JOINT MEETING OF THE IASC-*ARS* INTERIM CONFERENCE 2024 AND CSAT 2024

2024 國際統計計算學會亞洲區計算統計會議 暨113年統計學術研討會

PROGRAM BOOK

Website

Program Book

Table of Contents

JOINT MEETING OF
THE IASC-ARS INTERIM CONFERENCE 2024 AND CSAT 2024
2024 國際統計計算學會亞洲區計算統計會議
暨 113 年統計學術研討會

Organized by (主辦單位)

◎ Master Program in Statistics of National Taiwan University (國立臺灣大學統計碩士學位學程)
◎ Directorate General of Budget, Accounting and Statistics (DGBAS) of Executive Yuan (行政院主計總處)
◎ The Asian Regional Section of the International Association for Statistical Computing (IASC-ARS) (國際統計計算學會亞洲分會)

Co-organized by (合辦單位)

◎ Ministry of Education (教育部)
◎ The Chinese Institute of Probability and Statistics, CIPS (中華機率統計學會)
◎ B.A.S. Coordination and Development Society (財團法人主計協進社)
◎ Chinese Statistical Association (Taiwan) (CSAT) (中國統計學社)

Supporting Organized by (協辦單位)

◎ Institute of Statistical Science, Academia Sinica (中央研究院統計科學所)
◎ Department of International Cooperation and Science Education, NSTC (國科會科教發展及國際合作處)
◎ Mathematics, Natural Science and Sustainable Research Promotion and Engagement Center, NSTC (國科會自然科學及永續研究推展中心數學組)

Local Organizing Committee (本地籌備委員會)

Dr. Chen-An Tsai	National Taiwan University, Taiwan
Dr. Hsin-Chou Yang	Academia Sinica, Taiwan
Dr. Chun-houh Chen	Academia Sinica, Taiwan
Dr. Ray-Bing Chen	National Cheng Kung University, Taiwan
Dr. Ci-Ren Jiang	National Taiwan University, Taiwan
Dr. Charlotte Wang	National Taiwan University, Taiwan

Scientific Program Committee (學術委員會)

Dr. Erniel Barrios	School of Business, Monash University Malaysia, Malaysia
Dr. Ray-Bing Chen	Department of Statistics, National Cheng Kung University, Taiwan
Dr. Ying Chen	Department of Mathematics, National University of Singapore (NUS), Singapore
Dr. Donguk Kim	College of Economics, Sungkyunkwan University, Korea
Dr. Yuichi Mori	Department of Management Okayama, University of Science, Japan
Dr. David Nott	Department of Statistics and Data Science, National University of Singapore, Singapore
Dr. Mike So	Department of Information Systems, Business Statistics and Operations Management, Hong Kong University of Science and Technology, Hong Kong

GENERAL INFORMATION

會場資訊

Conference Venue: GIS NTU Convention Center 集思台大會議中心

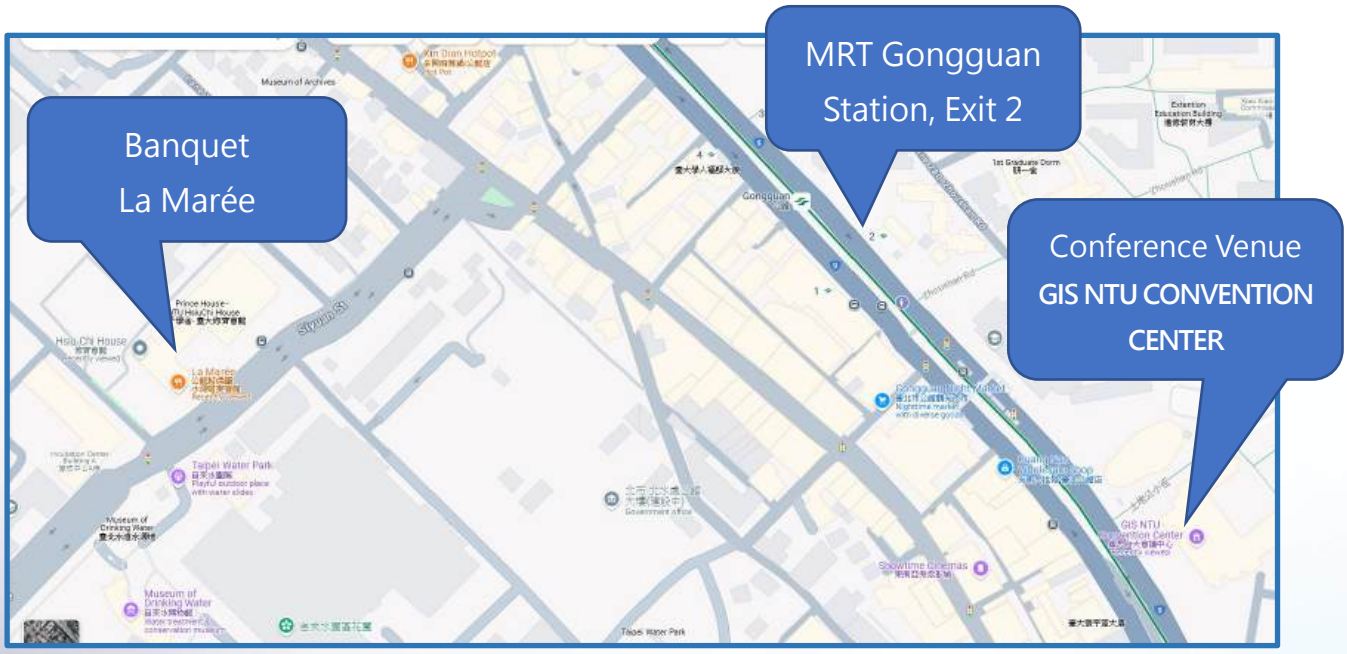
Conference Website:

https://www.stat.ntu.edu.tw/statweek2024

GETTING TO GIS NTU CONVENTION CENTER

MRT Gongguan Station, Exit 2

Songshan-Xindian Line (Line 3)：At the station exit 2, take the left turn to Roosevelt Rd. Here you will find the GIS NTU Convention Center on your left-hand side. (2 minutes walking)



Google Map Link

GIS NTU CONVENTION CENTER

Conference Venue

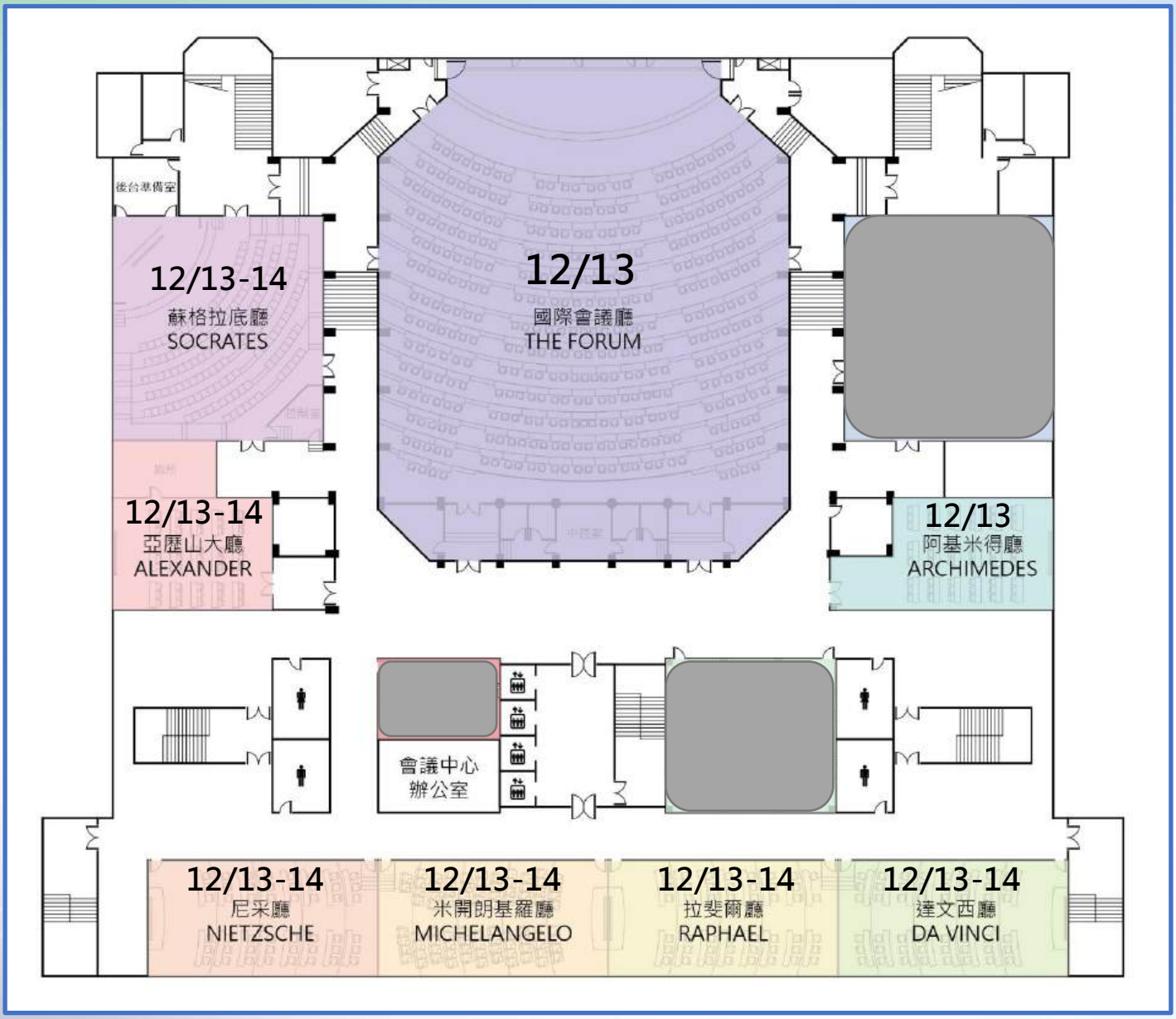
La Marée

Banquet

DECEMBER 13-14, 2024

GIS NTU CONVENTION CENTER

VENUE FLOOR PLAN 會場平面圖



SCIENTIFIC PROGRAM OVERVIEW

研討會議程

Scientific Program – Day 1

Venue: The GIS NTU Convention Center 集思台大會議中心

Friday, December 13, 2024

Time 時間	Activity 議程	Room 地點
08:20 - 17:00	Conference Registration 研討會註冊	B1F
08:50 - 09:30	**Opening Ceremony and Award Presentation of CSAT** **研討會開幕式致詞及中國統計學社頒獎** **Welcome Remarks 致詞貴賓** **Dr. Ray-Bing, Chen and Dr. Chen-An, Tsai** **陳瑞彬教授、蔡政安教授** LOC Chairman 籌備委員會主席 **Dr. Hung-Jen Wang 王泓仁教授** Vice President for Academic Affairs, National Taiwan University 台灣大學教務長 **Dr. Donguk, Kim 金東郁教授** IASC-ARS President 國際統計計算協會亞洲分會理事長 **Dr. Chun-houh, Chen 陳君厚教授** CSAT President 中國統計學社理事長	The Forum 國際會議廳
09:30 - 09:40	Coffee Break 中場休息	GIS Lobby 集思大廳
09:40 - 10:30	**Keynote Talk 1 大會主講 1** **Dr. Xuming He** President, International Statistical Institute Washington University in St. Louis Title: **Some Recent Developments in Expected Shortfall Regression** **Chair: Dr. Wen-Han Hwang 黃文瀚 教授** Director, Institute of Statistics and Data Science, National Tsing Hua University 清華大學統計與數據科學研究所所長	The Forum 國際會議廳

Time 時間	Activity 議程	Room 地點
10:30 - 10:50	Coffee Break 中場休息	GIS Lobby 集思大廳
10:50 - 12:20	**Parallel Sessions 同步演講場次 FAM-1**	
FAM-1-1	**Government Session 1 政府場次 1**	The Forum 國際會議廳
FAM-1-2	**Data Science and Statistical Challenges**	Socrates 蘇格拉底廳
FAM-1-3	**Statistical Analyses for Scientific Networks and Patent Networks**	Archimedes 阿基米德廳
FAM-1-4	**Recent Developments in Statistical Learning Theory and Applications**	Michelangelo 米開朗基羅廳
FAM-1-5	**Modeling in Complex Systems**	Raphael 拉斐爾廳
FAM-1-6	**Novel Statistical Models and Methods with Applications**	Nietzsche 尼采廳
FAM-1-7	**CSAT Thesis Award Session** **中國統計學社得獎論文場次**	Davinci 達文西廳
FAM-1-8	**Contributed Poster Presentation** **投稿海報場次**	Alexander 亞歷山大廳
12:20 - 13:30	Lunch Break (Boxed lunch provided) 午餐 (提供餐盒)	
13:30 - 15:00	**Parallel Sessions 同步演講場次 FPM-1**	
FPM-1-1	**Government Session 2 政府場次 2**	The Forum 國際會議廳
FPM-1-2	**Statistical Learning**	Socrates 蘇格拉底廳
FPM-1-3	**Statistical Methods and Applications in Clinical Trials**	Archimedes 阿基米德廳
FPM-1-4	**Fintech and Data Science**	Michelangelo 米開朗基羅廳
FPM-1-5	**Recent Advances in Regression-Based Approaches and Modelling Complex Data**	Raphael 拉斐爾廳

Time 時間	Activity 議程	Room 地點
FPM-1-6	**Advances in Reliability Science**	Nietzsche 尼采廳
FPM-1-7	**Bayesian Methods and Inference**	Davinci 達文西廳
FPM-1-8	**Clustering and Classification**	Alexander 亞歷山大廳
15:00 - 15:20	Coffee Break 中場休息	GIS Lobby 集思大廳
15:20 - 16:30	**Parallel Sessions 同步演講場次 FPM-2**	
FPM-2-1	**Machine Learning Algorithms in Analyzing Survey Data**	The Forum 國際會議廳
FPM-2-2	**Advanced Statistical Techniques for Multifaceted Data**	Socrates 蘇格拉底廳
FPM-2-3	**New Advances in Model Selection and Inference**	Archimedes 阿基米德廳
FPM-2-4	**Structural Equation Modeling**	Michelangelo 米開朗基羅廳
FPM-2-5	**Statistical Analysis of Complex Data**	Raphael 拉斐爾廳
FPM-2-6	**Statistical Machine Learning and Inference**	Nietzsche 尼采廳
FPM-2-7	**Innovative Designs and Analysis Methods for Clinical Trials and Biomedical Data**	Davinci 達文西廳
FPM-2-8	**Recent Advances on Interplay of Statistics and Optimization**	Alexander 亞歷山大廳
16:30 - 17:20	**Keynote Talk 2 大會主講 2** **Dr. Patrick John Fitzgerald Groenen** Erasmus University, Rotterdam Title: **Effective MM-Algorithms for Statistics and Machine Learning** **Chair: Dr. Donguk Kim 金東郁教授** IASC-ARS President 國際統計計算協會亞洲分會理事長	The Forum 國際會議廳
18:00 - 20:30	Banquet 晚宴	La Marée 水源文教會館

Scientific Program – Day 2

Venue: The GIS NTU Convention Center 集思台大會議中心

Saturday, December 14, 2024

Time 時間	Activity 議程	Room 地點
08:40 - 17:00	Conference Registration 研討會註冊	B1F
09:10 - 10:00	**Keynote Talk 3 大會主講 3** **Dr. Cathy W. S. Chen** Feng Chia University Title: **Advances in Spatial Integer-Valued Time Series Modeling for Dengue Fever** **Chair: Dr. Ray-Bing, Chen 陳瑞彬 教授** LOC Chairman 籌備委員會主席	Socrates 蘇格拉底廳
10:00 - 10:20	Coffee Break 中場休息	GIS Lobby 集思大廳
10:20 - 11:50	**Parallel Sessions 同步演講場次 SAM-1**	
SAM-1-1	**Young Scholar Session**	Socrates 蘇格拉底廳
SAM-1-2	**Inference on High Dimensional Covariance Matrix**	Nietzsche 尼采廳
SAM-1-3	**Particle Swarm Optimization and Its Application in Clinical Trials and Medical Imaging**	Michelangelo 米開朗基羅廳
SAM-1-4	**Advancements in Predictive Modeling and Data Integration with Applications**	Raphael 拉斐爾廳
SAM-1-5	**Dimension Reduction Methods**	Alexander 亞歷山大廳
11:50 - 13:30	Lunch Break (Boxed lunch provided) 午餐 (提供餐盒)	
13:30 - 15:00	**Parallel Sessions 同步演講場次 SPM-1**	
SPM-1-1	**Complex Data Analysis**	Socrates 蘇格拉底廳
SPM-1-2	**Recent Advances in Time Series and Spatial Statistics**	Nietzsche 尼采廳

Time 時間	Activity 議程	Room 地點
SPM-1-3	Advancements in Metaheuristic Algorithms and their Statistical Applications	Michelangelo 米開朗基羅廳
SPM-1-4	Advanced Methods for Complex Data Analytics in the AI Era	Raphael 拉斐爾廳
SPM-1-5	Recent Advances in Statistical Methods for Analyzing Time Series and Spatial Data	Alexander 亞歷山大廳
SPM-1-6	Recent Advances in Causal Inference and Applications to Biomedical Studies	Davinci 達文西廳
15:00 - 15:20	Coffee Break 中場休息	GIS Lobby 集思大廳
15:20 - 16:50	Parallel Sessions 同步演講場次 SPM-2	
SPM-2-1	Novel Approaches in Bayesian and Empirical Bayes Methods	Socrates 蘇格拉底廳
SPM-2-2	Causal Inference and Its Application on Statistics	Nietzsche 尼采廳
SPM-2-3	Innovative Statistical and Machine Learning Techniques for Enhanced Prediction and Optimization in Insurance, Finance, and Criminology	Michelangelo 米開朗基羅廳
SPM-2-4	Statistical and Mathematical Approaches in High Throughput Health Data Analytics	Raphael 拉斐爾廳
SPM-2-5	Causal Mediation analysis and Statistical Inference for Latent Class Models	Alexander 亞歷山大廳
SPM-2-6	Recent Advances in Symbolic Data Analysis and Computational Algorithms for Statistical Inference	Davinci 達文西廳
16:50 - 17:00	Closing Remarks 閉幕總結	

KEYNOTE TALK 1 \ 大會主講 1

Some Recent Developments in Expected Shortfall Regression

Xuming He, *President, International Statistical Institute* \ *Washington University in St. Louis*

Expected shortfall, measuring the average outcome (e.g., portfolio loss) above a given quantile of its probability distribution, is a common financial risk measure. The same measure can be used to characterize treatment effects in the tail of an outcome distribution, with applications ranging from policy evaluation in economics and public health to biomedical investigations. Expected shortfall regression is a natural approach of modeling covariate-adjusted expected shortfalls. Because the expected shortfall cannot be written as a solution of an expected loss function at the population level, computational as well as statistical challenges around expected shortfall regression have led to stimulating research. We discuss some recent developments in this area, with a focus on a new optimization-based semiparametric approach to estimation of conditional expected shortfall that adapts well to data heterogeneity with minimal model assumptions. The talk is based on joint work with Yuanzhi Li and Shushu Zhang.

KEYNOTE TALK 2 \ 大會主講 2

Effective MM-Algorithms for Statistics and Machine Learning

Patrick John Fitzgerald Groenen, *Erasmus University Rotterdam, Econometric Institute*

As an optimization method, majorization and minorization (MM) algorithms have been applied with success in a variety of models arising in the area of statistics and data science. A key property of majorization algorithms is guaranteed descent, that is, the function value decreases in each step. In practical cases, the function is decreased until it has converged to a local minimum. If the function is convex and coercive, a global minimum is guaranteed. The auxiliary function, the so-called majorizing function, is often quadratic so that an update can be obtained in one step. Here, we present a selection of useful applications of MM algorithms. We discuss its use multidimensional scaling, and in binary and multiclass classification such as logistic regression, multinomial regression, and support vector machines. In the case of regularized generalized canonical correlation analysis, its MM algorithm coincides with several partial least squares (PLS) algorithms thereby providing a previously unknown goal function for the PLS algorithms. We show how MM can also be effective in large scale optimization problems, such as the SoftImpute approach for dealing with missings in principal components analysis, and the MM algorithm for convex clustering.

