There are several studies of model selection for stochastic differential equations(SDEs), which includes the contrast-based information criterion for ergodic diffusion processes and the Schwarz type information criterion for locally asymptotically quadratic models. However, most of the existing theoretical literature has been developed in known model-time scale. We will first overview parameter estimation method for ergodic diffusion processes with unknown model-time scale and then propose the Schwarz type statistics for model selection.

EO075 Room U517 ADVANCES IN HAWKES PROCESSES AND THEIR APPLICATIONS Chair: Feng Chen

E0166: **Statistical inference for the doubly stochastic self-exciting process**
Presenter: **Yoann Potiron**, Keio University, Japan
Co-authors: Simon Clinet
The aim is to introduce and show the existence of a Hawkes self-exciting point process with exponentially-decreasing kernel and where parameters are time-varying. The quantity of interest is defined as the integrated parameter, and we consider the high-frequency asymptotics. To estimate it naively, we chop the data into several blocks, compute the maximum likelihood estimator (MLE) on each block, and take the average of the local estimates. The asymptotic bias explodes asymptotically, thus we provide a non-naive estimator which is constructed as the naive one when applying a first-order bias reduction to the local MLE. We show the associated central limit theorem. Monte Carlo simulations show the importance of the bias correction and that the method performs well in finite sample, whereas the empirical study discusses the implementation in practice and documents the stochastic behavior of the parameters.

E0185: **Asymptotic distribution of the score test for detecting marks in Hawkes processes**
Presenter: **Simon Clinet**, Keio University, Japan
Co-authors: William Dunsmuir, Gareth Peters, Kylie-Anne Richards
A general class of marked Hawkes processes is considered. We derive the asymptotic distribution of a previously proposed score statistic. The null hypothesis corresponds to the case where marks do not impact the stochastic intensity function. In that case, we prove that the score statistic converges to a chi-squared distribution with degrees of freedom equal to the number of parameters required to specify the boost function. Under a well-chosen sequence of local alternatives, we also prove that the statistic converges to a noncentral chi-squared distribution, allowing us to derive the local power of our test.

E0483: **Modeling extreme negative returns using marked renewal Hawkes processes**
Presenter: **Tom Stindl**, UNSW, Australia
Extreme return financial time series are often challenging to model due to the presence of heavy temporal clustering of extremes and strong bursts of return volatility. The marked self-exciting process has been used extensively to model these phenomena. However, it restricts the arrival times of exogenously driven returns to follow a Poisson process and may fail to provide an adequate fit. We introduce a modification to the marked Hawkes process by defining the arrival of exogenously driven extreme returns in terms of a renewal process. We discuss a direct likelihood evaluation approach to parameter estimation which poses some additional computational challenges but only requires quadratic computational time. As a by-product of the likelihood evaluation algorithm, we have a computationally efficient method for goodness-of-fit assessment and a simple, yet efficient procedure for future event prediction. The proposed model is applied to extreme negative returns for stocks traded on the ASX. The models identified for the stocks using in-sample data were found to be able to successfully forecast the out-of-sample risk measures such as the value at risk and expected shortfall and provides a better quality of fit than the competing Hawkes model.

E0516: **Predicting the popularity of tweets using internal and external knowledge: An empirical Bayes approach**
Presenter: **Feng Chen**, UNSW Syd, Australia
Co-authors: Wai Hong Tan
The problem of tweet popularity prediction is considered. We model the retweet time sequence using an inhomogeneous Poisson process with the intensity function depending on the age of the original tweet and the calendar time, through a parametric and a nonparametric function respectively. The functional parameter is estimated nonparametrically using the training data, and the finite-dimensional parameter using both internal knowledge on the times of historical retweets up to the censoring time, and external knowledge on complete retweet sequences in the training data set. The internal and external knowledge are then combined using a novel empirical Bayes type approach, where the prior distribution for the model parameter is constructed based on the external knowledge, and the likelihood is calculated based on the internal knowledge. The mode of the posterior distribution is then used as the estimator of the finite-dimensional parameter. Suitable functionals of the predictive distribution for the number of retweets implied by the estimated model are then used to predict the tweet popularity. The proposed methodology was applied to a large Twitter data set, and its performance is found to be superior to those of the competing approaches in the literature.

EO217 Room U602 INFERENCE ON NETWORKS Chair: Bhaswar Bhattacharya

E0626: **Adaptive estimation of multivariate piecewise constant functions**
Presenter: **Sabyasachi Chatterjee**, University of Illinois at Urbana Champaign, India
The estimation of multivariate piecewise constant functions is considered. A natural estimator is the Dyadic Cart estimator. We show that an extension of Dyadic Cart attains the best possible risk (up to log factors) adaptively for all piecewise constant functions in dimension 2. In higher dimensions, such a property continues to hold for a special subclass of all piecewise constant functions, but not all. Along the way, we show some new adaptive results for estimation of functions of bounded variation using dyadic cart and its extensions.

E0636: **Global testing against sparse alternatives under Ising models**
Presenter: **Rajarshi Mukherjee**, Harvard T.H. Chan School of Public Health, United States
Co-authors: Sumit Mukherjee, Ming Yuan, Gourab Ray
The effect of dependence on detecting sparse signals is studied. In particular, we focus on global testing against sparse alternatives for the magnetizations of an Ising models and establish how the interplay between the strength and sparsity of a signal determines its detectability under various notions of dependence (i.e. the coupling constant and underlying network of the Ising model). Moreover, we provide evidence that certain critical states of the model exhibit a subtle "blessing of dependence" phenomenon in that one can detect much weaker signals at criticality than otherwise. Furthermore, we develop testing procedures that are broadly applicable to account for general network dependence and show that they are asymptotically minimax-separation optimal in many examples.