KEYNOTE TALK 3 \ 大會主講 3

Advances in Spatial Integer-Valued Time Series Modeling for Dengue Fever

Cathy W. S. Chen, *Department of Statistics Feng Chia University,*

Female *Aedes aegypti* mosquitoes, the primary vectors of dengue, typically remain within 400 meters of their emergence site, while human movement facilitates the spread of the dengue virus to new areas, highlighting the importance of analyzing spatial and temporal patterns in dengue cases. This research integrates spatiotemporal dynamics into multivariate integer-valued GARCH models with generalized Poisson or zero-inflated generalized Poisson (ZIGP) distributions. By employing a flexible and continuous conceptualization of distance, the model represents spatial components without relying on a predefined spatial weight matrix. This approach effectively captures the non-separability of space and time, offering a more nuanced analysis of spatiotemporal relationships. In the second part, a spatial hurdle-type model is introduced, alongside the ZIGP model, incorporating two parameters to adjust the spatial weight decay rate and the shape of the decay curve between locations. These models are applied to time-series counts of dengue hemorrhagic fever within a Bayesian framework using Markov chain Monte Carlo (MCMC) algorithms. Performance evaluations through simulations and multivariate weekly dengue case data demonstrate their effectiveness in capturing spatial dependency, over-dispersion, and the high prevalence of zeros, offering a comprehensive framework for modeling the complex characteristics observed in dengue data.

PARALLEL SESSIONS 同步演講場次 FAM-1

10:50 - 12:20 Friday, December 13

FAM-1-1	**Government Session 1 政府場次 1** Organizer: CSAT Committee Chair: 李瓊慧處長	The Forum 國際會議廳
FAM-1-2	**Data Science and Statistical Challenges** Organizer: CSAT Committee Chair: Kengo Kamatani	Socrates 蘇格拉底廳
FAM-1-3	**Statistical Analyses for Scientific Networks and Patent Networks** Organizer: Frederick Kin Hing Phoa Chair: Junji Nakano	Archimedes 阿基米德廳
FAM-1-4	**Recent Developments in Statistical Learning Theory and Applications** Organizer: Xinyuan Song Chair: Ting Li	Michelangelo 米開朗基羅廳
FAM-1-5	**Modeling in Complex Systems** Organizer: Erniel Barrios Chair: Erniel Barrios Discussant: Joseph Ryan G. Lansangan	Raphael 拉斐爾廳
FAM-1-6	**Novel Statistical Models and Methods with Applications** Organizer: IACS-ARS Committee Chair: Li-Hsien Sun	Nietzsche 尼采廳
FAM-1-7	**CSAT Thesis Award Session** **中國統計學社得獎論文場次** Organizer: CSAT Committee Chair: Hsin-Chou Yang	Davinci 達文西廳
FAM-1-8	**Contributed Poster Presentation** **投稿海報場次**	Alexander 亞歷山大廳

FAM-1-1-1

近年我國 ICT 產業發展分析

魏銘佑, *經濟部統計處*

隨全球新興科技的蓬勃發展，產業技術日新月異，資訊科技已廣泛應用於各種領域，從企業到民眾對資訊科技之需求與日俱增，促使資訊與通信科技（Information and Communication Technology，簡稱 ICT）產業成為各國經濟發展的核心產業之一，也是企業提升全球競爭力的關鍵。近年來，在全球數位化轉型及人工智慧技術的推動下， ICT 產業將扮演國家經濟增長的重要驅動力。

本文透過觀察近年來我國 ICT 產業的附加價值、固定資產投資、研究投入及其對經濟成長的貢獻，並分別從製造部門及服務部門等不同範疇分析其發展情況。最後，本文將與主要國家進行比較，觀察各國 ICT 產業附加價值占 GDP 比重、ICT 產品出口概況及 ICT 服務輸出情形等不同面向，進而了解各國 ICT 產業發展狀況及我國 ICT 產業在全球市場中的競爭力，期望能對我國整體產業政策的推展有所助益。

Keywords: 資訊與通信科技、ICT

FAM-1-1-2

住宅租賃統計分析

詹蕙如, *內政部統計處*

依據 109 年人口及住宅普查顯示，全國約有 87.6 萬租賃家戶，因租賃屬民間契約行為，政府雖明定仲介業者須對租賃案件進行登錄，以納管不動產經紀業者，但對於租屋契約內容、承租人居住品質等則無法確實掌握；為改善租屋不透明現況，亟需整合供需數據，提供政府推動住宅政策所需資訊。

因此，內政部統計處盤點租賃住宅政策各業管單位（包含內政部地政司、國土管理署、國家住宅及都市更新中心），將分散之資料進行彙整，建立一個以政府補貼（租金補貼、包租代管）為主，民眾申報為輔（租賃實價登錄）的租屋契約資料集，透過連結內政大數據資料庫（包含戶籍、地籍、房屋稅籍、用水用電、就學就業及社會福利等），產製一完整統計，俾有效掌握租賃住宅市場現況。

本分析針對各縣市租金負擔及補貼效果進行深入探討，並從承租人之性別、年齡等特性呈現受惠族群及推動現況，作為後續政策調整之參考依據。

Keywords: 租金、租賃、包租代管、社會住宅

FAM-1-1-3

營造工程物價指數之趨勢評估

曾昭華, *新北市政府主計處*

營造工程物價指數（以下簡稱營造物價指數）係衡量營造工程之價格變動，而根據「112 年度中央政府總預算案整體評估報告」統計，110 至 111 年受 COVID- 19 影響工程採購案件流標、廢標情形趨增，計畫調增經費金額與幅度偏高之情 形，而其可能原因包含未按實際價格編列預算，而預算之推估常依賴預測能力優劣而定，因此本文進行營造物價指數預測研究。首先根據工程會公布指數調整及補貼門檻–總指數年增率漲跌幅超過 2.5%，建置 COVID-19 情境，再透過 ARIMA 及類神經網路（RNN、LSTM、BiLSTM 及 GRU）模型預測營造物價指數。觀察 MAPE 輔以 RMSE 研究結果，發現 ARIMA 整體表現普遍較類神經網路佳，尤以預測時間區段為一年且受 COVID-19 影響之時間序列較類神經網路更加準確，不過在預測時間區段為每月之時間序列 ARIMA 及類神經網路則不分軒輊；而在類神經網路 RNN、LSTM、BiLSTM 及 GRU 模型預測營造物價指數表現中，發現預測時間區段為每月時 LSTM 普遍較 RNN、BiLSTM 及 GRU 準確之情形，不過預測時間區段為一年時，則以 GRU 普遍較其他類神經網路模型預測能力佳。

Keywords: 營造工程物價指數、自我迴歸移動平均模型（ARIMA）、類神經網路、長短期記憶類神經網路（LSTM）、門控循環單元（GRU）

FAM-1-1-4

運用 AI 隨機森林技術分析高雄市行人交通事故特徵

廖祥凱，*高雄市政府主計處*

行人交通安全為現代化都市重要議題，隨著城市人口及交通運具數量增加，當發生行人交通事故，除對行人本身造成傷亡外，亦可能對其他交通參與者產生負面影響。經本市多年積極強化路口行人交通安全措施，行人傷亡人數已自 108 年逐年下降，但為消除所有行人交通事故，達成「行人交通事故零容忍」之終極目標，本文再深入探討行人交通事故發生模式，以供本市政策強化參用。

本文以交通部 Transport Data eXchange 運輸資料流通服務平臺資料，透過隨機森林技術篩選與本市行人有關交通事故之傷亡者重要特徵，並藉由該等重要特徵進行資料交叉分析。經觀察發現，本市行人相關之交通事故存在特定事故模式，包含高比例發生事故行人有違規穿越車道情形、高齡者易於交叉路口發生事故及無行人空間易導致事故等。期望本文結果能提供本市改善行人安全議題更多思考方向，協助降低反覆發生如「宿命式」的行人事故，保護市民生命，建立安全、宜居之友善城市環境。

Keywords: 行人、隨機森林、決策樹、AI、高雄市、交通事故、高齡、違規

FAM-1-2-1

Doubly Robust Counterfactual Classification

Kwangho Kim, *Department of Statistics, Korea University*

We study counterfactual classification as a new tool for decision-making under hypothetical (contrary to fact) scenarios. We propose a doubly-robust nonparametric estimator for a general counterfactual classifier, where one may incorporate flexible constraints by casting the classification problem as a nonlinear mathematical program involving counterfactuals. We go on to analyze the rates of convergence of the estimator and provide a closed-form expression for its asymptotic distribution. Our analysis shows that the proposed estimator is robust against nuisance model misspecification, and can attain fast root-n rates with tractable inference even when using nonparametric machine learning approaches.

We study the empirical performance of our methods by simulation and apply them for recidivism risk prediction.

FAM-1-2-2

Consistency of Matrix Decomposition Factor Analysis

Yoshikazu Terada, *Graduate School of Engineering Science, Osaka University*

For factor analysis, many estimators, starting with the maximum likelihood estimator, are developed, and the statistical properties of most estimators are well discussed. In the early 2000s, a new estimator based on matrix factorization, called Matrix Decomposition Factor Analysis (MDFA), was developed. Although the estimator is obtained by minimizing the principal component analysis-like loss function, this estimator empirically behaves like other consistent estimators of factor analysis, not principal component analysis. Since the MDFA estimator cannot be formulated as a classical M-estimator, the statistical properties of the MDFA estimator have not yet been discussed. In this talk, we establish the consistency of the MDFA estimator as the factor analysis. That is, we show that the MDFA estimator has the same limit as other consistent estimators of factor analysis.

FAM-1-2-3

On Monitoring and Post-Detection Diagnostics of Correlated Quality Variables of Different Types

Wei-Heng Huang, *Department of Statistics, National Taipei University*

Quality control applications in modern era, especially for non-manufacturing processes, often involve having to monitor correlated variables of different types, continuous, count and categorical. Most of the existing multivariate control charts implicitly assume that the correlated variables to be monitored are of the same type. Another equally challenging task in multivariate quality control which has received relatively little attention is identifying parameters that are actually out of control, when an out-of-control signal is detected on a control chart. In this talk, we will discuss how these two challenges present a unique opportunity to develop multivariate control charts which not only can monitor correlated variables of different types, but also can provide instantaneous diagnostics of out-of-control parameters. The talk will focus on discussing recent works which tackle these challenges by adopting multiple testing procedures in developing multivariate control charts. The proposed chart is shown to outperform the existing charts particularly in its ability to provide more accurate diagnostics. Future research directions along the same line will also be discussed.

Keywords: Diagnostics; Multiple testing procedure; Multivariate control chart; Phase-II monitoring

FAM-1-3-1

A new clustering method to enable exploratory analysis of research results and social issues for research IR

Hiroka Hamada, *The Institute of Statistical Mathematics*
Keisuke Honda, *The Institute of Statistical Mathematics*
Yoshiro Yamamoto, *Tokai University*

A significant challenge for research IR, which analyzes the research activities of institutions and supports decision making, is that there are few indicators for evaluating research.
One of the reasons why this problem has been difficult to solve is that the target to be evaluated in research is unclear. For example, in sports metrics, the smallest unit of evaluation is a player. Similarly, consider the smallest evaluation unit of a study as indicated by research metrics. It became clear that the author and the article did not meet the necessary elements for the player as the smallest unit for evaluating research. To solve this problem, we developed a method to define a unit of evaluation for research activities. We named it PLAYER. The PLAYER methodology combines papers into the smallest unit that can be compared and evaluated without specialized knowledge, with a kind of diversity distance between PLAYERs. However, the research results are not limited to articles. Therefore, the PLAYER methodology has been extended to include the other document (e.g., Japan Government Information) as well as articles in the analysis. First, one of the strengths of the PLAYER method is that, given the feature vectors to be analyzed, the same procedure can be used to define the evaluation units and distances between studies. Therefore, we use embedded representations by several language models for document to obtain feature vectors. Compare the PLAYERs from each result to identify the best method as an extended PLAYER. As an example of using this extended PLAYER methodology, we present the results of an exploratory analysis focusing on the relationship between social issues summarized in government information and research questions proposed by the academic community.

Keywords: Research metrics; Kernel method; Clustering; Optimal Transport; Transdisciplinary research

FAM-1-3-2

Characteristics changes of Scientists in AI-Related Fields for Several Countries Based on Non-Negative Matrix Factorization of Authorship of Scientific Papers

Yuji Mizukami, *Nihon University*
Junji Nakano, *Chuo University*

We extract information on articles and co-authors in AI-related research from the Web of Science (WoS) and compare the trends of research in the field. We analyze 320,063 authors in terms of the number of publications in AI-related papers from 2014 to 2023. We focus on the information of the authors considering the 250 fields of WoS. Our studies use non-negative matrix factorization (NMF), mainly because of the ease of interpretation of the results, from which we can extract the strength and trends of each country' s research in this field.

We must note that in the NMF analysis, there is a problem of having to specify the number of features, which represents the granularity of the model. In this study, the number of features was determined by heuristics.

Keywords: Bibliographic data analysis, Innovation, International comparison

FAM-1-3-3

Analysis of Word Co-occurrence Networks from Paper Abstracts in Semantic Scholar Database

Frederick Kin Hing Phoa, *Institute of Statistical Science, Academia Sinica*
Tiffany Yoonjin Lee, *Sungshin Women's University*
Judy Nakyung Shin, *Sungshin Women's University*
Hohyun Jung, *Sungshin Women's University*

The abstract is a crucial frontmatter element that provides readers with key insights into a manuscript' s core ideas and subject categories. Identifying the most important words in abstracts can offer valuable clues about the central themes and evolving trends within a particular subject area. This work introduces a novel analysis method to determine the importance of words within a subject category over time, based on various centrality measures in a word co-occurrence network. The network is constructed from words extracted from the abstracts of manuscripts within a specific scientific subject.We demonstrate the effectiveness of this method using a subset of the Semantic Scholar database, focusing on the field of Statistics from 2019 to 2023. This is a joint work with Tiffany Yoonjin Lee, Judy Nakyung Shin, and Dr. Hohyun Jung of Sungshin Women' s University, South Korea.

Keywords: Abstract Analysis, Word Co-occurrence Network, Semantic Scholar

FAM-1-3-4

A stochastic model of the citation mechanism of US patent documents

Junji Nakano, *Chuo University*
Yuichiro Yasui, *NIKKEI & ISM*

Citations among articles can be represented by a network structure called a citation network, where nodes and directed edges represent articles with discrete publication time and citations, respectively.
We have proposed a stochastic generative model for scientific articles, in which a citation between two articles is described by a probability based on the type of the citing article, the importance of the cited article, and the difference between their publication times. We consider the out-degree of an article as its type, and the in-degree as its importance. In the model, we assume three functions to approximate structures: a function to express the expected number of articles published in discrete time, a function to represent the aging effect, and a function to represent the out-degree distribution. We also assume two types of generative mechanisms, preferential attachment, and triad formation to perform edge generation.
After analyzing the patent data in detail, we find that some improvements are needed, especially for the triad formation ratio. We propose a model in which the triad formation ratio is a random variable. Triad formation counts follow a binomial distribution, with out-degree and triad formation ratio as parameters. We developed an algorithm to estimate these counts for each node based on citations. The observed distribution showed an inflation of both zeros and ones, and some dependence on out-degree. Considering these facts, we propose a stochastic generative model for patent data, as well as for some scientific article data.

Keywords: Citation analysis, Graph modeling, Stochastic graph simulation

FAM-1-4-1

ReHLine: Regularized Composite ReLU-ReHU Loss Minimization with Linear Computation and Linear Convergence

Ben Dai, *The Chinese University of Hong Kong*
Yixuan Qiu, *Shanghai University of Finance and Economics*

Empirical risk minimization (ERM) is a crucial framework that offers a general approach to handling a broad range of machine learning tasks. In this paper, we propose a novel algorithm, called ReHLine, for minimizing a set of regularized ERMs with convex piecewise linear-quadratic loss functions and optional linear constraints. The proposed algorithm can effectively handle diverse combinations of loss functions, regularizations, and constraints, making it particularly well-suited for complex domain-specific problems. Examples of such problems include FairSVM, elastic net regularized quantile regression, Huber minimization, etc. In addition, ReHLine enjoys a provable linear convergence rate and exhibits a per-iteration computational complexity that scales linearly with the sample size. The algorithm is implemented with both Python and R interfaces, and its performance is benchmarked on various tasks and datasets. Our experimental results demonstrate that ReHLine significantly surpasses generic optimization solvers in terms of computational efficiency on large-scale datasets. Moreover, it also outperforms specialized solvers such as liblinear in SVM, hqreg in Huber minimization and lightning(SAGA, SAG, SDCA, SVRG) in smooth SVM, exhibiting exceptional flexibility and efficiency.

Keywords: Convex optimization, Empirical risk minimization, SVM, Algorithm

FAM-1-4-2

Algorithm design for pure exploration in dueling bandit

Wei You, *HKUST*
Chao Qin, *Stanford University*

Pure exploration in dueling bandits enhances models like ChatGPT by efficiently identifying the most effective response strategies through pairwise comparisons. In this approach, the model explores different response options by "dueling" them against each other, gathering preference information without relying on explicit reward signals. This thorough exploration allows the model to learn which responses are more effective based on feedback or simulated interactions. In the context of pure-exploration problems in dueling bandits, the goal is to accurately identify the best response from multiple alternatives with high confidence, using minimal pairwise comparisons. Focusing on the fixed-confidence setting, we derive a sufficient condition for optimality based on a notion of strong convergence to the optimal allocation of samples. By directly incorporating the dual variables, we characterize the necessary and sufficient conditions for an allocation to be optimal. Our optimality conditions lead to a straightforward yet efficient selection rule, termed "information-directed selection," which adaptively picks candidate pairs based on the information gain of each pair. Numerical experiments highlight the exceptional efficiency of our proposed algorithms relative to existing ones.

Keywords: Pure exploration, dualing bandit, ChatGPT, fixed-confidence

FAM-1-4-3

Towards Non-Asymptotic Convergence for Diffusion-Based Generative Models

Gen Li, *The Chinese University of Hong Kong*

Diffusion models, which convert noise into new data instances by learning to reverse a Markov diffusion process, have become a cornerstone in contemporary generative modeling. While their practical power has now been widely recognized, the theoretical underpinnings remain far from mature. In this work, we develop a suite of non-asymptotic theory towards understanding the data generation process of diffusion models in discrete time, assuming access to ℓ_2-accurate estimates of the (Stein) score functions. We establish a convergence rate proportional to $1/T$ (with T the total number of steps), improving upon past results. Imposing only minimal assumptions on the target data distribution (e.g., no smoothness assumption is imposed), our results characterize how ℓ_2 score estimation errors affect the quality of the data generation process.

Keywords: diffusion models, score-based generative modeling, probability flow ODE

FAM-1-4-4

Heterogeneous Quantile Treatment Effect Estimation with High-Dimensional Confounding

Huichen Zhu, *The Chinese University of Hong Kong*
Zhixin Qiu, *East China Normal University*
Yanling Tang, *East China Normal University*

Understanding heterogeneous treatment responses is essential for advancing precision medicine. This is because individuals often respond differently to the same treatment due to their unique characteristics and circumstances, including patient demographics, genetic predispositions, and environmental exposures. Moreover, inferring causal relationships or associations from observational data can be compromised by the presence of confounding factors, which can sometimes be high-dimensional. In this paper, we focus on estimating heterogeneous quantile treatment effects in the context of high-dimensional confounding. Our innovative approach leverages quantile regression and random forests to capture the variability of treatment effects across both covariates and outcome distributions. Additionally, we employ orthogonal estimation equations to robustly adjust for high-dimensional confounding. We rigorously explore the theoretical properties of our proposed estimator and demonstrate its finite-sample performance through comprehensive simulations. By addressing these complexities, our work aims to enhance the reliability of treatment effect estimates, ultimately contributing to more personalized and effective medical interventions.

Keywords: Quantile Regression, Random Forest , Heterogeneous Treatment Effect, Double Machine Learning

FAM-1-5-1

Block Bootstrap Method for Sparse Spatiotemporal Models

Shirlee Ocampo, *De La Salle University*
Erniel Barrios, *Monash University*

Block bootstrap method and the Cochrane-Orcutt procedure are embedded into the backfitting algorithm to estimate the parameters of a sparse spatio-temporal model. The proposed block bootstrap methods include block residual bootstrap and block parameter bootstrap methods. These were used in Philippine daily COVID-19 data at the provincial level which are sparse, having many gaps and zeroes taking the island groups as the blocks. Results show that the root mean squared error (RMSE) of block residual bootstrap method is slightly lower than the block parameter bootstrap. Comparing block and no block boostrap methods indicated slightly lower RMSE for no block case. The no block residual bootstrap method has slightly lower RMSE compared to the no block parameter bootstrap methods. Moreover, the proposed methods resulted in identifying significant covariates for each block.

Keywords: block bootstrapping; sparse, spatiotemporal; backfitting; Cochrane-Orcutt

FAM-1-5-2

Characterising Brain Connectivity in Terms of Contagion: A Multiple Time Series Perspective

Nazirul Hazim A Khalim, *Monash University Malaysia*
Erniel Bayhon Barrios, *Monash University Malaysa*

In understanding the neurobiological underpinnings of alcoholism, we explore EEG data viewed as interrelated multiple time series parallel to the contagion effects observed in financial markets. This study proposed a new methodological approach to compare control and treatment (alcoholism) groups based on their EEG signals. The concept of volatility clustering is used to identify possible grouping of brain signals from EEG to differentiate the characteristics of the control and the treatment groups, following the test developed by (Barrios and Redondo, 2024). Contagion representing clustered volatility is equated to synchronised patterns of activities across different brain regions manifesting signals triggered by particular stimuli, e.g., alcohol intake.
The method is illustrated using EEG data taken from The UCI KDD Archive (Begleiter, 1999; Zhang et al., 1995) with 10 alcoholic and 10 control subjects, with 64 electrodes sampling at 256 Hz, under conditions involving either a single stimulus or dual stimuli (matched and non-matched), derived from the Snodgrass and Vanderwart picture set. Each subject completed 10 runs per paradigm, and the data were averaged over 10 trials for each stimulus condition. Volatility clustering in multiple EEG time series can separate the two groups significantly.

Keywords: multiple time series, volatility clustering, contagion effect, brain connectivity, EEG

FAM-1-5-3

Customer Utilization of Services by Providers: Insights from HMO Customer Claims Data

Francis Adrian Viernes, *University of the Philippines*
Juedi Myro Eugenio, *University of the Philippines*
Gian Karlo Torreno, *University of the Philippines*

Artificial Intelligence, particularly Machine Learning, has become increasingly important in transforming and evolving critical domains like the healthcare and insurance industries. More than the prediction of illnesses, machine learning can assist in predicting the utilization rates of Health Maintenance Organization' s (HMO) plan members. This benefits not only the insurance companies for upselling purposes but the insured or member, as well. Members who are expected to fully consume their plans before the year-end have the option to increase their coverage.

This paper analyzes the factors influencing members' healthcare usage relative to their benefit limits. The utilization rate, defined as the proportion of the annual limit consumed, was analyzed across several demographic and behavioral variables.

In addition, the paper compares the performance of the different regression algorithms to predict the utilization rate of healthcare services among members. Random Forest, Tobit Regression, and XGBoost were selected and tested for this paper. Among the models tested, the Tobit Regression model performed best, with a Mean Absolute Error (MAE) of 0.0210745. This model was particularly effective in handling the challenges of predicting a rate that could be censored at certain limits. While Random Forest and XGBoost also showed strong results, Tobit Regression provided the most accurate predictions.

Keywords: insurance, machine learning, tobit regression, xgboost, claims

FAM-1-6-1

Information geometry of determinantal point process

Hideitsu Hino, *The Institute of Statistical Mathematics*
Keisuke Yano, *The Institute of Statistical Mathematics*

This work investigates the information geometrical structure of a determinantal point process (DPP). It demonstrates that a DPP is embedded in the exponential family of log-linear models. The extent of deviation from an exponential family is analyzed using the e-embedding curvature tensor, which identifies partially flat parameters of a DPP. On the basis of this embedding structure, the duality related to a marginal kernel and an L-ensemble kernel is discovered.

Keywords: Curved exponential family, Discrete statistical model, Partially ordered set, Statistical curvature

FAM-1-6-2

Alternating Geometric Distribution Family and Its Application to Distribution-Valued Data Clustering

Fumitake Sakaori, *Chuo University*

In singles matches such as table tennis and tennis, rallies are formed by two players alternating strokes. The distribution of the number of strokes in such rallies is typically not well-modelled by a standard geometric distribution because the point-winning probabilities differ between the server and the receiver. Specifically, the server and receiver have different success rates, meaning that the success probabilities vary between even and odd strokes, making the standard geometric distribution inappropriate for modelling such scenarios. To address this, we build upon the probability distribution proposed by Sakaori and Ikebukuro (2021) and introduce the Alternating Geometric Distribution Family (AGD family). The AGD family is designed to capture the alternating nature of the success probabilities, where a random variable represents the server's points when the stroke count is even and the receiver's points when the count is odd. This alternating pattern results in a distribution that fundamentally changes its interpretation based on the parity of the stroke count. Given the unique structure of the AGD family, traditional methods for clustering distribution-valued data, such as those relying on Wasserstein distance or other optimal transport-based metrics, may not be well-suited for these distributions. Standard clustering techniques assume a uniform interpretation of the distribution across all instances, which is not the case with the AGD family. In this study, we thoroughly investigate the distance metrics between AGD family distributions and develop a novel clustering method specifically tailored to handle the alternating nature of these distributions. Our proposed method addresses the limitations of existing approaches and provides a more accurate and meaningful clustering of distribution-valued data within the AGD family, which is critical for applications in sports analytics and other domains where alternating success probabilities play a significant role.

Keywords: Alternating Geometric Distribution, Zero-Distorted Generalized Geometric Distribution, Wasserstein Clustering, Distribution-Valued Data

FAM-1-6-3

Category tree Gaussian process for computer experiments with multi-responses and application to cooling system design

Wei-Ann Lin, *Department of Statistics, National Cheng Kung University*

Computer experiments are commonly used in various engineering fields. These experiments can produce not just one response, but multiple responses. A key goal is to optimize these outcomes at the same time, making multi-objective optimization (MOO) increasingly important. Additionally, in practical situations, both quantitative and qualitative factors are often considered, which leads to a scenario with mixed-type input variables. When there are many different categories resulting from qualitative factors, common modeling approaches can become overly complicated. To address this, we are focusing on fitting multi-task models that can handle mixed inputs. The category tree Gaussian process (ctGP) model, which has been effective for mixed-input surrogate models, will be adapted for multi-objective responses. We demonstrate the effectiveness of this approach through several numerical experiments and a cooling system design problem.

Keywords: Computer experiments, Gaussian process, Quantitative and qualitative factors, Multiobjective responses

FAM-1-6-4

The Participatory Household Survey (PDP) of the Rio Doce Project: statistical matching at the design stage to support impact evaluation

Pedro Luis do Nascimento Silva, *SCIENCE*
Mauricio Teixeira Leite de Vasconcellos, *SCIENCE*

The Participatory Household Survey (PDP) of the Rio Doce Project was carried out in 2022 by SCIENCE, for the Getúlio Vargas Foundation. The PDP had two central objectives:

1. To provide a picture of the socioeconomic situation of the population of the region delimited as affected by the disaster of the collapse of the Fundão tailings dam, in Mariana – MG – in 2015;
2. To provide information that would allow testing the hypothesis of the occurrence of the economic impact of the disaster on the population of the affected region, and estimate the magnitude of this impact, if confirmed.

To meet objective #2, the sample design used an innovative approach of matching the primary sampling units (PSUs) of the sampling frame in the 'affected area' with PSUs in an area unaffected by the disaster but having similar characteristics called 'control area'. The matching was carried out using an algorithm developed to ensure that each PSU in the affected area would be matched with a PSU in the control area having small Mahalanobis distances between them. Matching was carried out without replacement, i.e. no PSU in either of the areas could be matched more than once. Then sampling of PSUs was carried out only in the affected area, with the corresponding control area matched PSUs included in the sample with certainty. This enabled matched PSU analysis after the survey data were collected, increasing efficiency of required impact estimates.

FAM-1-7-1

A Spatio-temporal Hierarchical PGEV Model for Extreme Value Analysis

彭紫涵, *中央大學統計研究所*

PoT-GEV 模型（Olafsdottir et al. 2021）是一種結合廣義極值（generalized extreme value; GEV）分佈和峰值超過閾值（peaks over threshold; PoT）方法的統計模型，近期已被廣泛應用於極端值分析。PoT-GEV 模型用於擬合最大值序列資料，可進一步地評估極端值資料發生的強度和頻率之趨勢。當 PoT-GEV 模型用於分析氣候和環境的資料時，將空間和時間效應納入模型中是不可或缺的。因此，本論文提出一個新穎的時空階層 PoT-GEV 模型，此模型使用潛在高斯隨機過程描述 PoT-GEV 模型的參數用以捕捉資料的空間訊息，同時結合時間相關的協變量用以考慮時間效應。此外，我們採用拉普拉斯近似（Laplace approximation）來取代貝式方法中馬可夫鏈蒙地卡羅（MCMC）的參數估計方法，有效地提高計算效率。我們透過各式的模擬情境來展示時空階層 PoT-GEV 模型的有效性，同時分析台灣的降雨數據和 PM2.5 濃度來說明所提方法的實用性。

Keywords: 貝氏推論；區塊最大序列數據；廣義極值分佈；潛在空間高斯過程；拉普拉斯近似

FAM-1-7-2

通用趨勢更新過程及其在可靠度分析之應用

Generic Trend Renewal Processes with Application to Reliability Analysis

曾柏善，*中央大學統計研究所*

A repairable system can be reused after repairs, but data from such systems often exhibit cyclic patterns. For instance, in the charge-discharge cycles of a battery, capacity decreases with each cycle, and the system's performance may not fully recover after each repair.

To address this issue, the trend renewal process (TRP) transforms periodic data using a trend function to ensure the transformed data satisfy independent and stationary increments.

This study explores random-effects models with a conjugate structure, achieved by reparameterizing the TRP models, called generic TRP (GTRP). These random-effects GTRP models, adaptable to any GTRP model with a renewal distribution possessing a conjugate structure, provide enhanced convenience and flexibility in describing sample heterogeneity.

An approximate formula for the end of performance is derived to make life inferences about recurrent systems, with a simulation study confirming the validity of these inferences for GTRP models. Moreover, the proposed random-effects models are extended to accelerated GTRP (AGTRP) for assessing the reliability of lithium-ion battery data, in addition to analyzing aircraft cooling system data and NASA battery data.

Keywords: Repairable system, trend renewal process, generic trend renewal process, random effects, end of performance (EOP)

FAM-1-7-3

應用於偵測微小偏移的移動平均-廣泛加權移動平均管制圖之研究

A Study on the MA-GWMA Control Chart for Detecting Small Shifts

唐子涵, *中興大學統計學研究所*

在製造業中，製程監控是確保產品品質和生產效率的關鍵步驟。其中管制圖是一種常用的工具，用於監測製程變異和檢測任何可能的製程偏移。隨著製程穩定性的提高，管制圖對於檢測微小偏移的敏感度變得更加重要。

傳統的 Shewhart 管制圖在檢測微小偏移方面可能不夠靈敏，因此 CUSUM、MA、EWMA、GWMA 等管制圖逐漸受到重視，這些方法能夠更及時地偵測到製程中的變化，有助於及早採取補救措施以維持製程穩定性。

本研究提出了一種新型的管制圖，即 MA-GWMA 管制圖，期望能夠提高對微小偏移的製程監控能力。並使用平均連串長度(ARL)、連串長度標準差(SDRL)和中位數連串長度(MRL)當作評估管制圖表現的指標。透過 Monte Carlo 模擬比較各個管制圖在不同偏移程度下的表現。最後，利用實際數據對 MA-GWMA 管制圖進行驗證，進一步評估其在實際應用中的有效性和實用性。

Keywords: 平均連串長度、中位數連串長度、MA 管制圖、GWMA 管制圖、MA-GWMA 管制圖

FAM-1-7-4

加速新生嬰兒腦部分割：利用多模態數據的深度學習方法

Accelerating Neonatal Brain Segmentation: A Deep Learning Approach Using Multimodality Data

馬欣蒂, *成功大學統計學系碩士班*

新生兒腦部形態與孩童的認知、語言發展及行為表現有重要關聯。主流的發展中人類連結組項目（The Developing Human Connectome Project, dHCP）結構管道雖能仔細分割核磁共振成像（Magnetic Resonance Imaging, MRI）腦影像，但運算耗時且易受影像差異影響。現有深度學習方法雖快，但無法多腦區分割，且不適用於跨數據集任務。

本研究提出分層式細節網路（Hierarchical Detail Network, HiDeNet），在跨數據集上提供詳細腦區分割，通過前處理模塊實現在高品質 MRI 影像上訓練，並在品質較差的早產影像上推理。它採用 3D U-Net 骨幹網絡學習特徵，準確分割 87 個腦部結構，同時降低計算複雜度。實驗顯示，HiDeNet 在跨數據集多腦區分割任務中保持 75% 準確率，運行時間僅為 dHCP 方法的 6%，在降低時間的同時維持良好穩定性與泛化能力。

Keywords: 深度學習、新生兒、腦、分割、跨數據任務

FAM-1-7-5

族群可歸因分率的機制分析 - 因果框架下的中介交互作用整合

Mechanism analysis of population attributable fraction - Integrated interaction-mediation analysis under causal inference framework

段宜辰, *陽明交通大學統計學研究所*

族群可歸因分率(Population Attributable Fraction, PAF)是流行病學領域中一個重要的衡量指標，旨在估計危險因子對於特定疾病所造成的可歸因比例，在計算疾病負擔與成本效益分析中佔有極重要的地位。過去全球疾病負擔計畫(Global Burden of Disease, GBD)在計算多個危險因子對於特定疾病所造成的聯合效應時，常假設危險因子之間分配互相獨立；當危險因子之間存在交互作用或者是因果關係的情況下，則使用經驗估計式 $PAF_{1,2,\ldots,K} = 1 - \prod_{k=1}^{K}(1 - PAF_k)$ 來計算多重危險因子所造成的聯合效應，以確保各個危險因子獨立的效應加總不超過 1。然而，此方法使用的正確性與時機不但不明確，其所對應到現實中的因果意義也不清楚；同時，過去也鮮有學者針對危險因子之間所對應的 PAF 指標提出相關的機制分析方法。本研究透過因果推論的反事實模型框架，整合過去學者針對 PAF 指標所提出之交互作用以及中介效應的機制分析，針對兩個危險因子的情況，將總效應拆解為五條不同的路徑，並且提出每一條路徑所對應到現實生活中的因果意義。同時也透過此路徑拆解方法，整合並且延伸過去學者對於 PAF 的機制所提出的拆解法；另外也說明 GBD 經驗估計式恰當的使用時機。此研究成果不但能夠了解危險因子背後相互作用的效應機制，更能夠幫助我們更正確的評估多重危險因子所造成的綜合疾病負擔，提供政策制定者進行更有效益的醫療衛生決策。

Keywords: 族群可歸因分率、交互作用、中介效應、疾病負擔、政策制定

FAM-1-7-6

分段平滑羅吉斯張量迴歸在醫學影像資料分析的方法與應用

Tensor smooth logistic regression with application in neuroimaging data analysis

鄔宜芳, *成功大學統計學系碩士班*

在醫學影像分析的分類問題中，不僅注重分類預測的準確率，也注重的是模型對於反應變數的解釋能力，從而找出與致病因子有關的影像特徵。針對二元反應變數，傳統的羅吉斯迴歸方法通常會將影像像素向量化，並將像素作為解釋變數進行配適。此方法可能破壞了影像的結構資訊，且無法應用於高階的醫學影像分析上。為了將高維度影像應用在處裡二元反應變數的問題上，本研究將提出針對二元反應變數的分段平滑羅吉斯張量迴歸，在模型中利用張量型態的解釋變數來保留影像的空間資訊，同時，考慮到影像具有平滑的特性，在模型中假設影像上座標位置相鄰的體素理論上會有相近的數值，則其所對應的係數也應具有相似的數值。除此之外，在估計參數的過程中，將運用二元反應變數的廣義 Lasso 問題來估計張量參數，本研究將結合迭代重新加權最小平方法(Iteratively Reweighted Least Squares, IRLS) 演算法與交替方向乘子法(Alternating Direction Method of Multipliers, ADMM) 演算法來處理二元反應變數的廣義 Lasso 問題，以實現分段平滑羅吉斯張量迴歸之估計。此外，我們透過模擬研究，全面評估了分段平滑羅吉斯張量迴歸的估計性能，考量了樣本數、訊號強度以及訊號之複雜程度的影響。最終藉由 ABIDE 公開資料集所提供的自閉症患者腦部影像資料進行實例分析，驗證了分段平滑羅吉斯張量迴歸在醫學影像訊號識別方面的有效性。

Keywords: 高維度影像、張量、分段平滑、羅吉斯迴歸、張量迴歸

FAM-1-8-01 Poster Presentation

Using Machine Learning to Identify Predictors of Adolescent Suicidal Ideation

Shou Chun Chiang, *Texas Tech University*
Wan-Chen Chen, *National Chengchi University*

Suicide is the second leading cause of death among adolescents and a major public health concern worldwide. Past research has often relied on predetermined predictors to explain the mechanisms of suicidal ideation and related behaviors. However, few studies have applied machine learning to examine a wide range of potential predictors to identify risk factors for adolescent suicidal ideation. This study aims to identify family-related measures that can serve as the best predictors of adolescent suicidal risk. The sample included 18,498 adolescents who participated in the first wave of the National Longitudinal Study of Adolescent to Adult Health (Add Health), a longitudinal, nationally representative study in the United States. Through a literature review of 167 articles using Add Health data, 73 family variables were identified as predictors of adolescent suicidal ideation. Using a supervised machine learning approach, a balanced random forest model was trained with cross-validation to develop a suicidal ideation risk model. Model performance was evaluated by classification accuracy and the area under the receiver operating characteristic curve (AUC). The results showed that the suicidal ideation model provided a cross-validated AUC of 0.89, with a sensitivity of 86% and a specificity of 78% at an optimized threshold. The strongest risk factors for suicidal ideation were a history of parental mental health problems, family income, adolescent health issues, family cohesion, parent-child relationship quality, academic stress, and neighborhood disadvantages. Overall, this study identified several well-recognized risk factors for suicidal ideation, such as parental mental health and family income, and also revealed new risk factors through the machine learning approach, such as family cohesion and neighborhood disadvantages. In conclusion, the findings advance our understanding of how family characteristics may impact adolescent suicidal risk and help guide future intervention programs and policies to prevent adolescent suicide.a

Keywords: Machine Learning; Adolescent; Suicidal Ideation, Family, Random Forest

FAM-1-8-02 Poster Presentation

Complex non-backtracking matrix for directed graphs and its application to clustering

Keishi Sando, *The Graduate University for Advanced Studies*
Hideitsu Hino, *The Institute of Statistical Mathematics\RIKEN AIP*

Graph representation matrices are essential tools in graph data analysis. Recently, Hermitian adjacency matrices have been proposed to study the structure of directed graphs. Previous studies have demonstrated that these matrices can extract valuable information for clustering. In this paper, we introduce the complex non-backtracking (CNBT) matrix corresponding to a Hermitian adjacency matrix. The proposed matrix shares similar properties with the non-backtracking matrix of undirected graphs. We reveal relationships between the complex non-backtracking matrix and the Hermitian adjacency matrix. Furthermore, we experimentally show that our spectral clustering method based on the CNBT matrix outperforms an existing method, particularly for sparse directed graphs, which are common in real-world data where the number of edges is only a few times the number of vertices.

Keywords: directed graphs, spectral clustering, Hermitian adjacency matrix, nonbacktracking matrix

FAM-1-8-03 Poster Presentation

KOO approach for variable selection of correlations in high-dimension.