E0639: **Inference in graphical models via semidefinite programming hierarchies**
Presenter: **Murat A Erdogdu**, University of Toronto, Canada
Maximum A posteriori Probability (MAP) inference in graphical models amounts to solving a graph-structured combinatorial optimization problem. Popular inference algorithms such as belief propagation (BP) and generalized belief propagation (GBP) are intimately related to linear programming (LP) relaxation within the Sherali-Adams hierarchy. Despite the popularity of these algorithms, it is well understood that the Sum-of-Squares (SOS) hierarchy based on semidefinite programming (SDP) can provide superior guarantees. Unfortunately, SOS relaxations for a graph with n vertices require solving an SDP with $n^{\Theta(d)}$ variables where d is the degree in the hierarchy. In practice, for $d \geq 4$, this approach does not scale beyond a few tens of variables. We propose binary SDP relaxations for MAP inference using the SOS hierarchy with two innovations focused on computational efficiency. Firstly, in analogy to BP and its variants, we only introduce decision variables corresponding to contiguous regions in the graphical model. Secondly, we solve the resulting SDP using a non-convex Burer-Monteiro method, and develop a sequential rounding

procedure. We demonstrate that the resulting algorithm can solve problems with tens of thousands of variables within minutes, and outperforms BP and GBP on practical problems such as image denoising and Ising spin glasses.

E0792: **Community estimation on non-binary networks**

Presenter: **Min Xu**, Rutgers University, United States

Community identification in a network is an important problem in fields such as social science, neuroscience, and genetics. Over the past decade, stochastic block models (SBMs) have emerged as a popular statistical framework for this problem. However, SBMs have an important limitation in that they are suited only for networks with binary edges. Network edges often carry additional information such as weights; disregarding such edge information may result in deteriorated performance in various scientific applications. We study a generalization of the SBM, in which a network is represented in the form of a non-binary adjacency array and the observation that correspond to each edge is generated independently from an unknown probability distribution determined by the community membership of its endpoints. In the case where the observations are real-valued, we characterize the optimal rate of misclustering error of the weighted SBM in terms of the Renyi divergence of order 1/2 between the distributions of within-community and between-community edges, substantially generalizing existing results for SBMs. Furthermore, we present a computationally tractable algorithm based on discretization that achieves the optimal error rate. Our method is adaptive in the sense that the algorithm, without assuming knowledge of the edge distributions, performs as well as the best algorithm that knows the edge distribution.

Thursday 27.06.2019 16:10 - 17:25 Parallel Session M – EcoSta2019

EO297 Room UB99(B1) STATISTICAL ANALYSIS FOR BIG DATA Chair: Jinfeng Xu

E0537: **Large dynamic covariance matrix modeling via adaptive local/global thresholdings**
Presenter: **Shaojun Guo**, Renmin University of China, China
The main aim is to develop a general class of sparse dynamic covariance matrix models for capturing the dynamic information in large covariance matrices. We apply adaptive local/global thresholding techniques to recover sparsity structures. We consider a bias-corrected local linear smoother to estimate the large covariance matrix locally, which is shown to be very useful for threshold selection in adaptive/global thresholdings. The nonasymptotic concentration bounds of the resulting estimators under different functional sparsity scenarios are established. We also demonstrate that our proposed method significantly outperforms possible competitors through intensive simulation studies and is also applied to a real data set, revealing some interesting findings.

E0664: **Linear discriminant analysis with high dimensional mixed variables**
Presenter: **Binyan Jiang**, The Hong Kong Polytechnic University, Hong Kong
Discriminant analysis is considered with both high dimensional discrete and continuous variables. Under the location-scale Gaussian model, we show that the optimal classification direction relies on the continuous variables via a functional coefficients and the contribution of the discrete variables appears in the intercept. Direct estimation methods is then proposed to estimate the directions and the intercepts. Asymptotic results on the estimation accuracy and the misclassification rates are established and preliminary numerical results will be presented to illustrate the competitive performance of our approach.

E0691: **Joint variable screening in accelerated failure time models**
Presenter: **Jinfeng Xu**, University of Hong Kong, Hong Kong
Variable screening has gained increasing popularity in high-dimensional survival analysis. Most existing methods for variable screening with survival data suffer from that variable importance is assessed based on marginal models that relate the time-to-event outcome to each variable separately, implying that the relevance of one variable is examined when other variables are excluded. These methods will preclude variables that only manifest their influence jointly and may retain irrelevant variables that are correlated with relevant ones. To circumvent these difficulties, we propose a new approach to evaluating joint variable importance in censored accelerated failure time models. We establish the sure screening properties of the proposed approach and demonstrate its effectiveness through simulation studies and a real data application. A novel stability selection-based procedure is also proposed for tuning.

EO131 Room S101 NEW ADVANCES IN STATISTICAL LEARNING AND THEIR APPLICATIONS Chair: Yuan Ke

E0229: **Online experiment design for mapping large-scale neural circuits**
Presenter: **Shizhe Chen**, University of California, Davis, United States
High throughput circuit mapping experiments are considered where subthreshold, postsynaptic responses of one neuron are recorded using whole-cell patch clamp, and optical stimulation is used to stimulate multiple genetically modified neurons per trial. In these experiments, we are interested in (i) inferring which neurons have synaptic connections with the patched neuron, and (ii) the properties of the presynaptic neurons. However, the amount of data one can collect is paltry compared to the extent of neural circuits because the preparations are short-lived. In addition, the patched neuron's responses are subject to intrinsic stochasticity due to the low spatial resolution of the optical stimulation and the biological variability in the responses of individual neurons to the optical stimulation. We propose an online procedure that automatically designs future trials during the experiment. Our procedure first focuses on detecting and eliminating disconnected cells with multi-spot stimulations, then learns properties of the connected cells with precise single-spot stimulations. To this end, we develop a robust method for fitting a physiobiologically plausible model for the observed postsynaptic events, which is used to learn the properties of the few connected cells. We derive a simplified working model that is fast to fit, which is used to detect the many disconnected cells.

E0685: **Measurement errors in the instrumental variable model with binary variables**
Presenter: **Zhichao Jiang**, Harvard University, United States
Instrumental variable methods can identify causal effects even when the treatment and outcome are confounded. We consider scenarios with imperfect measurements of the binary instrumental variable, treatment or outcome. For non-differentially measurement errors, we show that the measurement error of the instrumental variable does not bias the estimate, the measurement error of the treatment biases the estimate away from zero, and the measurement error of the outcome biases the estimate toward zero. Moreover, we derive sharp bounds on the causal effects without additional assumptions. These bounds are informative because they exclude zero. We also consider differential measurement errors, and focus on sensitivity analyses in those settings.