Takayuki Yamada, *Kyoto Women's University*
Tetsuro Sakurai, *Suwa University of Science*
Yasunori Fujikoshi, *Hiroshima University*

Statistical hypothesis testing methods for completely uncorrelation have been introduced in Schott (2005, Biometrika) and elsewhere.
We are interested in finding which correlations between variables are nonzero when the null hypothesis of complete uncorrelation is rejected. In this study, for the selection of nonzero correlation coefficients between variables in high-dimensional data, we apply variable selection approach introduced in Nishii et al. (1988, Hiroshima Math Journal) with the generalized information criterion (GIC), and a criterion based on the Fisher z transform of correlation coefficients. This variable selection approach is known as knock-one-out (KOO) method (cf., Bai et al., 2026, Statistica Sinica). We showed the consistency in terms of the selected model coincides with the true model. Small scale simulation was carried out to demonstrate the performance of our method.

Keywords: Variable selection, Correlations, KOO approach, high-dimension

FAM-1-8-04 Poster Presentation

RMST-Based Qualitative Interaction Trees for Survival Data

Yoko Sasayama, *Wakayama Medical University*
Kensuke Tanioka, *Doshisha University*
Ke Wan, *Wakayama Medical University*
Toshio Shimokawa, *Wakayama Medical University*

We propose a new means of applying the traditional QUINT (Qualitative Interaction Trees) method to survival analysis. Identifying subgroups with differential treatment effects is of paramount importance for the advancement of personalized medicine. QUINT (Dusseldorp et al., 2014) is considered to be a robust method for estimating qualitative interactions in randomized controlled trial (RCT) data, but it relies on effect size, such as the mean difference in treatment effects or Cohen's d. The applicability is therefore limited in trials where censored survival data are used as the endpoint.

To overcome this limitation, we propose a novel approach that extends QUINT to survival analysis. Our proposed method introduces an effect size based on the restricted mean survival time (RMST) within the QUINT framework. The RMST represents the average survival time within a specified boundary time τ, calculated as the area under the survival curve up to τ. By incorporating an RMST-based effect size, this method allows for comparisons across a wide range of treatments without relying on the proportional hazards assumption.

We have applied this method to several simulated scenarios and real-world data, confirming its ability to appropriately capture subgroups with qualitative interactions. This extended approach has shown promise in capturing differences in treatment effects within subgroups in trials with survival time as the endpoint, and in effectively detecting qualitative interactions. Our findings suggest that this RMST-based QUINT method could enhance the tools available for personalized medicine in survival analysis contexts.

FAM-1-8-05 Poster Presentation

An effective outlier detection method based on deep generative models by maximizing inlier-memorization effect

Seoyoung Cho, *Department of Statistics, Sungshin Women's University*
Jaesung Hwang, *SK Telecom*
Kwan-Young Bak, *Department of Statistics, Sungshin Women's University*
Dongha Kim, *Department of Statistics, Sungshin Women's University*

Outlier detection (OD) is the task of identifying outliers from a given or upcoming data by learning unique patterns of inliers. Recently, a study introduced a robust unsupervised OD (UOD) approach based on a novel observation of deep generative models, called inlier-memorization (IM) effect, which suggests that generative models memorize inliers before outliers during the early learning stages. The goal of our study is to develop an effective and theoretically well-grounded method for UOD tasks by maximally exploiting the IM effect. We begin by observing that the IM effect is more pronounced when the given training data has fewer outliers. This finding suggests that the IM effect can be enhanced in UOD settings by effectively excluding outliers from mini-batches during the loss function design. To this end, we introduce two main strategies: 1) progressively increasing the mini-batch size during training, and 2) applying an adaptive threshold to compute a truncated loss function. We theoretically demonstrate that these two techniques effectively remove outliers from the truncated loss function, enabling us to fully leverage the IM effect. Coupled with an additional ensemble technique, we propose our method, which we call Adaptive Loss Truncation with Batch Increment (ALTBI). We validate the superiority of our proposed method by analyzing an extensive set of 57 datasets, covering various data types. We prove that ALTBI achieves state-of-the-art performance in identifying outliers compared to other recent methods, even with lower computation costs. We also show that our method maintains strong performance when applied to privacy-preserving algorithms. As a result, our proposed method serves as a powerful OD solution, which is crucial for maintaining data quality and reliability in applications such as fraud detection, network security, and fault diagnosis.

Keywords: Outlier detection; Inlier-Memorization (IM) effect; Unsupervised learning; Deep generative models

FAM-1-8-06 Poster Presentation

Robust and efficient estimation of time-varying treatment effects using marginal structural models dependent on partial treatment history

Nodoka Seya, *Tokyo Medical University*
Masataka Taguri, *Tokyo Medical University*
Takeo Ishii, *Yokohama City University\Yokohama Daiichi Hospital Zenjinkai*

In real-world clinical practice, individuals do not always continue the same treatment but may switch between treatments based on their response to previous treatments. A representative method for estimating the effect of the whole series of treatment history from real-world data with treatment switching is inverse probability weighting (IPW) estimation for marginal structural models (MSMs). However, IPW estimators for MSMs have two difficulties: (i) inefficiency that becomes more pronounced as the number of time points increases; and (ii) bias if explanatory variables in MSMs are not correctly specified or selected. As a variant, IPW estimators for history-restricted MSMs (HRMSMs) have also been proposed, which targets the effect of recent partial treatment history, with how far back in the past partial treatment history goes is specified based on a priori knowledge. IPW estimators for HRMSMs also have two difficulties: (i) inefficiency that becomes more pronounced as the correlation between treatment variables increases; and (ii) a serious difference between the target parameter and the effect of the whole series of treatment history of clinical interest if partial treatment history is not appropriately specified. From this background, the purpose of this study is to propose methods for efficient and low-bias estimation of the effect of the whole series of treatment history. First, we propose closed testing procedures based on a comparison between IPW estimators for MSMs and HRMSMs to select how many time points back to include in treatment variables as explanatory variables in the MSM. Second, we propose an IPW estimator that is generally more efficient and robust to variable misselection than these existing estimators. Third, we propose an estimation framework in which IPW used for the final estimation differs depending on the selected variables. Furthermore, we evaluate the performance of the proposed methods through numerical experiments and real data analysis.

Keywords: time-varying confounding; inverse probability weighting; marginal structural models; history-restricted marginal structural models; variable selection

FAM-1-8-07 Poster Presentation

Graph-linked unified embedding with considering ordinal labels

Hiroshi Kobayashi, *Doshisha University, Japan*
Masaaki Okabe, *Doshisha University, Japan*
Hiroshi Yadohisa, *Doshisha University, Japan*

Graph-Linked Unified Embedding (GLUE) is designed to estimate the low-dimensional space shared across datasets obtained from multiple sources. GLUE leverages the prior knowledge between variables, represented as a knowledge graph. It uses graph neural networks to achieve dimension reduction from multiple data with different feature spaces to a common space. This method is particularly effective for analyzing related datasets, applied to multi-omics data, which integrates various types of omics data, such as gene expression, proteomics, and DNA methylation. These datasets typically consist of samples, such as cells, with variables representing features such as gene expression levels or other cellular characteristics. Therefore, applying GLUE to these data enables the elucidation of diseases at the cellular level. However, GLUE cannot consider label information, even when meaningful ordinal labels are obtained for the samples. To address this issue, we propose an approach that incorporates a penalty term that considers ordinal labels into the GLUE framework. This penalty is designed to preserve the relative order of labels by applying constraints to the embedding process that align with the ordinal nature of the labels. Specifically, the positions of observations within the low-dimensional embeddings generated by GLUE are constrained to correspond to their respective ordinal labels. Consequently, the low-dimensional representations can maintain the ordinal structure present among samples, thereby achieving embeddings consistent with the label information. This method is particularly beneficial in scenarios where data are collected across stages of disease progression or cellular differentiation, where labels reflecting such order are generated. Our proposed method is expected to not only capture the shared structure between datasets, but also respect the ordinal nature of the label information, enabling more insightful and reliable analyses, especially when dealing with complex biological data where understanding of order is important.

Keywords: Dimension reduction, Data integration, Multiple datasets, Multi-omics analysis, Data visualization

FAM-1-8-08 Poster Presentation

Evaluating the Impact of Teaching Methods on Students' Learning Motivation and Outcomes Using Structural Equation Modeling.

YU-HSUAN CHOU, *Department of Statistics and Information Science, College of Management, Fu Jen Catholic University*
JIA-REN, TSAI, *Department of Statistics and Information Science, College of Management, Fu Jen Catholic University*

Game-based learning has recently emerged as an innovative teaching approach. A number of educators and researchers have attempted to integrate this model into their teaching to ignite students' interest in learning. Numerous studies have shown that game-based learning yields better learning outcomes compared to traditional teaching methods, particularly in motivating students, which has led to positive feedback. This study aims to investigate the impact of integrating game-based learning into the curriculum on students' learning motivation and effectiveness. Data collection included administering a 'Learning Motivation Scale' (Pre-test) before the course. After the instructor introduced card games into the curriculum, a 'Learning Motivation Scale' (Post-test) was conducted to assess changes in students' motivation, along with tests to evaluate learning outcomes. Finally, this study offers practical recommendations for future research on integrating games into educational curricula.

Keywords: ARCS Motivation Model, Learning Motivation, Learning Achievement, Game-Based Learning, Structural equation modelling

FAM-1-8-09 Poster Presentation

An efficient causal structure learning algorithm for latent factors

Katsuya Hashimoto, *Graduate School of Human Sciences, Osaka University*
Michio Yamamoto, *Graduate School of Human Sciences, Osaka University\RIKEN AIP\Data Science and AI Innovation Research Promotion Center, Shiga University*

Exploring the causal structure among latent factors, which cannot be directly observed, from observational data is a particularly challenging problem in causal discovery. Recently, methods have been developed to effectively estimate directed acyclic graphs (DAGs) using constrained continuous optimization with continuous acyclicity constraints. However, these methods often require solving the subproblem optimization iteratively, which leads to a significant increase in computational time as the number of parameters, e.g., the number of latent factors, grows. To improve the computational efficiency, we propose a two-step causal discovery algorithm based on the DAG-NoCurl algorithm, assuming the Linear Non-Gaussian Acyclic Models for LAtent Factors (LiNA). Similar to the standard LiNA algorithm, the first step of the proposed algorithm involves locating the latent factors and estimating the factor loading matrix. In the second step, instead of the standard LiNA algorithm, we estimate the causal structure among the latent factors using continuous optimization based on Hodge theory (DAG-NoCurl). Numerical experiments on synthetic data show that the proposed algorithm can estimate the causal structure efficiently compared to existing algorithms. We make a contribution in this work in that we allow an efficient causal structure estimation among latent factors.

Keywords: Causal Discovery, Latent Factors, DAG-NoCurl, LiNA

FAM-1-8-10 Poster Presentation

Comparing predictions among competing risks models with rare events: application to KNOW-CKD study—a multicentre cohort study of chronic kidney disease

Jayoun Kim, *Medical Research Collaborating Center, Seoul National University Hospital, Seoul, Republic of Korea*
Soohyeon Lee, *Department of Statistics, Ewha Womans University, Seoul, Republic of Korea*
Ji Hye Kim, *Department of Internal Medicine, Chungbuk National University Hospital, Cheongju, Korea*
Dha Woon Im, *Department of Internal Medicine, Seoul National University Hospital, Seoul, Republic of Korea*
Donghwan Lee, *Department of Statistics, Ewha Womans University, Seoul, Republic of Korea*
Kook-Hwan Oh, *Department of Internal Medicine, Seoul National University Hospital, Seoul, Republic of Korea*

A prognostic model to determine an association between survival outcomes and clinical risk factors, such as the Cox model, has been developed over the past decades in the medical field. Although the data size containing subjects' information gradually increases, the number of events is often relatively low as medical technology develops. Accordingly, poor discrimination and low predicted ability may occur between low- and high-risk groups. The main goal of this study was to evaluate the predicted probabilities with three existing competing risks models in variation with censoring rates. Three methods were illustrated and compared in a longitudinal study of a nationwide prospective cohort of patients with chronic kidney disease in Korea. The prediction accuracy and discrimination ability of the three methods were compared in terms of the Concordance index (C-index), Integrated Brier Score (IBS), and Calibration slope. In addition, we find that these methods have different performances when the effects are linear or nonlinear under various censoring rates.

FAM-1-8-11 Poster Presentation

Empirical-Likelihood-Based Model Selection Criteria for Missing Longitudinal Data

Yi-Ju Chen, *Tamkang University*
Fu-Yu Ciou, *Tamkang University*
Pai-Ling Li, *Tamkang University*

Missing longitudinal data often arise in public health and biomedical sciences. The weighted generalized estimating equations (WGEE) approach is commonly used to analyze marginal models when data are missing at random. Model selection is a critical aspect of statistical data analysis, and several information criteria based on WGEE estimation have been developed for analyzing missing longitudinal binary data. However, the performance of these criteria, whether based on quasi-likelihood or empirical-likelihood, can vary with sample size, as shown by simulation results. This study proposes a stable model selection criterion based on empirical-likelihood that remains robust across varying sample sizes. The criterion is designed to simultaneously select both the marginal model and the working correlation matrix in the analysis of missing longitudinal binary or ordinal data. The proposed criterion's performance is evaluated through simulation studies and compared to existing criteria. Additionally, its practical utility is demonstrated through two real-world applications featuring missing longitudinal binary and ordinal data.

Keywords: Akaike information criterion, Bayesian information criterion, empirical likelihood, missing longitudinal data, weighted generalized estimating equations

FAM-1-8-12 Poster Presentation

Assessment of Serum Creatinine Prediction in Pregnant Women Based on Blood Pressure with Measurement Errors

BO-SHENG LI, *Fu Jen Catholic University*
JIA-REN TSAI, *Fu Jen Catholic University*

In the health management of pregnant women, blood pressure and serum creatinine are important indicators for predicting and preventing pregnancy-induced hypertension and related complications, which pose potential risks to both the mother and fetus. However, blood pressure measurements are influenced by various factors, including measurement errors, which can obscure the true relationship between these indicators. Therefore, a reliable approach is needed to better understand these associations and improve health monitoring and prevention strategies for pregnant women. This study adopts a process that starts by using bootstrap sampling to randomly select multiple data subsets. For each subset, we calculate the MSE (mean square error) using different estimation methods (ALS, MALS, CS, CSr, MCS, and SIMEX with linear, quadratic, and cubic extrapolation) to see which method performs best. We then rank the MSEs to find the most suitable estimation approach. In practical analysis, we use data from 450 pregnant women (Nab et al., 2021) to explore how systolic blood pressure measured at different time points (30, 60, 90, and 120 minutes), along with age, relates to serum creatinine levels. With this data, we expect to better understand how blood pressure measurement errors might affect the relationship between blood pressure and serum creatinine and to identify the best estimation methods to guide health recommendations.

Keywords: Measurement errors、Bootstrap、Modified adjusted least squares、Modified corrected score、SIMEX

FAM-1-8-13 Poster Presentation

A Study on Predicting Movement Paths in Search and Rescue Operations Using Transformer-Based RL

Junhee Kim, *Department of Statistics, INHA University*
Heedam Kwon, *Department of Statistics, INHA University*

In search and rescue operations, especially those conducted in localized settings, achieving operational success is often challenging, even when substantial resources are deployed relative to the size of the search area. The objective of this study is to develop a predictive model for missing person movement paths using reinforcement learning techniques based on deep neural networks within such complex terrain. Specifically, this research explores reinforcement learning approaches utilizing Vanilla Deep Q-Network (DQN) and Pointer Network DQN. To create a realistic simulation environment, the experiment incorporates detailed geographic data, including high-resolution digital elevation models, river networks, roads, and watersheds. The results demonstrate that while Vanilla DQN, with its simpler structure, can perform effectively, the Pointer Network DQN, which leverages an LSTM-based encoder-decoder architecture and an attention mechanism, achieves more accurate and consistent path predictions in complex terrain conditions.

Keywords: Reinforcement Learning, Deep Neural Networks(DQN), Search and Rescue Operations, PointerNet, PointerDQN, Path Prediction, LSTM, Attention Mechanism

FAM-1-8-14 Poster Presentation

Identification of functional dynamic brain states based on graph attention networks

Inyoung Baek, *Department of Statistics and Data Science, INHA University*
Jong Young Namgung, *Department of Data Science, INHA University*

Investigation of the functional dynamics of the human brain can help to unveil inherent cognitive systems. In this study, we adopted a graph attention network-based anomaly detection technique to identify abrupt changes in functional time series. We used the resting-state functional magnetic resonance imaging data of 1,010 participants from the Human Connectome Project. By applying multivariate time series anomaly detection using the graph attention network approach, we identified three distinct brain states, termed S1, S2, and S3. We further generated low-dimensional representations of functional connectivity (i.e., gradients) for each brain state, and compared these gradients among brain states. S1 and S3 exhibited segregated network patterns, whereas S2 displayed more integrated patterns. A topological analysis based on the betweenness centrality and path length revealed that the integrated state (S2) exhibited efficient network communication. Further, the two segregated states exhibited distinct patterns, with S1 and S3 showing higher centrality in the sensory regions and default mode regions, respectively. When we assessed the transitions between brain states, transitions between the low-level sensory (S1) and higher-order default mode states (S3), as well as between the sensory-focused segregated state (S1) and integrated state (S2), were associated with sensory/motor and memory-related tasks. In contrast, the transitions between the integrated (S2) and segregated states with higher centrality in the default mode region (S3) were found to be related to language tasks. These findings indicate that the proposed approach captures changes in individual participant-level brain dynamics, thereby enabling the assessment of inherently dynamic brain systems.

Keywords: dynamic connectivity; brain state transition; anomaly detection; graph attention network

PARALLEL SESSIONS 同步演講場次 FPM-1

13:30 - 15:00 Friday, December 13

FPM-1-1	**Government Session 2 政府場次 2** Organizer: CSAT Committee Chair: 鄭瑞成處長	The Forum 國際會議廳
FPM-1-2	**Statistical Learning** Organizer: CSAT Committee Chair: Shih-Feng Huang	Socrates 蘇格拉底廳
FPM-1-3	**Statistical Methods and Applications in Clinical Trials** Organizer: Tzy-Chy Lin Chair: Tzy-Chy Lin	Archimedes 阿基米德廳
FPM-1-4	**Fintech and Data Science** Organizer: Ying Chen; Ray-Bing Chen Chair: Shuen-Lin Jeng	Michelangelo 米開朗基羅廳
FPM-1-5	**Recent Advances in Regression-Based Approaches and Modelling Complex Data** Organizer: Hokeun Sun Chair: Hokeun Sun	Raphael 拉斐爾廳
FPM-1-6	**Advances in Reliability Science** Organizer: Chien-Tai Lin Chair: Chien-Tai Lin	Nietzsche 尼采廳
FPM-1-7	**Bayesian Methods and Inference** Organizer: IACS-ARS Committee Chair: Wan-Lun Wang	Davinci 達文西廳
FPM-1-8	**Clustering and Classification** Organizer: IACS-ARS Committee Chair: Shin-Fu Tsai	Alexander 亞歷山大廳

FPM-1-1-1

私立大專校院退場因素探討

程冠瑜, *教育部統計處*

面對少子化浪潮，私立大專校院面臨著多重挑戰。隨著招生人數銳減，學校經費收入也隨之下降，進而導致教職員人數減少等問題。因此，本次研究旨在利用統計模型探討哪些因素導致私立大專校院的退場。

本研究以教育部「大專校院校務資訊公開平臺」與統計處的公布數據作為基礎，將學生類、教職類和財務類等三類變數納入本次研究的數據範疇中。資料時間為 107 學年至 110 學年，透過這四個學年的數據，深入觀察影響學校退場的主要原因。首先利用逐步迴歸與主成份分析來識別影響私立大專校院退出的關鍵因素。逐步迴歸是從所有自變數中找出對應變數具有最大影響力的因素，而主成份分析則能將眾多相關變數轉換成少數個無關聯的主成份，從而更清晰地呈現數據的結構，並將其納入後續的模型建立過程中。同時也運用機器學習方法，包括羅吉斯迴歸模型、LASSO 迴歸模型、決策樹、隨機森林、支援向量機及多實例羅吉斯迴歸模型等，由各統計方法針對測試資料之預測結果，選擇最適合的模型，進一步找出影響退出學校的重要因素。期透過這些方法以不同角度解釋及分析資料，從而更全面地了解私立大專校院退場背後的原因。

Keywords: 私立大專校院退場、機器學習、多實例羅吉斯迴歸

FPM-1-1-2

桃園市詐欺案件被害人特徵

陳叡瑩, *桃園市政府警察局統計室*
黃加朋, *桃園市政府警察局統計室*

隨著時代變遷，詐欺手法也不斷改變，近年來因股市熱絡，出現了以股市名人、財經專家等人名義投放假投資廣告，更有詐騙集團透過社群網站或通訊軟體，結合假交友方式介紹假投資內線消息，造成被害人財產損失。
本文針對近 10 年桃園市詐欺案件被害人分析，透過統計方法，以性別、年齡別及教育程度別找出容易被害之特徵，以了解各類別是否具有顯著差異，並得以針對不同族群挑選相應高頻發生的詐欺手法宣導，以期透過各項指標分析詐欺被害人特性，提供政策之參考依據並營造良好治安環境。

Keywords: 詐欺、被害人

FPM-1-1-3

運用大數據探討臺南市機慢車及行人交通事故城鄉差異

吳維彬, *臺南市政府主計處*

政府一直以來非常重視交通安全的提升，尤其是在快速發展的都市區域，交通事故的頻發成為施政的重要挑戰之一。依據臺南市道安團隊 113 年重點工作，延續重點事故改善推動模式，以機車族群、年輕族群、高齡族群為改善對象，推出「113 年臺南市交通安全提升計畫」，短期以加強執法力度與準度、中長期以改善道路與交通設施改善，並搭配多元管道推動交安教育宣導，期能降低交通事故。在此基礎上，臺南市政府持續運用大數據技術來進行交通安全管理，依據大數據資料庫分析結果，以瞭解交通事故的分佈特徵，進一步提出交通設施及管理相關措施。

本文以城鄉差異的角度，利用臺南市政府警察局 110 年至 112 年間的交通事故原因傷亡統計資料，以及當事者區分統計資料 ，合計超過 27 萬多筆資料做大據分析。輔以內政部 109 年 11 月的電信人流大數據，利用 K-Means 演算法將臺南市 37 個行政區作城鄉劃分，並進行交通事故的比較分析。為瞭解不同區域在交通事故特徵之異同，本文進一步運用羅吉斯迴歸（Logistic regression）分析及隨機森林（Random Forest）演算法，系統性地探討臺南市城鄉交通事故的差異，並提出相應的結論與建議，以期為臺南市政府相關單位政策擬定提供有價值的參考。

Keywords: 臺南市交通安全政策、大數據分析、羅吉斯迴歸分析、機車慢車行人事故風險、城鄉道路事故差異

FPM-1-1-4

農業統計資料庫的視覺化革命

鄭絜文, *農業部統計處*

本文深入探討統計資料的視覺化圖表如何協助查詢與展現農業資訊。
本部視覺化圖表的特色，包括強化資料的互動性與彈性，使用者能根據需求自訂查詢，提升資料使用的便利性；深化階層查詢，提供更細緻的數據分析，促進決策的精確性；平台自力開發與維護，確保靈活應用彈性，以因應突發性輿情回應需求；統計資料之提供與切合施政所需，進一步檢視決策的成效與準確性。
本文將介紹本部開發的農產品視覺化查詢平台 (農產品量價及農情資料查詢平台)，並分享實證案例，展示視覺化圖表在農業政策實務中的應用。透過此次分享，我們期望激發對統計資料視覺化的深入探討，並彰顯農業統計在支援決策的應用價值。

Keywords: 農業統計、農產品量價、農業貿易、資料視覺化

FPM-1-1-5

運用循證決策，提升資源配置效益，增進市民福祉

朱宜寧, *臺北市政府主計處*

近年臺北市政府（簡稱市府）為因應疫後經濟發展、淨零排放、物價上漲、調薪等，致須推動多項新興計畫、調整原有計畫內容及配合增加經費，如：因應衛生福利部針對 0-3 歲發放之育兒津貼（5,000-7,000 元），市府考量臺北市物價水準較高，爰加碼推出協力照顧補助（2,000-3,000 元），故如何在有限的資源下，精進資源配置效益，以增進市民福祉，實為主計處（簡稱本處）籌編年度總預算案之挑戰，亟需研謀精進策略。

市府各機關在編製概算時，係本零基預算精神核實編列，惟對於部分內容複雜且經費需求較高之政策，如托育補助、2-4 歲幼兒就學費用補助等，因未能精確掌握受益對象人數，故以主觀概估提報概算，致概算審查陷入膠著。準此，為突破困局，本處研析國內外推動循證決策的做法，再結合上述本處動機及相關實務作業後，歸納出在地化的循證資料分析過程，包括：問題意識、知識領域、資料來源及分析技術 4 大部分，因其過程需要大量統計專業支撐，且為主計專業的一部分，故本處在籌編 112 年度的第一次追加（減）預算及 113 年度總預算案時，擇選數額較大且以往無法刪減的福利補助項目（如：2 至 4 歲幼兒就學費用補助），主動了解政策內涵及應用臺北市 112 至 141 年人口推估報告及相關統計指標，以統計方法推估年度所需經費，同時參考以前年度執行情形、政策改變緣由及相關數據後，提出具體審查建議意見，供機關與其原提報數作一比較，使其了解減列之具體理由，以提升本處審查意見的說服力。

經以前述方法推估 20 項重要施政項目，並運用該等分析與機關協商結果，已協助各機關於 112 年度總預算第一次追加（減）預算案及 113 年度總預算案分別減編 2.6 億元及 10.2 億元，所減編數額不僅可用於其他優先施政項目，且均不須舉債，亦提升市民福祉。

Keywords: 循證決策、人口推估

FPM-1-2-1

High dimensional linear discriminant analysis using estimation of mean vector and covariance matrix

Junyong Park, *Department of Statistics, Seoul National University*

The classification of high-dimensional data is a critical problem that has been extensively studied over time. Numerous studies have proposed linear classifiers based on Fisher’ s linear discriminant analysis (LDA), which involves estimating the unknown covariance matrix and the mean vector of each group. In this talk, we categorize existing methods into three cases, discussing the shortcomings of each, and comparing various LDA methods that use different estimations of the mean vector and covariance matrix. Specifically, we focus on the empirical Bayes methods for estimating the mean vector in Fisher's LDA, and we compare their performances both theoretically and through numerical studies.

FPM-1-2-2

Controlling Algorithm Trajectory Through Free Energy Landscapes: Overcoming Difficulties in Nonconvex Sparse Penalties

Ayaka Sakata, *The Institute of Statistical Mathematics*

The trajectories of algorithms for inference problems with a large number of parameters are typically defined in high-dimensional spaces, making it challenging to precisely trace. In this presentation, we explore how free energy landscapes-based understanding can provide valuable insights into controlling and optimizing these algorithmic trajectories. Specifically, we introduce an estimation problem using nonconvex sparse penalties, which often induce convergence difficulties. By analyzing the typical behavior of algorithms, particularly using message passing algorithms and based on Bethe free energy, we can identify effective methods for controlling parameters in sparse penalties to achieve globally stable fixed points. Our findings demonstrate that controlling algorithms using the free energy landscapes is an efficient method to improve the performance of algorithms.

FPM-1-2-3

Interpreting Ambiguous Clustered Images Through Heterogeneous Analysis

Yan-Bin Chen, *Master Program in Statistics, CGE/National Taiwan University*

In recent years, significant advancements in data clustering have been achieved through deep neural network. However, the issue of data clustering for ambiguous images remains a persistent challenge. This study identifies heterogeneous data structures using similarity measurements and aims to address this problem by applying multiple regression techniques to ambiguous images. The research provides a detailed analysis of similarity measurements, particularly on the distribution of consistent neighbor counts among image data points, revealing that clean images have a higher fraction of consistent neighbors. Exploring the consistent neighbor distribution has inspired us to improve clustering techniques by adjusting multiple regressions. The findings underscore the importance of statistical methods in managing data heterogeneity to enhance the reliability of image clustering.

Keywords: clustering, similarity measurement

FPM-1-3-1

Mixture of multivariate t linear Mixed Models With Missing Information

Tzy-Chy Lin, *Division of New Drugs, Center for Drug Evaluation*

Linear mixed-effects (LME) models have been widely used for longitudinal data analysis as it 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, normal distribution is sensitive to outliers and intolerance of outliers may greatly affect the model estimation and inference.

In this presentation, 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.

Keywords: Longitudinal data; mixture model; multivariate t distribution; ECME algorithm

FPM-1-3-2

Regression-Based Signaling Pathway Analysis for Promising Inhibitor Compound Selection in Early Phase Clinical Trials

Wei-Quan Fang, *CDE*

Novel therapeutics for high unmet need areas require innovative drug targets. While biologicals expand druggable molecules, appropriate targets remain limited, escalating unmet need. This constraint underpins the importance of new target discovery, allowing transformative therapies where they're mostly needed.

Analyzing signaling pathways in preclinical studies does identify actionable therapeutic targets. Since multi-target drugs, though often more effective than single-target ones, can increase adverse events, highly selective targeting with a best risk-benefit is ideal. Therefore, success of pathway analysis hinging on the elucidation of specific targets may facilitate promising therapeutic interventions.

New regression-based approaches to model signaling pathways were introduced and demonstrated robustness with biological noises, based on metrics of unbiasedness and consistency, as well as simulation results. Analyzing real data, the models aligned with existing literature evidence, thereby confirming their potential. These robust modeling techniques have the capacity to unravel signaling pathways within the intricate realm of genome biology.

Keywords: druggable compound; immunotherapy; signaling pathways; regressions

FPM-1-3-3

The impact of covariate misclassification on robust tests under covariate-adaptive randomization for survival clinical trials

Chun Fan Liu, *Center for Drug Evaluation, Taiwan*

In clinical trials, Covariate-Adaptive Randomization (CAR) is a method for optimizing the randomization design. The concept behind CAR is to balance the number of subjects in the treatment and control groups while effectively controlling the influence of important covariates on the experimental outcomes. CAR ensures a more balanced randomization process, reduces experimental bias, and improves the statistical power of the analysis. However, in practice, misclassification of covariates is inevitable. When the covariates used in CAR are misclassified, not only does it prevent the balance of covariates but also affects the effectiveness of subsequent tests. As CAR is often used in survival analysis, in this study, we utilize the Cox proportional hazards model to derive the impact of misclassified covariates on the asymptotic distribution of robust test statistics under CAR. Additionally, we conducted simulation studies to evaluate the Type I error rates and power curves of different test statistics under two common CAR methods and simple randomization when the covariates are subject to varying degrees of misclassification.

Keywords: Covariate, Adaptive randomization, Misclassification matrix, Cox model

FPM-1-3-4

Sample Size Estimation in Clinical Trials

Chen-Fang Chen, *Center for Drug Evaluation*

In clinical trials, it is important to have a sufficient number of participants (patients) to ensure reliable answers to the key clinical questions. Sample size is usually determined by the primary objective of the trial. Different sample sizes are considered for different research objectives. In long-term clinical trials, the assumptions underlying the original design and sample size calculation would be review at interim. Some adjustments are particularly important if the planning of the trial is based on preliminary or uncertain information. For example, an interim review of the data may reveal that the total number of response variants, event rates, or survival status obtained is not as expected, and the assumptions should be revised and the sample size recalculated as appropriate. This topic is to provide an introduction for the methods of sample size estimation, sample size re-estimation and relative statistical considerations.

Keywords: clinical trial, sample size estimation, sample size re-estimation

FPM-1-4-1

Stock Trend Prediction Assisted by Sentiment Features and Polarity Scores of News Articles

Shuen-Lin Jeng, *Department of Statistics/ National Cheng Kung University*

This study integrated news information into the daily stock price up/down prediction. The predictors included the carefully selected technical features of stock prices, the word level and sentence level sentiment features, and the automatically defined polarity scores of news articles. The main goals of this study are to establish the functional polarity scores of news articles, build models with high prediction accuracy, and identify the important news feature effect for the price up/down prediction. Compared with the Long Short-Term Memory (LSTM) model of Recurrent Neural Networks, we take advantage of Multivariate Adaptive Regression Splines (MARS) as our primary prediction model for its capability in local feature building and its interpretability of selected features. An empirical analysis was conducted on the selected 100 individual stocks in Standard & Poor’ s 500.

FPM-1-4-2

Applying spatial entropy to define forest coverage and applications

Thi Tuan Anh Tran, *University of Economics Ho Chi Minh City*

Assessing the forest coverage is very important for conserving the environmental sustainability and understanding ecological system. This study introduces a new approach using spatial entropy to analyze the distribution and density of forest coverage across varied landscapes. Spatial entropy is applied to preprocess satellite images to classify land cover into forested and non-forested areas. A higher entropy value indicates a more uniform distribution of forested areas, suggesting effective conservation practices, while lower values signify concentrated patches of forests, potentially indicating ecological or anthropogenic pressures.

By providing a robust quantitative measure, this approach leads to implications for policymakers and environmental activist in identifying priority areas for conservation, monitoring changes in forest density over time.

Keywords: Spatial entropy; forest coverage; environmental sustainability, satellite imagery

FPM-1-4-3

Enhanced Customer Service System for Thailand Online Shops Using Case-Based Reasoning

Thacha Lawanna, *International College of Digital Innovation, Chiang Mai University*

Integrating Case-Based Reasoning into the Thai online shops customer service is a great business strategy. The process involves five key phases, beginning with the CBR system working alongside domain experts to extract essential information and develop clear rules, forming the groundwork for precise problem-solving. Next, in the case retrieval and similarity evaluation phase, the system applies algorithms to find past cases that are similar to the current issue. In the adaptation stage, the system adjusts the solutions from those previous cases to fit the unique aspects of the current problem. Next is the solution evaluation and case counting phase which ensures the accuracy of the solutions and track metrics for performance assessment. The final phase focuses on system optimization and continuous improvement where case outcomes are analysed, the knowledge base is updated, and retrieval and adaptation algorithms are adjusted to facilitate on going system enhancement. This iterative learning process allows the CBR system to improve over time, increasing its ability to handle diverse cases while enhancing the overall online shopping experience. CBR can achieve accuracy, precision, and recall rates between 99.93% and 99.98%, approximately exceeding Machine Learning by 9%, Knowledge-based approach by 15%, and Rule-based approach by 20%, which highlights its outstanding performance.

FPM-1-4-4

A Modified VAR-deGARCH Model for Asynchronous Multivariate Financial Time Series via Variational Bayesian Inference

Wei-Ting Lai, *Graduate Institute of Statistics, National Central University*
Ray-Bing Chen, *Department of Statistics, National Cheng-Kung University*
Shih-Feng Huang, *Graduate Institute of Statistics, National Central University*

This study proposes a modified VAR-deGARCH model, denoted by M-VAR-deGARCH, for modeling asynchronous multivariate financial time series with GARCH effects and simultaneously accommodating the latest market information. A variational Bayesian (VB) procedure is developed to infer the M-VAR-deGARCH model for structure selection and parameter estimation. We conduct extensive simulations and empirical studies to evaluate the fitting and forecasting performances of the M-VAR-deGARCH model. The simulation results reveal that the proposed VB procedure produces satisfactory selection performances. In addition, our empirical studies find that the latest market information in Asia can provide helpful information to predict market trends in Europe and South Africa, especially when momentous events occur.

FPM-1-5-1

Integrative sparse reduced-rank regression via orthogonal rotation for analysis of high-dimensional multi-source data

Kipoong Kim, *Department of Statistics, Changwon National University*
Sungkyu Jung, *Department of Statistics, Seoul National University\Institute for Data Innovation in Science, Seoul National University*

Recent advancements in technology have allowed for the collection of large amounts of data from multiple sources for individuals, leading to an increase in interest in developing statistical methods for analyzing the multi-source data. One of the main interests is to identify the structural association of multiple sources on multiple correlated responses, including individual, joint, and partially-joint structures. In this work, we propose a novel integrative sparse reduced-rank regression (iSRRR) model for identifying the structural associations between multi-source data and multiple responses. The model is based on the assumption of a structured decomposition of the coefficient matrix, and utilizes a new constraint based on orthogonal rotation to ensure model identifiability. The constraint imposes a specific structure, quartimax-simple, on the loading matrix, which enhances interpretability when identifying the multi-source structures relevant to specific responses. An iterative algorithm for estimating the iSRRR model parameters is also proposed. Simulation studies have demonstrated the ability of the proposed method to identify the underlying structured associations between multi-source data and multiple responses. The method has been applied to multi-omics dataset with multiple drug responses, and has been shown to be capable of detecting structured association patterns.

Keywords: Multi-source data, Integrative analysis, Orthogonal rotation, Reduced-rank regression, Structural learning

FPM-1-5-2

Variational Inference Aided Variable Selection For Spatially Structured High Dimensional Covariates

Minwoo Kim, *Pusan National University*
Siddhartha Nandy, *Case Western Reserve University*
Shrijita Bhattacharya, *Michigan State University*
Tapabrata Maiti, *Michigan State University*

We consider the problem of Bayesian high dimensional variable selection under linear regression when a spatial structure exists among the covariates. We use an Ising prior to model the structural connectivity of the covariates with an undirected graph and the connectivity strength with Ising distribution parameters. Ising models which originated in statistical physics, are widely used in computer vision and spatial data modeling. Although a Gibbs solution to this problem exists, the solution involves the computation of determinants and inverses of high dimensional matrices rendering it unscalable to higher dimensions. Further, the lack of theoretical support limits this important tool's use for the broader community. This paper proposes a variational inference-aided Gibbs approach that enjoys the same variable recovery power as the standard Gibbs solution all the while being computationally scalable to higher dimensions. We establish strong selection consistency of our proposed approach together with its competitive numerical performance under varying simulation scenarios.

Keywords: Bayesian Variable Selection, Structured Covariates, Variational Bayes, Highdimensional, Ising distribution

FPM-1-5-3

GBOSE: Generalized Bandit Orthogonalized Semiparametric Estimation

Gi-Soo Kim, *UNIST*
Mubarrat Chowdhury, *UNIST*
Elkhan Ismayilzada, *UNIST*
Khalequzzaman Sayem, *UNIST*

In sequential decision-making scenarios i.e., mobile health recommendation systems revenue management contextual multi-armed bandit algorithms have garnered attention for their performance. But most of the existing algorithms are built on the assumption of a strictly parametric reward model mostly linear in nature. In this work we propose a new algorithm with a semi-parametric reward model with state-of-the-art complexity of upper bound on regret amongst existing semi-parametric algorithms. Our work expands the scope of another representative algorithm of state-of-the-art complexity with a similar reward model by proposing an algorithm built upon the same action filtering procedures but provides explicit action selection distribution for scenarios involving more than two arms at a particular time step while requiring fewer computations. We derive the said complexity of the upper bound on regret and present simulation results that affirm our methods superiority out of all prevalent semi-parametric bandit algorithms for cases involving over two arms.

FPM-1-5-4

Nonparametric Bayesian Poisson hurdle random effects model with application to environmental epidemiology

Jinsu Park, *Chungbuk National University*
Yeonseung Chung, *Korea Advanced Institute of Science and Technology*

In environmental epidemiology, the short-term association between temperature and suicide has been examined by analyzing daily time-series data on suicide and temperature collected from multiple locations. A two-stage meta-analytic approach has been conventionally used. A Poisson regression with splines is fitted for each location in the first stage, and location-specific association parameter estimates are pooled, adjusted, and regressed onto location-specific variables using meta-regressions in the second stage. However, several limitations of the conventional two-stage approaches have been reported. First, the Poisson distribution assumption may be inappropriate because the daily number of suicides is often zero. Second, the normal assumption in the second-stage meta-regression is not sufficiently flexible to describe between-location heterogeneity when subgroups exist. Third, the two-stage approach does not properly account for the statistical uncertainty associated with first-stage estimates. In this study, we propose a nonparametric Bayesian Poisson hurdle random effects model to investigate heterogeneity in the temperature-suicide association across multiple locations. The proposed model consists of two parts, binary and positive, with random coefficients specified to describe heterogeneity. Furthermore, random coefficients combined with location-specific indicators were assumed to follow a Dirichlet process mixture of normals to identify the subgroups. Our methodology provides a general approach for the nonparametric modeling of zero-inflated data. The proposed methodology was validated through a simulation study and applied to data from a nationwide temperature-suicide association study in Japan.