E0690: **Long-run covariance matrix estimator for high-dimensional time series**
Presenter: **Haotian Xu**, University of Geneva, Switzerland
Co-authors: Yuan Ke
An estimator of long-run covariance matrix estimator for high-dimensional stationary time series is proposed. This estimator can be represented as the sum of autocovariance matrix estimators up to lag M. Generalizing the idea of adaptive Huber regression to dependent data, we study the nonasymptotic deviation properties of our estimator under the functional dependence measure. Moreover, in order to be user-friendly in practice, we provide the strategy of deciding the lag M and give the optimal tuning parameters which depend on the sample size, dimensionality, moment and the dependence. Our result allows heavy-tailed marginal distribution and the dimension to be increasing exponentially with the sample size.

EO175 Room S102 BIOSTATISTICS Chair: Jin-Jian Hsieh

E0248: **Efficient estimation of a hazard-based partial sufficient dimension reduction model for right-censored data**
Presenter: **Ming-Yueh Huang**, Academia Sinica, Taiwan
In many applications, it is important to summarize the hazard ratio of certain primary exposure variables, while controlling for many other covariates flexibly. When the number of controlled covariates is large, existing methods usually lead to stringent parametric assumptions or unstable nonparametric estimation. To address this issue, we introduce partial sufficient dimension reduction for survival data by introducing a nested family of multivariate baseline proportional hazards models. The family contains the Cox proportional hazards model and the continuously stratified proportional hazards model as special cases. The model maintains the practically desirable hazard-ratio interpretation of target parameters, while allowing data-adaptive dimension reduction of multi-dimensional covariates to reduce the effect of curse of dimensionality. The goal is to strike a balance between flexibility and parsimony, similar to the existing partial sufficient dimension reduction methods for uncensored data. Under

the proposed model, we characterize the semiparametric efficiency bound and propose an efficient estimator. The efficiency gain compared to the continuously stratified proportional hazards model is also proved.

E0298: **A Bayesian algorithm for case-control association study with copy number variation**
Presenter: **Yu-Chung Wei**, Feng Chia University, Taiwan
Copy number variations (CNVs) are genomic mutations consisting of abnormal numbers of gene fragment copies. Current algorithms for CNV association study for whole genome sequencing are restricted to a specific size or common/rare CNVs. We propose a Bayesian marker-level testing procedure to detect disease-associated CNVs. First, the absolute copy number of each window is estimated from sequencing read depths for every sample. And then the absolute copy numbers from case and control are compared to select candidate disease-associated windows. Finally, the information from neighboring windows is combined to identify the disease-associated copy number regions. We evaluate the performance and compare with competing approaches via simulations and real data.

E0504: **On summary models for meta-analysis of diagnostic studies**
Presenter: **ShengLi Tzeng**, National Sun Yat-sen University, Taiwan
Summarizing performance metrics is essential to a systematic review of a diagnostic performance. When a gold standard is available, every individual study in a meta-analysis has merely four numbers from a dichotomized test, i.e., number of true positives, false negatives, true negatives, and false positives. The goal of such a meta-analysis is to produce the summarized sensitivity, specificity, and a summary line of the receiver operating characteristic (ROC) curve. There are various summary models for the performance metrics in the literature, and hence one inevitably faces the problem of which model(s) to use. However, the suitability of existing asymptotic model selection approaches is doubtful since the meta-analysis here typically has a rather small sample size. Directly applying information criteria may fails to select or to combine good models for meta-analyses. Novel simulation scenarios mimicking a typical data collection process are conducted. The simulation avoids generating data from a certain model that would be biased towards specific assumptions. Even though the data never follow any probabilistic mechanism of a candidate model, we do know the underlying ROC curve. Then several model determination methods were compared accordingly, including simply using the most popular one, model averaging, and criterion-based model selections. Some suggestion and discussion about the better model determination strategy will be given based on the simulation results.

EO273 Room S104 HIGH DIMENSIONALITY AND TIME SERIES Chair: Chi Seng Pun

E0165: **Integrated volatility matrix estimation with nonparametric eigenvalue regularization**
Presenter: **Cheng Qian**, London School of Economics and Political Science, United Kingdom
Co-authors: Clifford Lam
When the number of assets p is large compared with the sample size n, trivially extending univariate volatility matrix estimators is not advised, since they are all modifications of a sample realized covariance matrix, which suffers from bias in its extreme eigenvalues under the high dimension framework. Without implicit assumptions on the structure of the true integrated volatility matrix, we propose a nonparametric eigenvalue regularization on the multi-scale (NER-MSRVM), the kernel (NER-KRVM) and the pre-averaging (NER-PRVM) realized volatility matrix estimators. We show that our regularization can shrink nonlinearly those extreme eigenvalues on all three estimators, and are positive definite in probability. Incidentally, the bias-corrected versions of kernel and pre-averaging estimators, which have faster rate of convergence at $n^{-1/4}$, but are not guaranteed to be positive definite, are now regularized to be positive definite in probability, and we prove their rates of convergence to an "ideal" estimator under the spectral norm are also at $n^{-1/4}$ in high dimension scenario. Our results are extended to a jump-diffusion model for the log-price processes with jumps removed using a previous wavelet method. All methods are applied to a simulated data and real data.