Keywords: Bayesian inference, temperature-suicide association, Poisson hurdle model, Drichlet process mixture, model-based clustering

FPM-1-6-1

The First-Passage-Time Moments for Hougaard Process and its Birnbaum-Saunders Approximation

YiShian Dong, *Department of Statistics, National Chengchi University*
Chien-Yu Peng, *Institute of Statistical Science, Academia Sinica*
Tsai-Hung Fan, *Graduate Institute of Statistics, National Central University*

Hougaard processes, which include gamma and inverse Gaussian processes as special cases, as well as the moments of the corresponding first-passage-time (FPT) distributions, are commonly used in many applications. Because the density function of a Hougaard process involves an intractable infinite series, the Birnbaum-Saunders (BS) distribution is often used to approximate its FPT distribution. This article derives the finite moments of FPT distributions based on Hougaard processes and provides a theoretical justification for BS approximation in terms of convergence rates. Further, we show that the first moment of the FPT distribution for a Hougaard process approximated by the BS distribution is larger and provide a sharp upper bound for the difference using an exponential integral. The conditions for convergence coincidentally elucidate the classical convergence results of Hougaard distributions. Some numerical examples are proposed to support the validity and precision of the theoretical results.

Keywords: characteristic function; contour integration; exponential dispersion model; residue; Stirling numbers

FPM-1-6-2

Optimal Measurement Interval Planning for Gamma Degradation Tests

Hung-Ping Tung, *National Yang Ming Chiao Tung University*
Yu-Wen Chen, *National Yang Ming Chiao Tung University*

Gamma degradation tests are widely used in industry to evaluate product lifetimes. To design an efficient test plan under budget constraints, several studies have explored optimal design strategies, including considerations for the number of test units, measurement times, number of measurements, and testing termination time. However, most existing literature focuses on designs with equal measurement interval. In this paper, we present an optimal measurement interval and demonstrate that the equal measurement interval is the least effective. Furthermore, we propose optimal designs for both measurement interval and provide an example that shows the superiority of the optimal design with the optimal measurement interval compared to the equal measurement interval.

Keywords: Degradation test, Gamma process, Measurement time, Optimal design

FPM-1-6-3

Degradation Analysis of Accelerated Multivariate Inverse Gaussian Process with Random Effects

Yi-Fu Wang, *National Cheng Kung University*
Tzu-Erh Huang, *National Cheng Kung University*

High-tech products have become an integral part of human life. To understand information such as the warranty period and lifetime of these products, reliability analysis is crucial for manufacturers. For high-reliability products, collecting sufficient failure data within a limited time is challenging. Therefore, lifetime estimation is performed based on quality characteristics, known as degradation analysis. Additionally, accelerated degradation tests use controlled stress factors to complete experiments in a shorter time. However, with the diversification of products and advancements in measurement technologies, analyzing a single quality characteristic for degradation is no longer sufficient, making the analysis of multiple quality characteristics necessary. In this study, we first assume that the degradation paths of each quality characteristic follow an inverse Gaussian process. We then construct a common dependent random effect across the various quality characteristics. This approach enables us to capture the heterogeneity among samples and the correlations between quality characteristics, which we refer to as the Multivariate Inverse Gaussian Process with Random Effects (MIGP) model. Furthermore, considering accelerated degradation tests, we propose the Accelerated Multivariate Inverse Gaussian Process with Random Effects (AMIGP) model. Finally, the feasibility of these models is validated through simulation studies and two case studies.

Keywords: accelerated degradation test, multiple quality characteristics, inverse Gaussian process, random effects

FPM-1-6-4

Reliability inference for devices with dependent components under an Inverted Dirichlet distribution

Man Ho Ling, *The Education University of Hong Kong*

Modern systems are often composed of multiple interdependent components, making multivariate distributions critical for reliability analysis. In this work, the inverted Dirichlet distribution is considered to model dependent component lifetimes. This distribution offers intuitive interpretability and considerable flexibility for modeling multivariate lifetime data. Despite these advantages, the inverted Dirichlet model has seen limited use in the presence of censoring, due to the lack of explicit forms for the associated reliability function. A frailty-based approach is presented to efficiently perform maximum likelihood inference for one-shot device test data with dependent components under constant-stress accelerated life tests to address this challenge. The maximum likelihood estimates and asymptotic confidence intervals are derived for the reliability analysis. An extensive Monte Carlo simulation study is conducted to evaluate the performance of the proposed inferential methods.

Keywords: Frailty model, one-shot devices, censored, inverted Dirichlet distribution

FPM-1-7-1

Loss-based Bayesian Sequential Prediction of Value at Risk with a Long-Memory and Non-linear Realized Volatility Model

Rangika Iroshani Peiris, *University of Sydney Business School*
Minh-Ngoc Tran, *University of Sydney Business School*
Chao Wang, *University of Sydney Business School*
Richard Gerlach, *University of Sydney Business School*

A long-memory and non-linear realized volatility model class is proposed for direct Value-at-Risk (VaR) forecasting. This model, referred to as RNN-HAR, extends the heterogeneous autoregressive (HAR) model, a framework known for efficiently capturing long memory in realized measures, by integrating a Recurrent Neural Network (RNN) to handle non-linear dynamics. Loss-based generalized Bayesian inference with Sequential Monte Carlo is employed for model estimation and sequential prediction in RNN-HAR. The empirical analysis is conducted using daily closing prices and realized measures from 2000 to 2022 across 31 market indices. The proposed model' s one-stepahead VaR forecasting performance is compared against a basic HAR model and its extensions. The results demonstrate that the proposed RNN-HAR model consistently outperforms all other models considered in the study.

Keywords: HAR model, Recurrent Neural Network, Quantile Score, Sequential Monte Carlo, Generalized Bayesian inference

FPM-1-7-2

Fast Laplace Variational Inference Method for Bayesian Envelope Model

Seunghyeon KIM, *Department of Mathematics and Statistics, Chonnam National University, Gwangju, South Korea*
Kwangmin Lee, *Department of Big Data Convergence, Chonnam National University, Gwangju, South Korea*
Yeonhee Park, *Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, Madison, WI 53726, United States*

The envelope model is an estimation-efficient method for estimating regression coefficients within the context of a multivariate linear regression model. It splits material and immaterial parts of responses(or predictors) using an orthogonal basis, thereby enhancing efficiency by removing the immaterial variation. The envelope model has been extended to the Bayesian approach, but has the limitation of slow estimation of parameter by Metropolis-Hastings (MH) method. Variational inference (VI) is well known as a fast alternative method, but previous studies applying automatic differentiation VI (ADVI) have reported that the accuracy is very low. In this study, we discuss and solve common computational issues that arise when directly applying VI to Bayesian envelope models. We propose a Laplace VI method for the comprehensive Bayesian envelope model that enables fast estimation. Applying Laplace VI requires the second derivative of the log posterior distribution with respect to the parameter. We suggest a parametric formulation where it can be easily derived. Simulation studies show that the proposed method is almost 20 times faster than the MH method. Also performance of proposed method is similar with MH method, more accurate than ADVI.

Keywords: Variational inference; envelope model; Laplace approximation

FPM-1-7-3

Exploring Bayesian Methodologies in Random Forest Classification for Predicting Clinical Trial Outcomes

Paula Angelica Tagumasi Ramos, *Pfizer\University of the Philippines Diliman*
Maria Kudela, *Pfizer*

Development of a new compound and assessment of its success requires rigorous approach that integrates historical data from clinical trials and published literature. Leveraging machine learning techniques combined with thorough understanding of the patient population and careful feature selection can synergistically drive sound decision-making and yield robust predictions of clinical trial outcomes. Conventional machine learning methods often face challenges such as low predictive accuracy, overfitting, and lack of reliable uncertainty estimates. We therefore propose the use of Bayesian priors (e.g. split weights to improve feature selection) and a causal inference framework to address these limitations. Positive Predictive Value (PPV) and Negative Predictive Value (NPV) were used as metrics in assessing the performance of a Bayesian-augmented random forest algorithm against the regular random forest to predict binary outcome (Clinical Remission) in a simulated Ulcerative Colitis trial containing only baseline features. The test dataset was severely imbalanced (N=148), consisting of 38 remission and 110 non-remission patients. The performance of the proposed algorithm was also analysed using other simulated clinical trial data. Additionally, conformal prediction was explored to generate statistically rigorous uncertainty intervals for these models, further enhancing their predictive reliability. Bayesian-augmented random forest (BRF) algorithm yielded better performance [PPV: 60.52%; NPV: 79.45%] compared to the regular random forest (RF) algorithm [PPV: 45.45%; NPV: 75.91%]. The incorporation of Bayesian weights within the random forest framework demonstrated clear advantages in predictive performance, particularly in the context of imbalanced datasets common in clinical trial data. Additionally, the integration of conformal prediction for generating uncertainty intervals has further solidified the reliability of the predictions, offering a more robust and scientifically sound approach to model uncertainty. This study highlights the potential of Bayesian-augmented ML in improving decision-making processes in drug development and clinical trial outcome prediction, leading to more precise and reliable models in future applications.

Keywords: Variational Bayes; Random Forest; Machine Learning; Conformal Prediction; Remission

FPM-1-7-4

Bayesian Stochastic Frontier Models under the Skew Normal Settings

Boris Choy, *The University of Sydney*
Zheng Wei, *Texas A&M University*
Tonghui Wang, *New Mexico State University*
Xiaonan Zhu, *University of North Alabama*

Recently, a skew normal-based stochastic frontier model has emerged as a promising tool for efficiency analysis. In this paper, a Bayesian framework for statistical inference is presented, incorporating both informative and non-informative prior knowledge. The efficacy of the Bayesian approach is evaluated through rigorous examination using both simulation data and real data from a manufacturing productivity study. Comparisons with the conventional maximum likelihood approach are conducted. Results from both simulated and empirical investigations unequivocally demonstrate the superior performance of the Bayesian methodology.

Keywords: Bayesian inferences, Efficiency , Markov chain Monte Carlo, Skew normal distribution, Stochastic frontier model

FPM-1-8-1

MicroFisher: Fungal taxonomic classification for metatranscriptomic and metagenomic data using multiple short hypervariable markers

Steven Hung-Hsi Wu, *Departement of Agronomy, National Taiwan University, Taipei, Taiwan*
Haihua Wang, *North Florida Research and Education Center, University of Florida, FL, USA*
Kaile Zhang, *North Florida Research and Education Center, University of Florida, FL, USA*
Ko-Hsuan Chen, *Biodiversity Research Center, Academia Sinica, Taipei, Taiwan*
Rytas Vilgalys, *Department of Biology, Duke University, NC, USA*
Hui-Ling Liao, *North Florida Research and Education Center, University of Florida, FL, USA*

Profiling the taxonomic and functional composition of microbes using metagenomic (MG) and metatranscriptomic (MT) sequencing is advancing our understanding of microbial functions. However, the sensitivity and accuracy of microbial classification using genome- or core protein-based approaches, especially the classification of eukaryotic organisms, is limited by the availability of genomes and the resolution of sequence databases. To address this, we propose the MicroFisher, a novel approach that applies multiple hypervariable marker genes to profile fungal communities from MGs and MTs. This approach utilizes the hypervariable regions of marker genes for fungal identification with high sensitivity and resolution. Simultaneously, we propose a computational pipeline (MicroFisher) to optimize and integrate the results from classifications using multiple hypervariable markers. We applied MicroFisher to the synthetic community profiling to test the performance of our method, and found high performance in fungal prediction and abundance estimation. In addition, we also used MGs from forest soil and MTs of root eukaryotic microbes to test our method and the results showed that MicroFisher provided more accurate profiling of environmental microbiomes compared to other classification tools. Overall, MicroFisher serves as a novel pipeline for classification of fungal communities from MGs and MTs.

Keywords: fungal communities, hypervariable marker genes, metagenomic, pipeline

FPM-1-8-2

Clustering for high-dimensional categorical data

somi cha, *Department of Mathematics and Statistics, Chonnam National University of Korea*

Kwangmin Lee, *Department of Mathematics and Statistics, Chonnam National University of Korea*

Clustering high-dimensional categorical data is a challenging problem. Applying traditional cluster methods, such as k-means, to high-dimensional data is ineffective. In this paper, we propose a 3-step procedure for clustering high-dimensional categorical data. The main idea is identifying variables irrelevant to clustering and applying a clustering algorithm to the reduced data. The proposed method first applies spectral clustering to achieve an initial clustering result. With this initial clustering result, we select informative variables with large values of Cramér's V to exclude non-informative variables. Finally, we apply a multiple correspondence analysis (MCA) based method on the reduced dimensional datasets with selected features for the final clustering. We demonstrate that the proposed model-based approach shows better performance in synthetic datasets.

Keywords: High-dimensional clustering, Categorical data clustering, Variable selection, Cramér's V, Spectral clustering

FPM-1-8-3

Enhancing Digital Health Assessments through Classification of Text and Acoustic Features in High-Dimensional Mixed-Type Data

Ngai Lam Chan, *Department of Information Systems, Business Statistics and Operations Management, The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong, China*
Amanda M. Y. Chu, *Department of Social Sciences and Policy Studies, The Education University of Hong Kong, Tai Po, Hong Kong, China*
Benson S. Y. Lam, *Department of Mathematics, Statistics and Insurance, The Hang Seng University of Hong Kong, Shatin, Hong Kong, China*
Agnes Tiwari, *School of Nursing, The University of Hong Kong, Pokfulam, Hong Kong, China*|*School of Nursing, Hong Kong Sanatorium & Hospital, Happy Valley, Hong Kong, China*
Helina Yuk, *Department of Social Work, The Chinese University of Hong Kong, Shatin, Hong Kong, China*
Mike K. P. So, *Department of Information Systems, Business Statistics and Operations Management, The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong, China*

We developed an automatic speech analytics program (ASAP) to detect psychosocial health issues from interviews. which were conducted and recorded with 100 Cantonese-speaking family caregivers recruited for the study. We analyzed speech using text and acoustic features. Text features convey the speech content, and the transcript was extracted using Google Cloud Speech API from audio tracks. As the text features contain lots of irrelevant information, we employed cross validation (CV) method to identify relevant text features for the given psychosocial instrument. For the acoustic features, it contains emotional information about the caregivers. We applied popular signal processing techniques including Fourier Transform and spectral methods to extract useful information for further processing. After obtaining the textual and acoustic features, we combined the two sets of features using principal component analysis (PCA) to address the redundant information often carried by text and acoustic features from the same person, which can cause bias in the analysis. We

can remove redundancy and bias by merging highly correlated features into a set of uncorrelated features using PCA. Finally, we adopted linear support vector machine (LSVM) to perform classification. We applied the proposed method to classify three different psychosocial instruments using the two sets of features. Caregiver burden, depressive symptoms, and family resilience were assessed using the CBI (24 items), BDI-II (21 items), and FRAS (54 items), respectively. Scores were totaled and categorized as High or Low based on set thresholds as ground truth.

The correct classification rates of 10-Fold CV of these three instruments are 87%, 80% and 91% respectively. The findings show that digital tools like ASAP leverage acoustic and text features from interviews to enhance psychosocial assessments by enabling early detection of high risk of psychosocial health issues, allowing for quicker referrals and bypassing more intensive conventional assessments in the social health sector.

Keywords: High-Dimensional Data, Mixed-Type Data, Text Analysis, Acoustic Analysis, Dimensionality Reduction

FPM-1-8-4

Biplots for visualising multidimensional data

Sugnet Lubbe, *Department of Statistics and Actuarial Science, Stellenbosch University*

In a multidimensional world it is important to visualise our data in two or three dimensions. Especially in a classification setting a picture is worth a thousand words Biplots, as the prefix ‘bi-’ suggests, represent the cases as well as the variables. Such a visual display allows for intuitive interpretation of relationships between samples, between variables and between samples and variables.

After a brief introduction, the focus will be on observations grouped into different classes. In this case the visual display of the samples and variables allows interpretation on which variables separate classes and to what extent classes overlap.

We will step through an array of options related to classification settings. In traditional multivariate statistics courses, classification is introduced through linear discriminant analysis. We will start from this point onward and look at biplot options for linear and nonlinear classification, continuous and categorical data, and make our way to more recently developed classification methodology.

PARALLEL SESSIONS 同步演講場次 FPM-2

15:20 - 16:30 Friday, December 13

FPM-2-1	**Machine Learning Algorithms in Analyzing Survey Data** Organizer: Erniel Barrios Chair: Erniel Barrios Discussant: Joseph Ryan G. Lansangan	The Forum 國際會議廳
FPM-2-2	**Advanced Statistical Techniques for Multifaceted Data** Organizer: Tsung-I Lin Chair: Tsung-I Lin	Socrates 蘇格拉底廳
FPM-2-3	**New Advances in Model Selection and Inference** Organizer: Hsueh-Han Huang Chair: Hsueh-Han Huang	Archimedes 阿基米德廳
FPM-2-4	**Structural Equation Modeling** Organizer: IACS-ARS Committee Chair: Ming-Che Hu	Michelangelo 米開朗基羅廳
FPM-2-5	**Statistical Analysis of Complex Data** Organizer: Hsin-wen Chang Chair: Ming-Yueh Huang	Raphael 拉斐爾廳
FPM-2-6	**Statistical Machine Learning and Inference** Organizer: IACS-ARS Committee Chair: Han-Ming Wu	Nietzsche 尼采廳
FPM-2-7	**Innovative Designs and Analysis Methods for Clinical Trials and Biomedical Data** Organizer: IACS-ARS Committee Chair: Steven Hung-Hsi Wu	Davinci 達文西廳
FPM-2-8	**Recent Advances on Interplay of Statistics and Optimization** Organizer: IACS-ARS Committee Chair: Yan-Bin Chen	Alexander 亞歷山大廳

FPM-2-1-1

How Filipinos Value Democracy, Ethics, and Governance: Insights for Social Development Advocacy and Planning

John Erwin Bañez, *University of the Philippines*
Erniel Barrios, *Monash University Malaysia*
Karl Arvin Hapal, *University of the Philippines*

The outcomes of the 2022 Philippine elections triggered various questions regarding the way Filipinos value democracy and ethics. What drives the formation of Filipino ideation of democracy and ethics? Using the Philippine segment of the 2019 World Values Survey, we used machine learning algorithms to provide indicative responses to some of these issues. In the initial analysis, exploratory factor analysis (EFA) was used to identify latent factors related to Ethics and Norms (ENVS), Governance (GFS) and Democracy (DVFS). Using Factorial ANOVA, these factors were then analyzed to identify drivers of values formation. Results show, Filipinos generally adhere to traditional values, authoritarian rule, and general satisfaction with governance. Values ideation among Filipinos is significantly influenced by demographic characteristics of households. There is also a significant geographic clustering among these latent factors of values. The results deepen our understanding of Southeast Asia' s longest democracy and can serve as a baseline in designing a more targeted and nuanced social development advocacy and plans aimed at attaining more liberal values - i.e. believing in equality, individual liberty, and related values. For example, communication campaigns, education intervention, and voter education programs can be guided by these results.

Keywords: Ethics and Norms, Democracy, Exploratory Factor Analysis, Factorial ANOVA

FPM-2-1-2

Testing Congruence of Scale Scores Using Principal Components Analysis

Martin Augustine Borlongan, *University of the Philippines School of Statistics*
Geselle Manguiat, *De La Salle University Manila*
Erniel Barrios, *Monash University Malaysia*

We propose a method of testing the congruence of individual scores from a scale of items using principal components analysis. Prediction intervals of PC scores are constructed via bootstrap sampling to establish the plausible scores of an individual based on their "baseline" response to the scale. These are then used to assess whether a new set of response is still congruent with the "baseline" or whether a change in the individual has occurred. Simulation runs are done to compare the proposed method with the usual test of means (parametric and non-parametric). The performance of the proposed method is also assessed using survey data.