E0204: **A sparse learning approach to relative-volatility-managed portfolio selection**
Presenter: **Chi Seng Pun**, Nanyang Technological University, Singapore
A self-calibrated sparse learning approach is proposed for estimating a sparse target vector, which is a product of a precision matrix and a vector, and investigates its application to finance to provide an innovative construction of relative-volatility-managed portfolios. The proposed iterative algorithm, called DECODE, jointly estimates a performance measure of the market and the effective parameter vector in the optimal portfolio solution, where the relative-volatility timing is introduced into the risk exposure of an efficient portfolio via the control of its sparsity. The portfolio's risk exposure level, which is linked to its sparsity in the proposed framework, is automatically tuned with the latest market condition without using cross-validation. The algorithm is efficient as it costs only a few computations of quadratic programming. We prove that the iterative algorithm converges and show the oracle inequalities of the DECODE, which provide sufficient conditions for a consistent estimate of an optimal portfolio. The algorithm can also handle the curse of dimensionality that the number of training samples is less than the number of assets. Our empirical studies of over-12-year backtest illustrate the relative-volatility timing feature of the DECODE and the superior out-of-sample performance of the DECODE strategy, which beats the equally-weighted strategy and improves over the shrinkage strategy.

E0528: **Constrained sparse network autoregressive model in high dimensions**
Presenter: **Nazgul Zakiyeva**, National University of Singapore, Singapore
Co-authors: Ying Chen, Thorsten Koch , Bangzhu Zhu
A Constrained Sparse Network AutoRegressive (CSNAR) model is proposed for large scale network, where supply and demand in the network are balanced. Simultaneously considering three challenges in the CSNAR framework, namely 1) high dimensionality and 2) unknown adjacency matrix and 3) equality constraints, we estimate the constrained adjacency matrix under sparsity assumption. After discussing the estimation procedures and theoretical properties of the CSNAR model, we demonstrate its forecasting performance using real life network data and compare with alternative high dimensional models.

EO101 Room S106 STATISTICAL ANALYSIS FOR COMPLEX HIGH DIMENSIONAL DATA Chair: Heng Lian

E0315: **Confidence region of singular subspaces**
Presenter: **Dong Xia**, Hong Kong University of Science and Technology, Hong Kong
Spectral methods are prevalent and powerful in low-rank statistical models. Unlike the elegantly established results on the accuracy of point estimates, little is known about statistical inference of low-rank models. We will introduce a simple technical tool for investigating the empirical singular subspaces. Basically, an explicit representation formula is developed for the empirical spectral projector. We then prove the normal approximation of the joint projection distance between the empirical singular subspace and the true singular subspace when the noise matrix has i.i.d. Gaussian entries. We calculate, up to the fourth order, the approximation of the expected joint projection distance. Data-dependent confidence regions are then proposed which achieves any pre-determined confidence level asymptotically.

E0525: **Community detection for network with high dimensional attributes**
Presenter: **Wanjie Wang**, National University of Singapore, Singapore
Community detection in social network is a topic with much interest nowadays due to the high demand. With the observed connections between nodes, it is of interest to cluster the nodes into different communities, so that nodes within the same community have larger probability to connect. With the development of technology, the observed data are not only the connections between nodes, but also the attributes of each node. For example, the abstract of each paper in the paper citation network. With the attributes, some works present new methods for a better community detection result. However, most of the works are on low-dimensional attributes, while now the attributes are quite high-dimensional. We propose a new algorithm for the community detection problem for social network with high dimensional attributes, with good theoretical properties and simulation results.

E0619: **Generalization error bound of deep learning via spectral analysis and its application to model compression**
Presenter: **Taiji Suzuki**, University of Tokyo / RIKEN-AIP, Japan
To efficiently execute a high performance deep learning system on edge-computing devices, model compression methods have been gathering much attention. However, there have been a limited number of studies that simultaneously offer a practically effective compression method and its rigorous theoretical back-ground that guarantees its compression ability in connection with generalization ability. To resolve this issue, we develop a new theoretical frame-work for model compression,and propose a new pruning method called Spectral-Pruning based on the theory. We define "degree of freedom" to quantify an intrinsic dimensionality of the model by using the eigenvalue distribution of the covariance matrix across the internal nodesand show that the compression ability is essentially controlled by this quantity. For this bound, the theory of the kernel quadrature rule plays the essential role. Along with this, we give a sharp generalization error bound of the compressed model, and characterize a bias-variance trade-off induced by the compression procedure.We apply our method to several datasets to justify our theoretical analyses and show that the proposed method achieves the state-of-the-art performance.

EO183 Room S1A01 EMERGING METHODS IN DATA SCIENCE Chair: Yuan Ke

E0482: **Random regression coefficients model for small area estimation**
Presenter: **Gauri Datta**, University of Georgia and US Census Bureau, United States
Co-authors: Hee Cheol Chung, Jerry Maples
Small area estimation methodology has been found to be an indispensable tool to reliably estimate various government statistics such as income, employment, poverty status, health care availability, disease prevalence, etc. for various segments of a population. Sample surveys have been effectively used to provide suitable statistics not only for the population as a whole targeted by a survey but also for a variety of subpopulations, often called domains or areas. Domains may be geographical areas such as states, districts, or socio-demographic groups or other subpopulations. Often samples collected from some domains are not large to produce on their own accurate statistics for those domains. These domains are considered "small areas" which need alternative estimates with better accuracy. Small area estimation methodology benefited immensely from Stein's shrinkage estimation by borrowing strength from the direct estimates of the other areas and appropriate auxiliary variables available for all the areas. A part of the variability of the population small area means is explained through a suitable regression model based on auxiliary variables. We present a noninformative Bayesian analysis of random regression coefficients model to produce reliable point estimates of population means. Our method generalizes the standard Fay-Herriot model.