Keywords: principal component analysis, bootstrap sampling, congruence

FPM-2-1-3

Statistical Matching for Modeling of Count Data using Philippine National Surveys

Honeylet Santos, *University of the Philippines*
Joseph Ryan Lansangan, *University of the Philippines*
Erniel Barrios, *Monash University Malaysia*

Statistical matching deals with methods of combining different datasets to get information on variables not observed in a single data source. With the goal of estimating a Poisson regression model, this study explores statistical matching techniques and model estimation procedures involving bootstrap. Simulation studies reveal that Poisson regression imputation and Markov chain Monte Carlo (MCMC) imputation as matching methods produce comparable results. Simulations also showed that the bootstrap within method performs well regardless of the matching method used. As an illustration, statistical matching was applied using Philippine national surveys.

Keywords: statistical matching, count data, national surveys, MCMC, bootstrap

FPM-2-2-1

A Bayesian approach for mixed state-space models under skewness and heavy-tails

Luis Mauricio Castro, *Pontificia Universidad Catolica de Chile*
Lina Hernandez-Velasco, *Universidad Santiago de Cali*
Carlos A. Abanto-Valle, *Federal University of Rio de Janeiro*
Dipak K. Dey, *University of Connecticut*

Human immunodeficiency virus (HIV) dynamics have been the focus of epidemiological and biostatistical research during the past decades to understand the progression of acquired immunodeficiency syndrome (AIDS) in the population. Although there are several approaches for modeling HIV dynamics, one of the most popular is based on Gaussian mixed-effects models because of its simplicity from the implementation and interpretation viewpoints. However, in some situations, Gaussian mixed-effects models cannot (a) capture serial correlation existing in longitudinal data, (b) deal with missing observations properly, and (c) accommodate skewness and heavy tails frequently presented in patients' profiles. For those cases, mixed-effects state-space models (MESSM) become a powerful tool for modeling correlated observations, including HIV dynamics, because of their flexibility in modeling the unobserved states and the observations in a simple way. Consequently, our proposal considers an MESSM where the observations' error distribution is a skew-t. This new approach is more flexible and can accommodate data sets exhibiting skewness and heavy tails. Under the Bayesian paradigm, an efficient Markov chain Monte Carlo algorithm is implemented. To evaluate the properties of the proposed models, we carried out some exciting simulation studies, including missing data in the generated data sets. Finally, we illustrate our approach with an application in the AIDS Clinical Trial Group Study 315 (ACTG-315) clinical trial data set.

Keywords: Bayesian inference, Heavy-tailed distribution, Longitudinal data, Mixedeffects, Skewness

FPM-2-2-2

Multivariate Contaminated Normal Censored Regression Model: Properties and Maximum Likelihood Inference

Wan-Lun Wang, *Department of Statistics and Institute of Data Science, National Cheng Kung University, Taiwan*

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

Keywords: Censored data, ECM algorithm, Mild outliers, Multivariate models, Truncation

FPM-2-2-3

Advanced AI in Medical Imaging: Applications in Chest X-Ray, CT Liver Tumors, and Coronary Stenosis

Tai-Been Chen, *Department of Radiological Technology, Teikyo University\Infinity Co. Ltd. Taiwan*
Takahide Okamoto, *Department of Radiological Technology, Teikyo University*
Koharu Kimura, *Department of Radiological Technology, Teikyo University*
Akari Matushima, *Department of Radiological Technology, Teikyo University*

This presentation delves into the revolutionary impact of deep learning on medical imaging, spotlighting three critical applications: chest X-ray classification, computed tomography (CT) liver tumor segmentation, and the detection of stenosis in X-ray coronary angiography. For chest X-ray classification, a hybrid artificial intelligence model, Fusion Convolutional Neural Network (CNN), was developed, integrating five distinct CNN architectures through transfer learning. Trained on a dataset of 5,260 images, including 1,792 normal, 1,658 COVID-19, and 1,800 bacterial pneumonia images, the Fusion CNN model achieved outstanding performance with an accuracy of 99.4% and a Kappa value of 99.1%. In the CT liver tumor segmentation task, fully convolutional networks (FCNs) were utilized, employing backbones such as Xception, InceptionResNetv2, MobileNetv2, ResNet18, and ResNet50. Analyzing 7,190 2D CT images from 131 patients, the ResNet50-based FCN delivered superior results, with a global accuracy of 99.9%, a mean intersection over union (IoU) of 95.4%, and a weighted IoU of 99.8%. For detecting stenosis in X-ray coronary angiography, a real-time YOLO model was applied to video-derived static images. Processing 2,708 images from 120 patients, this model achieved an optimal IoU of 78.8%, with a detection speed of approximately 24 frames per second using ResNet-50. These case studies underscore the effectiveness of advanced deep learning models in enhancing diagnostic accuracy and streamlining medical imaging workflows, marking a significant leap forward in healthcare technology.

Keywords: Chest X-ray Classification, CT Liver Tumor Segmentation, X-ray Coronary Angiography, Convolutional Neural Networks (CNN), Fully Convolutional Networks (FCN), YOLO (You Only Look Once)

FPM-2-3-1

Variable selection for high-dimensional regression models with time series and heteroscedastic errors

Hai-Tang Chiou, *National Chung Cheng University*
Meihui Guo, *National Sun Yat-sen University*
Ching-Kang Ing, *National Tsing Hua University*

Although existing literature on high-dimensional regression models is rich, the vast majority of studies have focused on independent and homogeneous error terms. In this article, 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 the product of two components: one time series component, allowing for a short-memory, long-memory, or conditional heteroscedasticity effect, and a high-dimensional dispersion function accounting for exogenous heteroscedasticity. By making use of the orthogonal greedy algorithm and the high-dimensional information criterion, we propose a new model selection procedure that consistently chooses 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.

Keywords: Heteroscedasticity, High-dimensional information criterion, Orthogonal greedy algorithm, Long-range dependence

FPM-2-3-2

High-dimensional model selection via Chebyshev’ s greedy algorithm

Chien-Tong Lin, *Feng Chia University*
You-Lin Chen, *Cognitive Computing Lab, Baidu Research*
Chi-Shian Dai, *University of Wisconsin-Madison*
Ching-Kang Ing, *National Tsing Hua University*

Assuming sparsity on the regression coefficient is fundamental to ultra-high dimensional variable selection. However, the true sparsity of practical data is typically uncertain, making it necessary to device a variable selection technique that performs well under various sparsity settings. In this talk, we investigate the convergence rate of Chebyshev's greedy algorithm (CGA) for regression models when the true coefficient vector satisfies a general weak sparsity condition. We determine the iteration number of CGA using our developed data-driven approach and demonstrate that the optimal convergence rate can be achieved even when the actual sparsity level is unknown. Our convergence theory relies on the convexity and the smoothness of the population loss function, allowing for the analysis of a broad family of regression models and providing optimality guarantees under weak assumptions. As a specific example, we apply our method to generalized linear models (GLM) and composite quantile regression (CQR) models, and offer the sufficient conditions under which the optimal rate can be achieved. Thorough simulation studies as well as data analysis are provided to support the obtained theory.

Keywords: Variable selection, Weak sparsity, Chebyshev's Greedy Algorithm, High-dimensional Akaike's Information Criterion

FPM-2-3-3

Analyze Additive and Interaction Effects via Collaborative Trees

Chien-Ming Chi, *Institute of Statistical Science, Academia Sinica*

We present Collaborative Trees, a novel tree model designed for regression prediction, along with its bagging version, which aims to analyze complex statistical associations between features and uncover potential patterns inherent in the data. We decompose the mean decrease in impurity from the proposed tree model to analyze the additive and interaction effects of features on the response variable. Additionally, we introduce network diagrams to visually depict how each feature contributes additively to the response and how pairs of features contribute interaction effects. Through a detailed demonstration using an embryo growth dataset, we illustrate how the new statistical tools aid data analysis, both visually and numerically. Moreover, we delve into critical aspects of tree modeling, such as prediction performance, inference stability, and bias in feature importance measures, leveraging real datasets and simulation experiments for comprehensive discussions. On the theory side, we show that Collaborative Trees, built upon a ``sum of trees'' approach with our own innovative tree model regularization, exhibit characteristics akin to matching pursuit, under the assumption of high-dimensional independent binary input features (or one-hot feature groups). This newfound link sheds light on the superior capability of our tree model in estimating additive effects of features, a crucial factor for accurate interaction effect estimation.

Keywords: Sensitivity analysis, Sum of trees, Matching pursuit, Feature interaction, Visual network diagram

FPM-2-4-1

Estimation methods for Three-mode GMANOVA Model with Unobserved Design Matrices

Rei Monden, *Hiroshima University*
Isamu Nagai, *Chukyo University*
Hirokazu Yanagihara, *Hiroshima University*

When n individuals provide data of m items for p measurement time points, three-mode data of size n×m×p is obtained. In order to analyse such three-mode data, the Three-mode Generalized Multivariate Analysis of Variance (3mGMANOVA) model have been proposed. This model describes three-mode data by means of three design matrices, i.e., a design matrix for between individuals, items and measurement time points. However, the available algorithm for this model assumes that a design matrix for between-individuals and/or between-items to be defined prior to the analysis. This is unfortunate since it limits the applicability of 3mGMANOVA model, in practice. To resolve this issue, we propose estimation methods for the 3mGMANOVA model under the following three cases: i) between-individuals design matrix is known, but between-items design matrix is unknown, ii) between-individuals design matrix is unknown, but between-items design matrix is known, and iii) both between-individuals and -items design matrices are unknown. Note that we are focusing on analysing three-mode data where one entity is time points, i.e., we limit ourselves to consider multivariate longitudinal data. Therefore, we can fix a design matrix for measurement time points using a basis function (e.g., polynomial basis or Fourier polynomial basis) depending on a chronological trend that data has. These bases are known to effectively capture chronological trends. In short, we do not need to estimate a design matrix for measurement time points. This study can expand the applicability of 3mGMANOVA model to wider cases. Moreover, there is another model which summarize three-mode data, called three-mode Principal Component Analysis (3mPCA). The current study illustrated similarities and differences between the 3mGMANOVA when both between-individuals and between-items design matrices are unknown and the 3mPCA by applying them to real data.

Keywords: GMANOVA model; Three-mode GMANOVA model; Three-mode Principal Component Analysis model; three-mode data; longitudinal data

FPM-2-4-2

Matrix decomposition structural equation modeling and its generalization

Naoto Yamashita, *Kansai University*

Structural equation modeling (SEM) is a framework that models inter-variable relationships using latent variables. A new formulation of SEM is introduced in the current research. The proposed method, called matrix decomposition structural equation modeling (MDSEM), is a natural extension of matrix decomposition factor analysis. In MDSEM, the data matrix is directly approximated by the product of parameter matrices, called a model part. At the same time, the conventional formulation of SEM fits the covariance structure derived by a hypothetical model to sample the covariance matrix. The research proved that MDSEM shares important properties with MDFA, such as factor score indeterminacy. An iterative algorithm for estimating the parameter matrices in MDSEM is presented. Further, MDSEM is generalized to penalized factor analysis, and some procedures are derived by modifying the penalty term. The variety of penalty terms serves to add special properties to the factor analysis model, such as unique factor scores. Numerical examples are provided to demonstrate the proposed procedures.

Keywords: factor analysis, structural equation modeling, clustering, penalized estimation

FPM-2-4-3

An efficient estimation of functional structural equation models

Yao Zhao, *Temple University*
Kuang-Yao Lee, *Temple University*

Multivariate functional data is increasingly common across various applications, yet discovering causal relationships among these functional entities remains a challenge. In this work, we propose a new algorithm to estimate functional structural equation models (SEM). Our method builds on a newly proposed Non-combinatorial Optimization via Trace Exponential and Augmented lagRangian for Structure learning (NOTEARS, Zheng et al., 2018), and extends it to functional data. A key element of our approach is to introduce a novel continuous constraint whose argument is entirely based on linear operators. Additionally, our algorithm is highly scalable. Simulation studies show that it improves both accuracy and computational efficiency, particularly in large graphs. We further demonstrate its efficacy through an application to an electroencephalogram (EEG) dataset for brain effective connectivity analysis.

Keywords: Brain connectivity analysis; Continuous constrained optimization; Directed acyclic graphs; Multivariate functional data; Structural equation models

FPM-2-5-1

Cox regression with a MNAR covariate via Heckman selection model

Chiu-Hsieh (Paul) Hsu, *University of Arizona*
Mandi Yu, *NCI*
Valentina Petkov, *NCI*

We consider the situation of estimating Cox regression, where a covariate is subject to missing not at random (MNAR), and some of the observed data (e.g. survival data and auxiliary variables) may be predictive of the missing covariate. All existing approaches for handling MNAR data require a full specification of the relationship between missing values and missingness probabilities. Under the Heckman selection modelling framework, where the correlation coefficient between missing values and selection probabilities is a standardized measure for the magnitude of MNAR, we propose an approach using a nonparametric multiple imputation strategy. Specifically, we propose to fit the Heckman selection model to the observed data to derive the correlation coefficient between missing values and selection probabilities and two working models: one for predicting the missing covariates and the other for predicting the missing probabilities. For each missing covariate observation, the correlation coefficient and these two working models are used to define a nearest neighbour imputing set. This set is then used to non-parametrically impute covariate values for the missing observation. Upon the completion of imputation, Cox regression is performed on the multiply imputed datasets to estimate the regression coefficients. The proposed approach does not directly use the Heckman selection model to perform imputation. Hence, the proposed approach might be more robust against misspecifications of the Heckman selection model. We conduct a simulation study to evaluate the performance of the proposed approach. We show the proposed approach can reduce bias due to MNAR and is less susceptible to misspecification of the Heckman selection model than the existing approaches. We also apply the proposed approach to study the effect of HPV infection on survival of head and neck cancer patients using SEER 2010-2017, where HPV infection status is not always collected, and the missing mechanism is suspected to be MNAR.

Keywords: Cox regression; Heckman selection model; Missing not at random; Multiple imputation

FPM-2-5-2

Efficient Data Integration Under Prior Probability Shift

Ming-Yueh Huang, *Institute of Statistical Science, Academia Sinica*
Jing Qin, *National Institute of Health*
Chiung-Yu Huang, *University of San Francisco*

Conventional supervised learning usually operates under the premise that data are collected from a homogeneous underlying population. However, challenges may arise when integrating new data from different populations, resulting in a phenomenon known as dataset shift. This talk focuses on prior probability shift, a specific form of dataset shift, where the distribution of the outcome varies across different datasets but the conditional distribution of features given the outcome remains the same. To tackle the challenges posed by this shift, we propose a maximum likelihood estimation method that efficiently amalgamates information from multiple sources under prior probability shift. Unlike existing methods that are restricted to discrete outcomes, the proposed approach accommodates both discrete and continuous outcomes. It also handles high-dimensional covariate vectors through variable selection using an adaptive LASSO penalty, producing efficient estimates that possess the oracle property. Moreover, a novel semiparametric likelihood ratio test is proposed to check the validity of prior probability shift assumptions by embedding the null conditional density function into Neyman' s smooth alternatives and testing study-specific parameters. We demonstrate the effectiveness of our proposed method through extensive simulations and two real data examples. The proposed methods serve as a useful addition to the repertoire of tools for addressing challenges that arise from dataset shifts in machine learning.

FPM-2-5-3

Fatty Liver Classification via Risk Controlled Neural Networks Trained on Grouped Ultrasound Image Data

Tso-Jung Yen, *Institute of Statistical Science, Academia Sinica*
Chih-Ting Yang, *Institute of Statistical Science, Academia Sinica*
Yi-Ju Lee, *Institute of Statistical Science, Academia Sinica*
Chun-houh Chen, *Institute of Statistical Science, Academia Sinica*
Hsin-Chou Yang, *Institute of Statistical Science, Academia Sinica*

Ultrasound imaging is a widely used technique for fatty liver diagnosis as it is practically affordable and can be quickly deployed by using suitable devices. When it is applied to a patient, multiple images of the targeted tissues are produced. We propose a machine learning model for fatty liver diagnosis from multiple ultrasound images. The machine learning model extracts features of the ultrasound images by using a pre-trained image encoder. It further produces a summary embedding on these features by using a graph neural network. The summary embedding is used as input for a classifier on fatty liver diagnosis. We train the machine learning model on a ultrasound image dataset collected by Taiwan Biobank. We also carry out risk control on the machine learning model using conformal prediction. Under the risk control procedure, the classifier can improve the results with high probabilistic guarantees.

FPM-2-6-1

TLRR-TF: A Fast Tensor Low-Rank Representation via Tri-Factorization

Youngwook Kwon, *Department of Statistics, Seoul National University, Korea, yung9171@snu.ac.kr*
Hee-Seok Oh, *Department of Statistics, Seoul National University, Korea, heeseok@stats.snu.ac.kr*

Low-rank representation (LRR) is effective in segmenting data points into their intrinsic subspaces and representing the original signal using the lowest-rank criterion. Recently, LRR methods for tensor data have gained increasing attention. Still, most tensor LRR algorithms require computing the singular value decomposition (SVD), which is computationally expensive when dealing with large tensor data. In this paper, we propose a new robust tensor LRR model using a fast tri-factorization approach that approximates the representation tensor as the product of three smaller tensor components. The main advantage of the proposed method is that it mitigates the computational cost of applying nuclear norm minimization directly to the large original tensor. Furthermore, it simultaneously processes sparse and dense Gaussian noises to recover clean tensors effectively. Through extensive experimental experiments, including synthetic data and real-world image data, we demonstrate the promising performance of the proposed method in both clustering/denoising accuracy and computing time.

Keywords: Low-rank representation, Tensor decomposition, Subspace clustering, Tensor recovery, Tri-factorization

FPM-2-6-2

Image generator for tabular data based on non-Euclidean metrics for CNN-based deep learning classification

Han-Ming Wu, *Department of Statistics, National Chengchi University*
Yu-Rong Lin, *Department of Statistics, National Chengchi University*

Statistical analysis methods have been widely applied in various fields such as finance, industry, biomedicine, public health, and social and environmental sciences, and its primary data format is tabular data. However, traditional statistical methods face challenges when dealing with high-dimensional data and nonlinear relationships among variables. In contrast, Convolutional Neural Networks (CNNs) are primarily used for tasks such as image recognition and object detection in computer vision, handling data in the form of images. CNNs do not require preset assumptions about the dataset and excel in learning complex feature relationships, but it has weaker model interpretability. Therefore, this study aims to integrate statistical analysis with CNNs, particularly focusing on methods to transform tabular data into sequences of images to enhance the interpretability and predictive accuracy of deep learning algorithms. We propose the use of non-Euclidean distances such as Geodesic Distance, Jensen-Shannon Distance, and Wasserstein Distance to capture the nonlinear feature structures of the data, thereby improving the existing techniques such as the Image Generator for Tabular Data (IGTD) based on Euclidean distance. By combining the powerful predictive capabilities of deep learning and the interpretability of statistical methods, we aim to address challenges encountered by traditional methods in handling high-dimensional tabular data. Furthermore, we expect to apply CNNs in more fields which are primarily focused on tabular data research, to provide a flexible and user-friendly tool for interdisciplinary research.

Keywords: Convolutional Neural Networks, Tabular Data, Wasserstein Distance

FPM-2-6-3

A Comparision on the Abstractive Summarization of Large Language Models

Hansol Lee, *Dongguk University, Seoul, KOREA*
Yung-seop Lee, *Dongguk University, Seoul, KOREA*

This study conducts a comparative evaluation of the abstractive summarization capabilities of Large Language Models (LLMs) using a diverse set of datasets. As LLMs have demonstrated significant advancements in natural language processing, their ability to generate coherent and accurate summaries has garnered increasing attention. This research employs eight datasets across various domains, including news articles (CNN/DailyMail, XSum), scientific papers (arXiv, PubMed), legal documents (BillSum), government reports (GovReport), and online content (Reddit TIFU, WikiHow). Nine state-of-the-art LLMs, including GPT-3.5, Llama 3.1, and Mistral, were selected based on their prominence and varied architectures. The summarization quality of each model was assessed using widely recognized metrics such as ROUGE, METEOR, and BERTScore, which consider both n-gram overlap and semantic similarity. The findings highlight performance variations across different text types, with certain models excelling in specific domains. Additionally, the study reveals the generalization capabilities of these models when faced with diverse and complex textual data. The results provide valuable insights into the strengths and limitations of current LLMs in abstractive summarization and suggest pathways for further improvements in model development and evaluation methodologies. This work contributes to the ongoing exploration of LLM applications in various fields, offering a comprehensive assessment of their potential to enhance automated summarization tasks.