E0669: **Robust estimation of conditional variance of time series using density power divergences**
Presenter: **TN Sriram**, University of Georgia, United States
Suppose Z_t is the square of a time series Y_t whose conditional mean is zero. We do not specify a model for Y_t, but assume that there exists a $p \times 1$ parameter vector $\blacksquare$ such that the conditional distribution of $Z_t|\mathbb{Z}_{t-1}$ is the same as that of $Z_t|\lesseqgtr^T\mathbb{Z}_{t-1}$, where $\mathbb{Z}_{t-1} = (Z_{t-1},\ldots,Z_{t-p})^T$ for some lag $p \geq 1$. Consequently, the conditional variance of Y_t is some function of $\lesseqgtr^T\mathbb{Z}_{t-1}$. To estimate $\lesseqgtr$, we propose a robust estimation methodology based on Density Power Divergences (DPD) indexed by a tuning parameter $\alpha \in [0,1]$, which yields a continuum of estimators, $\{\widehat{\lesseqgtr}_\alpha; \alpha \in [0,1]\}$, where α controls the trade-off between robustness and efficiency of the DPD estimators. For each α, $\widehat{\lesseqgtr}_\alpha$ is shown to be strongly consistent. We develop data-dependent criteria for the selection of optimal α and lag p in practice. We illustrate the usefulness of our DPD methodology via simulation studies for ARCH-type models, where the errors are drawn from a gross-error contamination model and the conditional variance is a linear and/or nonlinear function of $\lesseqgtr^T\mathbb{Z}_{t-1}$.

E0798: **A minimax hypothesis test in smoothing spline ANOVA models**
Presenter: **WenXuan Zhong**, University of Georgia, United States
Smoothing spline ANOVA (SSANOVA) model has been a popular choice for building the nonparametric regression model with multiple predictors. Extensive research efforts have been devoted to model fitting. However, testing the significance of components in SSANOVA is still lacking. We will present a test for testing the significance of the interaction in a bivariate SSANOVA model. We derive the limiting distribution of our test statistics which unveils a new version of Wilks phenomenon. We prove that the proposed test achieves the minimax rate for hypothesis testing. Simulation studies are conducted to investigate the empirical performance of the proposed test in the context. Application analysis offers numerical support to DNA methylation and neuroimaging studies.

EO312 Room AT241 RECURRENT EVENT ANALYSIS UNDER INFORMATIVE CENSORING Chair: Chiung-Yu Huang

E0178: **Flexible accelerated time modeling of recurrent events data in the presence of a dependent terminal event**
Presenter: **Limin Peng**, Emory University, United States
Accelerated time modeling provides a useful prospective for assessing covariate events on recurrent event outcomes that have physical interpretations. The generalized accelerated recurrence time model (GART) significantly extends the traditional accelerated failure time model for recurrent events, offering extra flexibility in accommodating heterogeneous covariate effects. In practice, the observation of recurrent events is often stopped by a dependent terminal event. To address such a realistic scenario, we discuss two extensions of the GART models that can appropriately account for the presence of a dependent terminal event. We develop estimation and inference procedures for both extensions of the GART model, and establish desirable asymptotic properties. The proposed estimation and inference procedures can be readily implemented based on existing software. Simulation studies demonstrate satisfactory finite-sample performance of the proposed methods. We illustrate the proposed methods via an application to a dataset from the Cystic Fibrosis Foundation Patient Registry (CFFPR).

E0223: **Rank-based inference for censored quantile regression**
Presenter: **Tony Sit**, The Chinese University of Hong Kong, Hong Kong
A class of quantile regression models is proposed for time-to-event observations subject to censoring. By observing similarities between the AFT and the quantile regression models and borrowing techniques that have long been developed for the AFT model to the current setup, our framework aims at developing a more efficient for the estimating parameters of interest. Asymptotic properties including consistency and weak convergence

of the proposed estimator are established via the martingale-based argument. Numerical studies are presented to illustrate the outperformance of the proposed estimator over existing contenders under various settings.

E0818: **U-processes of diverging dimensional parameters**
Presenter: **Yuanyuan Lin**, The Chinese University of Hong Kong, Hong Kong
Co-authors: Zhanfeng Wang, Wenxin Liu, Qi-Man Shao
A general theory for estimation methods based on U-process typed objective functions with parameter of increasing dimensions is derived. Under reasonable conditions, we establish a maximal inequality for degenerate U-processes with increasing dimensional parameter, and prove the sqrt(n/p)-consistency and asymptotic normality of each component of the resultant estimator of their linear combinations. Moreover, with increasing dimensionality, a general theory for random weighting resampling for U-process typed objective function is proposed and justified rigorously for inference. The theory is illustrated with the Han's maximum rank correlation estimation and the Gehan's estimation. Particularly, when the dimension of the parameter space diverges at the order of $o(\sqrt{n}/\log(n))$, the Han's maximum rank correlation estimator and the Gehan's estimator are shown to be $\sqrt{n/p}$-consistent, and their component-wise asymptotic normality remains valid.

EO227 Room AT335 NEW APPROACHES IN BAYESIAN ECONOMETRIC MODELING Chair: Thomas Zoerner

E0535: **Credit market sentiments as driving force of economic fluctuations**
Presenter: **Maximilian Boeck**, Vienna University of Economics and Business, Austria
The role of credit market sentiments and investor beliefs on credit cycle dynamics and their propagation to business cycle fluctuations are investigated. Building strongly on recent theoretical contributions that introduce 'animal spirits' or credit market sentiments as driving force of cyclicality and instability in the financial market. Those sentiments can be characterised by extrapolation of past credit market outcomes and exhibit a mean-reverting behaviour. We model the Euro area credit and real economy within a Bayesian vector autoregressive framework, where non-linearities are introduced to the model allowing for transitions between an 'optimistic' and 'pessimistic' regime. Using data from the Bank Lending Survey of the European Central Bank allows to capture forward-looking credit market sentiments, which govern the transitional dynamics between the regimes. A flexible Bayesian approach permits fully probabilistic inference and provides regime probabilities that the credit market behaviour is 'optimistic' or 'pessimistic'. Moreover, due to the high dimensionality of the proposed model the Normal-Gamma shrinkage prior is applied in order to get more precise estimates. Furthermore, this framework is suited for investigating the impact of accommodative monetary policy in both regimes of the credit cycle by means of structural impulse response analysis.

E0536: **Approximate Bayesian inference in a semiparametric copula model of income and wealth**
Presenter: **Anna Stelzer**, Vienna University of Economics and Business, Austria
The aim is to evaluate economic well-being by means of estimating the joint distribution of income and wealth. While both income and wealth are of crucial interest individually, analysing them together draws a fuller picture of people's economic possibilities. Recent studies investigate the dependence structure between income and wealth by using copula models, which allow estimating both marginal distributions as well as the joint distribution separately. The tail behaviour of the joint distribution is of special interest, as it indicates whether the relationship between income and wealth for individuals in the lower part of the distribution is different from the ones in the top part. Measures of dependence, both overall and in the upper as well as in the lower tail of the distribution, are estimated in a flexible way. Applying a Bayesian approach to a semiparametric copula model and using HFCS data for several Euro Area countries, inference on summaries of the dependence structure can be obtained by only partially specifying the model. By using this approach, choosing a complete copula structure and the risk of misspecification can be avoided while still answering important questions about the joint distribution of income and wealth.

E0648: **Sparse finite mixture-of-experts modeling**
Presenter: **Gregor Zens**, Vienna University of Economics and Business, Austria
A sparse Bayesian mixture-of-experts model is developed where the number of components is estimated endogenously. Bayesian inference is carried out through the implementation of a Gibbs sampler. By constructing a suitable prior density for the component weights, we allow the sampler to apply shrinkage to the component weights when necessary. Moreover, additional information is allowed to enter the model via a set of independent covariates. This is achieved by combining the clustering information of the mixture-of-experts linking function and the idea of a deliberately overfitting mixture model. An identified model is obtained by relabeling the MCMC output in the point process representation of the draws. The model is evaluated in a simulation setup with artificial data. In a real data application, the model is used to find homogenous clusters of women in Mozambique based on their information sources on HIV.

EO103 Room AT337 MIXED EFFECTS MODEL AND MODEL SELECTION Chair: Dalei Yu

E0494: **Conditional Akaike information under covariate shift with application to small area estimation**
Presenter: **Yuki Kawakubo**, Chiba University, Japan
Co-authors: Shonosuke Sugasawa, Tatsuya Kubokawa
The problem of selecting explanatory variables of fixed effects is considered in linear mixed models under covariate shift, which is when the values of covariates in the model for prediction differ from those in the model for observed data. We construct a variable selection criterion based on the conditional Akaike information. We focus especially on covariate shift in small area estimation and demonstrate the usefulness of the proposed criterion. In addition, numerical performance is investigated through simulations, one of which is a design-based simulation using a real dataset of land prices.

E0680: **Modelling spatially correlated binary data**
Presenter: **Youjun Huang**, Sichuan University, China
Co-authors: Gabrielle Kelly, Jianxin Pan
Generalized estimating equations lead to consistent estimators of the mean parameters for spatially correlated binary data, but the estimation of covariance matrix is also of interest in spatial data analysis. A specific parametric form is proposed to model the correlation matrix for spatially correlated binary data. An iterative approach based on generalized estimating equations is developed to estimate the mean and correlation parameters simultaneously. Asymptotic normality for the estimators of the mean and correlation parameters is provided. Simulation studies are conducted through considering various model parameters such as different working correlation matrices, correlation parameters and dimensions of the mean parameters. The proposed approach is used to analyze the spatial bovine tuberculosis infection data in Ireland, aiming to quantify the influence of some important factors on the infection for both badgers and cattle, as well as the correlation between their setts and herds.

E0583: **Robust estimation and confidence interval in meta-regression models**
Presenter: **Dalei Yu**, Yunnan University of Finance and Economics, China
Co-authors: Chang Ding, Na He, Ruiwu Wang, Xiao-Hua Zhou, Lei Shi
Meta-analysis provides a quantitative method for combining results from independent studies with the same treatment. However, existing estimation

methods are sensitive to the presence of outliers in the datasets. We study the robust estimation for the random effects meta-regression. Hubers rho function and Tukeys biweight function are adopted to derive the formulae of robust maximum likelihood estimators. The corresponding algorithms are developed. The asymptotic confidence interval and second-order-corrected confidence interval are investigated. Simulation studies show that the robust estimators are promising and outperform the conventional maximum likelihood and restricted maximum likelihood estimators when outliers exist in the dataset. Results in case studies further support the eligibility of our methods in practical situations.

EO247 Room U302 MODEL AVERAGING AND RELATED TOPICS Chair: Fang Fang

E0442: **Semiparametric model averaging prediction for dichotomous response**
Presenter: **Fang Fang**, East China Normal University, China
Co-authors: Jialiang Li, Xiaochao Xia
Model averaging has attracted abundant attentions from researchers in the past decades as it becomes a powerful forecasting tool in areas such as econometrics, social sciences and medicine. So far, most developed model averaging methods focus only on either parametric models or nonparametric models with a continuous response. We propose a semiparametric model averaging prediction (SMAP) method for a dichotomous response. The idea is to approximate the unknown discrimination score function by a linear combination of one-dimensional marginal score functions. The weight parameters involved in the approximation are constructed by an initial stage nonparametric smoothing estimation of the marginal scores and then applying the familiar parametric model averaging on the dichotomous response based on the likelihood estimation. The proposed model averaging procedure provides more flexibility than parametric models while being more stable than a fully nonparametric approach. Some theoretical properties are investigated and the weight optimality is established under certain technical assumptions. In particular, the weight optimality result requires a condition that reflects the difficulty of model averaging with nonparametric estimates. Empirical evidences from simulation studies and a real data analysis are presented to illustrate our methods.

E0591: **Deviance information criterion for model selection: Justification and variation**
Presenter: **Tao Zeng**, Zhejiang University, China
Co-authors: Yong Li, Jun Yu
Deviance information criterion (DIC) has been extensively used for making model selection based on the MCMC output. Although it is understood as a Bayesian version of AIC, a rigorous justification has not been provided in the literature. We show that when the plug-in predictive distribution is used, DIC can have a rigorous decision-theoretic justification in a frequentist setup. Under a set of regularity conditions, we show that DIC chooses a model that gives the smallest expected Kullback-Leibler divergence between the data generating process (DGP) and the plug-in predictive distribution asymptotically. An alternative expression for DIC, based on the Bayesian predictive distribution, is proposed. The new DIC has a smaller penalty term than the original DIC and is very easy to compute from the MCMC output. It is invariant to reparameterization and yields a smaller expected loss than the original DIC asymptotically.

E0205: **Frequency-domain cross-validation for determining the number of the common factors in factor models**
Presenter: **Natalia Sirotko-Sibirskaya**, University of Bremen, Germany
A frequency-domain-based cross-validation (FDCV) criterion is proposed to determine the number of common factors driving the observable multivariate data process with respect to the appropriately defined loss function. The suggested method is based on the theoretical property of the coefficients of the Fourier transforms which are known to be approximately independent under certain conditions. An alternative scheme for dealing with missing values in cross-validation is suggested under the assumption that the empirical Fourier transforms of the time series are smooth functions of frequency. Properties of the proposed criterion are studied both at the theoretical level and in simulations. The performance of the method is tested on the U. S. macroeconomic and financial data and compared with other commonly used criteria on determining the number of common factors in static and dynamic factor models.

EO097 Room U414 RECENT ADVANCES IN ECONOMETRICS Chair: Hon Ho Kwok

E0470: **Identification and estimation of linear social interaction models**
Presenter: **Hon Ho Kwok**, The University of Hong Kong, Hong Kong
Firstly, the identification conditions for higher-order social interaction models are considered. In the case where social effects depend on the distance between individuals, the upper bounds on the network diameters for non-identified models are derived. Many network properties of non-identified models in the literature can be derived from these upper bounds. Fixed effect elimination methods which require less restrictive identification conditions are analyzed. Secondly, estimation with panel data is considered. An estimator is developed which is computationally simple and asymptotically as efficient as the maximum likelihood estimator under normality.

E0715: **Two stage lasso-based least square estimation of spatial autoregressive model with many instrument variables**
Presenter: **Pu Wang**, The University of Hong Kong, China
A lasso and post-lasso based two stage least square estimation of spatial autoregressive models are studied when some or all regressors are endogenous in the presence of many instruments. In order to handle the bias-variance trade off caused by many instruments, we use Lasso and post Lasso methods in the first-stage to find the most informative instruments and obtain prediction of conditional expectations of endogenous variables given instruments. We show that if the conditional expectation is approximate sparse, i.e., only a small set of instruments can explain the most portion of conditional expectation, our Lasso based estimation is root n consistent and asymptotically normal. The method will be valid even when the number of instruments increases at the same rate or faster than the sample size.

E0720: **On Obamacare: A fuzzy difference-in-discontinuities approach**
Presenter: **Guy Tchuente**, University of Kent, United Kingdom
The use of fuzzy regression-discontinuity design is explored in the context where multiple treatments are applied at the threshold. It derives the conditions for identification of the effects of one of the treatments. The identification result shows that, under the strong assumption that the change in the probability of treatment at the cut off is equal across treatments, a difference-in-discontinuities estimator identifies the treatment effect of interest. Point identification of the treatment effect using fuzzy difference-in-discontinuities is impossible if the changes in the treatment probabilities are not equal across treatments. Using data from the National Health Interview Survey (NHIS), we apply this new identification strategy to evaluate the causal effect of the Affordable Care Act (ACA) on older Americans' health care access and utilization. Our results suggest that the implementation of the Affordable Care Act has (1) led to 5% increase in the hospitalization rate of elderly Americans, (2) increased by 3.6% the probability of delaying care for cost reasons, and (3) exacerbated cost-related barriers to follow-up and continuity of care–7% more elderly could not afford prescriptions, 7% more could not see a specialist and, 5.5% more could not afford a follow-up visit- as a result of the ACA.

EO019 Room U517 NEW DEVELOPMENTS IN TIME SERIES ANALYSIS Chair: Ke Zhu

E0300: **Spectral distribution of the sample covariance of high-dimensional time series with unit roots**
Presenter: **Chen Wang**, University of Hong Kong, Hong Kong
Co-authors: Alexei Onatski

The aim is to study the empirical spectral distributions of two sample-covariance-type matrices associated with high-dimensional time series with unit roots. The first matrix is $S = XX'/T$; where X is an $n \times T$ data matrix with rows represented by n i.i.d. copies of T consecutive observations of a difference-stationary process. The second matrix is $W = n\int_0^1 B_n(t)B_n(t)'dt$; where $B_n(t)$ is an n-dimensional vector with i.i.d. Brownian motion components. We show that, as n and T diverge to infinity proportionally, the two distributions weakly converge to non-random limits. The limit corresponding to S has a density $f(x)$ that decays as $x^{-3/2}$ when $x \to \infty$. The limit corresponding to W is a Feller-Pareto distribution.

E0314: **A novel partial-linear single-index model for time series data**
Presenter: **Lei Huang**, Southwest Jiaotong University, China

Partial-linear single-index models have been widely studied and applied, but their current applications to time series modeling still need some strong and inappropriate assumptions. A novel method which relaxes those assumptions is proposed. It extends the applicability of partial-linear single-index models to time series modeling, taking both lag variables and autocorrelated errors into consideration. An estimation procedure based on Whittle likelihood is proposed and some asymptotic properties of the corresponding estimators are derived. In addition, some simulation studies are conducted to elaborate that the proposed model is necessary in certain situations. The proposed models are also shown to be useful and reasonable in real data analysis, indicating the feasibility and practicability of the proposed estimation method.

E0339: **Non-standard inference for augmented double autoregressive models with null volatility coefficients**
Presenter: **Ke Zhu**, University of Hong Kong, Hong Kong

An augmented double autoregressive (DAR) model is considered, which allows null volatility coefficients to circumvent the over-identification problem in the DAR model. Since the volatility coefficients might be on the boundary, the statistical inference methods based on the Gaussian quasi-maximum likelihood estimation (GQMLE) become non-standard, and their asymptotics require the data to have a finite sixth moment, which narrows applicable scope in studying heavy-tailed data. To overcome this deficiency, a systematic statistical inference procedure is developed based on the self-weighted GQMLE for the augmented DAR model. The entire procedure is valid as long as the data is stationary, and its usefulness is illustrated by simulation studies and one real example.

EO181 Room U602 RECENT DEVELOPMENTS IN IMAGES, NETWORKS, AND HIGH-DIMENSIONAL DATA ANALYSIS Chair: Qiang Liu

E0275: **Accelerating Metroplis-within-Gibbs sampler with localized computations of differential equations**
Presenter: **Qiang Liu**, National University of Singapore, Singapore
Co-authors: Xin Tong

Bayesian inverse problem is widely encountered when quantifying uncertainty for underlying parameters in practice. For high dimensional spatial models, classical Markov chain Monte Carlo (MCMC) methods are ususlly slow to be applied, while it has been shown that Metropolis-within-Gibbs (MwG) sampling works when the parameters are locally dependent. The problem is that its implementation requires $O(n^2)$ calculation, where n is the number of parameters. Our target in this paper is to reduce the computation cost to an optimal scalability of $O(n)$, in the framework of stochastic differential equation (SDE) with local dependence structure. The key is that MwG proposal is only different from the original at local entries, and the difference caused also evolves locally. This inspires us to approximate the solution for the proposal with a surrogate updated only within a local domain, which brings down the computation to our targeting level. Both theoretically and numerically, we prove that the induced errors can be controlled by the local domain size. Implementions of our computation scheme by using Euler-Maruyama and 4th order Runge-Kutta method are also discussed. We demonstrate the finite sample performance of our method in numerical examples of Lorenz 96 and a linear stochastic flow model.

E0340: **Signal subgraph learning for longitudinal structural brain networks**
Presenter: **Lu Wang**, Central South University, China

Modern neuroimaging technologies, combined with state-of-the-art data processing pipelines, have made it possible to collect longitudinal observations of an individual's brain connectome at different ages. It is of substantial scientific interest to study how brain connectivity varies over time in relation to human cognitive traits. In brain connectomics, the structural brain network for an individual corresponds to the strength of connection between each pair of brain regions. We propose a symmetric bilinear logistic regression to learn a set of small outcome-relevant subgraphs from subjects' longitudinal structural brain networks as well as estimating the time effects of the subgraphs. The clique structure of the extracted signal subgraphs can be related to some neurological circuits and hence gives interpretable results. Time effect of each signal subgraph reflects how the predictive effect on the outcome varies with age, which may guide the optimal age for taking a brain scan to diagnose a neurological disorder. We apply the method to longitudinal brain connectome and cognitive capacity data.

E0385: **Posterior contraction and credible sets for filaments of regression functions**
Presenter: **Wei Li**, Syracuse University, United States
Co-authors: Subhashis Ghosal

A filament consists of local maximizers of a smooth function when moving in a certain direction. Filamentary structures are important features of the shape of objects and considered as a useful lower dimensional characterization of multivariate data. There have been some recent theoretical studies of filaments in the nonparametric kernel density estimation context. We shall discuss a Bayesian approach to the filament estimation in regression context and present some results on posterior contraction rates obtained using a finite random B-splines series. Compared with the kernel-estimation method, the bias can be better controlled using the series method when the function is smoother, which allows obtaining better rates. In addition, we discuss a way to construct a credible set with sufficient frequentist coverage for the filaments and demonstrate the proposed method in simulations and one application to earthquake data.

EC354 Room U501 CONTRIBUTIONS IN RESAMPLING Chair: Hong-Dar Wu

E0197: **Frequency-based bootstrap methods for modeling risk and return of DC pension plan strategies (virtual presentation)**
Presenter: **Aya Ghalayini**, Lancaster University, United Kingdom
Co-authors: Rami Chehab, Richard Harris

The use of conventional bootstrap methods, such as Standard Bootstrap (SB) and Moving Block Bootstrap (MBB), to produce long-run returns to rank pension strategies based on its associated reward and risk, might be misleading. Therefore, a simple pension model, that is mainly concerned with the long-term accumulation of wealth, is used to assess, for the first time, different bootstrap methods in this context. One of the main findings is that the Multivariate Fourier Bootstrap gives the most satisfactory result in its ability to mimic the 'true' distribution using Cramer-von-mises statistics. Also, the disagreement in the pension literature on selecting the best pension plan strategy is addressed. A comprehensive study is presented to compare different strategies using different bootstrap procedures with different cash-flow performance measures across a range of countries. Indeed, the other finding is that bootstrap methods play a critical role in determining the optimal strategy. Additionally, different cash-flow performance (CFP) measures rank pension plans differently across countries and bootstrap methods.

E0190: **A frequency domain wild bootstrap for dependent data (virtual presentation)**
Presenter: **Rami Chehab**, University of Exeter, United Kingdom

A resampling method for stationary dependent time series is proposed which based on Rademacher wild bootstrap draws from the Fourier transform of the data. The main distinguishing feature of our method is that the bootstrap draws share their periodogram identically with the sample, implying good properties under dependence of arbitrary form. A drawback of the basic procedure, that the bootstrap distribution of the mean is degenerate, is overcome by a simple Gaussian augmentation. Monte Carlo evidence indicates a favourable comparison with alternative methods in tests of significance and of location in a regression model with autocorrelated shocks, and also of unit roots.

E0749: **On some resampling procedures with the empirical Beta copula**
Presenter: **Hideatsu Tsukahara**, Seijo University, Japan
Co-authors: Johan Segers, Anna Kiriliouk

Because of many nice properties of the empirical beta copula, it is reasonable to expect that our smoothing procedure might also have a beneficial effect on the accuracy of resampling schemes for the empirical copula process. More specifically, testing procedures based on the empirical copula typically rely on the bootstrap for the computation of the critical values of the test statistic. For finite samples, the accuracy is often not very good: the actual size of the test may differ greatly from the nominal one. We first show the asymptotic equivalence of several related bootstrapped processes. Then we investigate the accuracy of the bootstrap resampling schemes based on the empirical beta copulas with small sample sizes; we consider confidence intervals for some functionals such as rank correlation coefficients and dependence parameters of several well-known copula families, and goodness-of-fit testing problems including testing exchangeability. And we compare the performance of three methods: standard asymptotic approximation, standard bootstrapping and resampling from the empirical beta copula.

Authors Index