Keywords: Abstractive Summarization, Large Language Models, Natural Language Processing, Model Evaluation, Dataset Evaluation

FPM-2-7-1

A novel class of nonparametric control charts with exceedance probability criterion for Phase I

Hong-Ji Yang, *National Cheng Kung University*
Chung-I Li, *National Cheng Kung University*

This study advances Phase I statistical process monitoring by integrating estimation uncertainty into control charts through the exceedance probability criterion (EPC), a method rapidly adopted in recent literature. The EPC ensures in-control performance with a high nominal coverage probability (Pn), thus reducing false alarms. We propose a new class of nonparametric methods using fractional order statistics, improving existing semiparametric approaches for tail-extrapolated quantile estimation. Our methods, which include both the use and non-use of activation functions commonly found in deep neural networks, demonstrate superior performance over existing nonparametric methods, particularly in small to moderate sample sizes, typical of Phase I studies, due to cost and time constraints. Our approaches allow practitioners to implement adjustable control limits based on the EPC, with the flexibility to tune the desired Pn for specific applications in production environments. We also provide practical guidance, supported by simulations, real-case studies, and R code examples, to facilitate the adoption of these methods.

Keywords: Statistical process monitoring, Adaptive activation function, Exceedance probability criterion, Nonparametric control chart, Phase I stage

FPM-2-7-2

Evaluation of methods for approximating posterior probability of dose-response relationship models in Bayesian benchmark dose methods for risk assessment

Sota Minewaki, *Department of Information and Computer Technology, Tokyo University of Science Graduate School of Engineering*
Takashi Sozu, *Department of Information and Computer Technology, Faculty of Engineering, Tokyo University of Science*
Tomohiro Ohigashi, *Department of Information and Computer Technology, Faculty of Engineering, Tokyo University of Science*

The benchmark dose (BMD) method is used to determine the point of departure for the acceptable daily intake of substances for humans using experimental animal data. In the BMD method, multiple dose-response relationship models are considered, and the BMD (a dose associated with a specified change in response relative to the control group) is estimated. Bayesian model averaging (BMA), in which models are averaged based on their posterior probabilities, has recently been commonly used. In BMA, the marginal likelihood (ML) is calculated to determine the posterior probability of models. Several methods of ML approximation have been proposed, because the ML is not calculated analytically. Some software packages, such as BBMD and ToxicR, have been developed for the BMD method using BMA. However, the performances of the ML approximations in the BMD method have not been compared.
We evaluated the performance of ML approximation methods in the BMD method and their effects on BMD estimation through numerical examples using four real experimental datasets. Eight models and three prior distributions used in BBMD and ToxicR were assumed. Five ML approximation methods were assumed: (1) MLE-based Schwarz criterion, (2) MCMC-based Schwarz criterion, (3) Laplace approximation, (4) density estimation, and (5) bridge sampling. We evaluated the ML approximation bias (the difference between the approximation method using the ML approximation method and reference value calculated using Monte Carlo integration) and the estimation bias of BMD (the differences between the estimates obtained using the approximation value and the reference value).
The approximation and estimation biases of bridge sampling were the smallest

regardless of the dataset or prior distributions. Both the approximation and estimation biases were large for some datasets under the implemented software settings (MCMC-based Schwarz criterion and Laplace approximation). The approximation biases of the density estimation were overall relatively small but were large for some datasets.

Keywords: Benchmark dose estimation, Risk assessment, Bayesian model averaging, Marginal likelihood approximation, Bridge sampling

FPM-2-7-3

Compositional regression analysis for categorical predictors not jointly observed with response

Juyeon Oh, *Department of Mathematics and Statistics, Chonnam National University*
Kwangmin Lee, *Department of Big Data Convergence, Chonnam National University*

In real-world data analysis, there are cases where variables are not jointly observed. In particular, when each variable is investigated by an independent survey, it is difficult to estimate the correlation between the variables. In our study, we propose a methodology to address this issue by combining individuals with similar characteristics into a group and treating this group as a single data point. We present the theoretical foundation and verify this approach. Additionally, in surveys that use multiple-choice questions, the proportion of responses to each question can be treated as compositional data. Compositional data is useful when the composition of components is important, but standard statistical models are difficult to apply because it has a sum constraint. Furthermore, these surveys often contain numerous questions, and including all of them in a regression analysis can lead to a complex model and reduce interpretability. To address this, we use a log contrast model that relaxes the sum constraint of compositional data with a bi-level selection method using the exponential penalty. We demonstrate the superiority of our model compared to other methods via simulations and conduct real data analysis using two separate national statistical datasets.

Keywords: non-jointly observed, compositional regression, log-contrast model, bi-level selection

FPM-2-8-1

Empirical comparison of maximum likelihood and minimum Cramér-von Mises for input analysis in stochastic simulations

David Fernando Munoz, *Instituto Tecnologico Autonomo de Mexico*

In a previous paper, Muñoz and Villafuerte (2015) introduced a software (Simple Analyzer) to fit sample data to the most frequently used probability distribution families as well as to generate sample data from each distribution to test how fitting procedures perform. This software was intended for input analysis of stochastic simulations and, in addition to the estimation of the parameters of the corresponding family of distributions, the joint estimation of a shift and a scale parameter is considered for each family of distributions. In most cases, the maximum likelihood (ML) method was applied to estimate the parameters of a family of distributions, except for the case of the lognormal distribution since, as is well known, maximum ML estimators for a shifted lognormal distribution may not exist. For this case, the Simple Analyzer incorporated three additional estimation methods, the one proposed by Nagatsuka et al. (2013), a method proposed by Muñoz et al. (2016) and using the method of Cooke (1979) to estimate the shift and then maximum likelihood to estimate the other parameters.

In this paper, we introduce a new version of the Simple Analyzer that incorporates graphs of the adjusted density functions and P-P plots, for a set of selected families (in a single figure) to facilitate visual comparison of fitting the sample data for different distribution families. Furthermore, for all the distribution families considered in the Simple Analyzer, we incorporated parameter estimation based on the minimization of the Cramér-von Mises statistic (see, e.g., Luong and Blier-Wong, 2017). We present experimental results to compare the performance of ML estimation versus minimum Cramér-von. Our preliminary results show that the new implementation performs very well, with no requirements to produce reasonable estimates.

Keywords: Simulation input analysis, Cramér-von Mises statistic, distribution fitting, optimization, stochastic simulation

FPM-2-8-2

Multi-objective Optimization method combined AWA and U-NSGA-3

Masami Murakami, *Doshisha university, Japan*
Hiroshi Yadohisa, *Doshisha university, Japan*

A multi-objective optimization problem involves simultaneously optimizing multiple objective functions. Such problems have multiple optimal solutions, called Pareto solutions, which form a set. Each Pareto solution exhibits a trade-off relationship, wherein the optimization of one objective function worsens another objective function' s optimization.

A key challenge in the algorithmic study of multi-objective optimization is how close the solutions obtained are to the true Pareto solution, and how well these solutions cover the wide range of Pareto solutions. To address these challenges, U-NSGA-3 and adaptive weighted aggregation (AWA) have been proposed as effective solutions.

U-NSGA-3 is a genetic algorithm (GA) capable of generating a wide range of solutions that approximate the true Pareto front beyond the local Pareto solution problems with up to 15 objectives. However, due to the inherent nature of GAs, it becomes difficult to achieve high accuracy in terms of how close they are to the true Pareto solution (hereafter referred to as "accuracy").

However, the AWA can produce highly accurate solutions from a small set of initial solutions by adjusting the weight of each objective function, and repeating scalarization and optimization. However, these advantages have been primarily demonstrated in problems with two to six objectives. For problems with more than seven objectives, the solution accuracy may deteriorate.

In this paper, we propose a method that combines U-NSGA-3 and AWA using a serial approach. Specifically, the proposed method is expected to obtain more accurate solutions than U-NSGA-3 for problems with 2–6 objectives due to the features of the AWA. Additionally, for problems with 7–15 objectives, it is expected to yield more accurate solutions than the AWA due to the features of U-NSGA-3.

Keywords: multi-objective optimization, genetic algorithm, scalarization, multi-start strategies

FPM-2-8-3

Probabilistic Decoding Algorithms with Efficient Pooling Designs for Two Types of Defectives

Hiroyasu Matsushima, *Data Science and AI Innovation Research Promotion Center, Shiga University\Center for Training Professors in Statistics, The Institute of Statistical Mathematics*
Yusuke Tajima, *Data Science and AI Innovation Research Promotion Center, Shiga University\Center for Training Professors in Statistics, The Institute of Statistical Mathematics*
Xiao-Nan Lu, *Department of Electrical, Electronic and Computer Engineering, Gifu University*
Masakazu Jimbo, *Center for Training Professors in Statistics, The Institute of Statistical Mathematics\Data Science and AI Innovation Research Promotion Center, Shiga University*

Group testing is a test method that identifies defectives from the results of testing pools of various combinations of items. This paper considers the case where there are two types of defective items, A and B types, and addresses to concurrently identify them in group testing. In order to screen two types of defectives in group testing, this paper proposes a hybrid algorithm with belief propagation (BP) and Markov chain Monte Carlo (MCMC). Concretely, group testing for two types of defectives consists of a set of pools that test for defectives of A and B and a set of pools that test for either type of defectives. A collection of pools is constructed by using finite affine geometry. Through simulation experiments, the screening performance of the BP and MCMC algorithms is evaluated. Experimental results show that the proposed algorithm has high screening and identification abilities even when the number of defectives exceeds the identification ability of the pooling design.

Keywords: Group testing, Pooling design, Belief Propagation, Markov Chain Monte Carlo

PARALLEL SESSIONS 同步演講場次 SAM-1

10:20 - 11:50 Saturday, December 14

SAM-1-1	**Young Scholar Session** Organizer: CSAT Committee Chair: Yi-Ting Hwang	**Socrates** **蘇格拉底廳**
SAM-1-2	**Inference on High Dimensional Covariance Matrix** Organizer: Johan Lim Chair: Johan Lim	**Nietzsche** **尼采廳**
SAM-1-3	**Particle Swarm Optimization and Its Application in Clinical Trials and Medical Imaging** Organizer: Grace Hyun Kim Chair: Grace Hyun Kim	**Michelangelo** **米開朗基羅廳**
SAM-1-4	**Advancements in Predictive Modeling and Data Integration with Applications** Organizer: Ching-Ti Liu Chair: Ching-Ti Liu	**Raphael** **拉斐爾廳**
SAM-1-5	**Dimension Reduction Methods** Organizer: IACS-ARS Committee Chair: Li-yu Daisy Liu	**Alexander** **亞歷山大廳**

SAM-1-1-1

Generalized Kernel Two-Sample Tests

Hoseung Song, *Department of Industrial and Systems Engineering, KAIST*

Kernel two-sample tests have been widely used for multivariate data to test equality of distributions. However, existing tests based on mapping distributions into a reproducing kernel Hilbert space mainly target specific alternatives and do not work well for some scenarios when the dimension of the data is moderate to high due to the curse of dimensionality. We propose a new test statistic that makes use of a common pattern under moderate and high dimensions and achieves substantial power improvements over existing kernel two-sample tests for a wide range of alternatives. We also propose alternative testing procedures that maintain high power with low computational cost, offering easy off-the-shelf tools for large datasets. The new approaches are compared to other state-of-the-art tests under various settings and show good performance. All proposed methods are implemented in an R package kerTests.

SAM-1-1-2

Asymptotic Expansion of the Distributions of Estimators Related to Stochastic Processes Driven by a Fractional Brownian Motion and Their Statistical Applications

Hayate Yamagishi, *Graduate School of Mathematical Sciences, University of Tokyo*

We study the asymptotic expansions of the distributions of estimators related to stochastic processes driven by fractional Brownian motion (fBm). Specifically, we examine an estimator of the Hurst parameter of fBm ($H\in(0,1)$), the same Hurst estimator for the solution process of a stochastic differential equation (SDE) driven by fBm ($H\in(1/2,1)$), and the power variation of the solution of the SDE.

The main methodology involves a series of techniques for the asymptotic expansion of the distribution of Wiener functionals, recently developed following advances in the field of limit theorems on the Wiener space. To apply the general theory, we estimate the order of functionals sharing a certain common structure. To this end, we introduce a systematic method using exponents defined via a graphical representation of the structure of these functionals.

Numerical experiments demonstrate that the approximate density functions obtained from the asymptotic expansions exhibit higher accuracy than normal approximations. Furthermore, we discuss applications to statistical problems, such as second-order modifications of the estimators and bootstrap approximations.

SAM-1-1-3

Sigma-pattern and optimal regular fractional designs for computer experiments

Cheng-Yu Sun, *Institute of Statistics and Data Science, National Tsing-Hua University*

Space-filling designs play a vital role in computer experiments. Common criteria for selecting such designs are either distance- or discrepancy-based. Recently, Tian and Xu (2022) introduced a minimum aberration-type criterion known as the Space-Filling Pattern (SFP). This criterion examines whether a design exhibits stratifications on a series of grids, and can effectively distinguish strong orthogonal arrays of the same strengths. Subsequently, Shi and Xu (2023) refined the SFP to the Stratification Pattern (SP). They showed that designs excelling under the SFP can yield better surrogate models compared to designs optimized for many other uniformity criteria. In this paper, we propose a novel pattern, referred to as the Sigma-pattern, which is closely related to the SP. We demonstrate that the Sigma-pattern has advantages over the SP in certain scenarios, and provide a new justification for both patterns. Then, our focus shifts to the construction of space-filling regular designs. We show that the Sigma-pattern of a regular design can be determined by counting different types of words of given lengths. This result allows for a complete search for the most space-filling regular designs of moderate run sizes.

SAM-1-2-1

High-dimensional Missing Data Imputation Via Undirected Graphical Model

Seongoh Park, *Sungshin Women's University*
Yoonah Lee, *Sungshin Women's University*

Multiple imputation is a practical approach in analyzing incomplete data, with multiple imputation by chained equations (MICE) being popularly used. MICE specifies a conditional distribution for each variable to be imputed, but esti- mating it is inherently a high-dimensional problem for large-scale data. Existing approaches propose to utilize regularized regression models, such as lasso. How- ever, the estimation of them occurs iteratively across all incomplete variables, leading to a considerable increase in computational burden, as demonstrated in our simulation study. To overcome this computational bottleneck, we propose a novel method that estimates the conditional independence structure among variables before the imputation procedure. We extract such information from an undirected graphical model, leveraging the graphical lasso method based on the inverse probability weighting estimator. Our simulation study verifies the pro- posed method is way faster against the existing methods, while still maintaining comparable imputation performance.

Keywords: Incomplete data, Gaussian graphical model, MICE, Conditional independence, Inverse probability weighting estimator

SAM-1-2-2

Bayesian inference on spiked eigenstructure of high-dimensional covariances

Kwangmin Lee, *Chonnam National University*
Sewon Park, *Samsung SDS*
Jaeyong Lee, *Seoul National University*

We consider Bayesian inference on the spiked eigenstructures of high-dimensional covariance matrices; in other words, we are interested in estimating the eigenvalues and corresponding eigenvectors of high-dimensional covariance matrices in which a few eigenvalues are significantly larger than the rest. We impose an inverse-Wishart prior distribution on the unknown covariance matrix and derive the posterior distributions of the eigenvalues and eigenvectors by transforming the posterior distribution of the covariance matrix. We justify the proposed method by demonstrating that the posterior distribution of the spiked eigenvalues and corresponding eigenvectors converges to the true parameters under the spiked high-dimensional covariance assumption. Furthermore, we extend the proposed method to Bayesian principal component regression analysis. To achieve this, we reparameterize the principal component regression model as a multivariate Gaussian model with a spiked covariance matrix for the joint vectors of the response and predictors. Next, we obtain the posterior distribution of the regression coefficients by transforming the inverse-Wishart posterior on the covariance matrix. Compared to existing principal component regression methods, the proposed approach improves the interval estimation of the regression coefficients by accounting for uncertainties in the principal scores. We prove that the derived posterior distribution of the regression coefficients converges to the true parameters, theoretically validating the principal component regression analysis in high-dimensional settings. Simulation studies and real data analysis demonstrate that our proposed method outperforms all existing methods in quantifying uncertainty.

Keywords: Bayesian inference, High-dimensional statistics, Spiked covariance, Principal component analysis, Principal component regression

SAM-1-2-3

HP-ACCORD: Learning Massive-scale Partial Correlation Networks with High-performance Computing

Sungdong Lee, *National University of Singapore*
Joshua Bang, *University of California, Santa Barbara*
Youngrae Kim, *National University of Singapore*
Hyungwon Choi, *National University of Singapore*
Sang-Yun Oh, *University of California, Santa Barbara*
Joong-Ho won, *Seoul National University*

While graphical models are effective in estimating sparse partial correlation structure, achieving the computational scalability required to handle modern massive-scale data demands further consideration. We introduce a novel pseudolikelihood-based graphical model framework called Asymmetric ConCORD (ACCORD). To ensure both statistical consistency and computational scalability on the estimation, ACCORD reparameterizes the target precision matrix that preserves sparsity pattern and minimizes an L1-penalized empirical risk based on a new loss function. The resulting optimization problem facilitates a provably fast computation algorithm using a novel operator-splitting approach and communication-avoiding distributed matrix computation. HP-ACCORD, our implementation in high-performance computing environments, can manage datasets with up to 1 million dimensions with underlying graphical structures.

Keywords: High-performance statistical computing, Graphical model selection, Ultrahigh-dimensional data, Pseudolikelihood, Communication-avoiding linear algebra

SAM-1-2-4

Scalable Bayesian inference on high-dimensional multivariate linear regression

Xuan Cao, *University of Cincinnati*
Kyoungjae Lee, *Sungkyunkwan University*

We consider jointly estimating the coefficient matrix and the error precision matrix in high- dimensional multivariate linear regression models. Bayesian methods in this context often face computational challenges, leading to previous approaches that either utilize a generalized likelihood without ensuring the positive definiteness of the precision matrix or rely on maximization algorithms targeting only the posterior mode, thus failing to address uncertainty. In this work, we propose two Bayesian methods: an exact method and an approximate two-step method. We first propose an exact method based on spike and slab priors for the coefficient matrix and DAG-Wishart prior for the error precision matrix, whose computational complexity is comparable to the state-of-the-art generalized likelihood-based Bayesian method. To further enhance scalability, a two-step approach is developed by ignoring the dependency structure among response variables. This method estimates the coefficient matrix first, followed by the calculation of the posterior of the error precision matrix based on the estimated errors. We validate the two-step method by demonstrating (i) selection consistency and posterior convergence rates for the coefficient matrix and (ii) selection consistency for the directed acyclic graph (DAG) of errors. We demonstrate the practical performance of proposed methods through synthetic and real data analysis.

Keywords: Generalized likelihood, Selection consistency, Posterior convergence rate

SAM-1-3-1

Particle Swarm Optimization as a general-purpose optimization tool

Weng Kee Wong, *University of California at Los Angeles*
Ray-Bing Chen, *Natinal Cheng Kung University*
Ping Yang Chen, *New Taipei University*

Particle Swarm Optimization (PSO) algorithm is based on swarm intelligence and widely used in the field of Artificial Intelligence. Like many other nature-inspired metaheuristic algorithms, it is already widely used to tackle all sorts of hard optimization problems across disciplines, particularly in engineering and computer science. Interestingly, it is less used in the statistical sciences. Their meteoric rise in popularity is due to their ease of use, speed, availability of codes across different platforms and above all, their apparent lack of technical assumptions for them to work reasonably well. I focus on an exemplary algorithm PSO and, as examples, present some of the recent applications of PSO to tackle challenging problems in the biomedical sciences. If time permits, I will also discuss PSO variants and design strategies for problems with multiple objectives.

SAM-1-3-2

Development and Validation of a CT based Imaging Biomarker – Single Time point Prediction – Using Quantum Particle Swam Optimization for Predicting Disease Progression in Idiopathic Pulmonary Fibrosis

Grace Hyun Kim, *UCLA*

Disease progression in idiopathic pulmonary fibrosis (IPF) is unpredictable at the time of diagnosis, underscoring the need for reliable prognostic markers. This study aims to evaluate a novel machine-learning-based radiomic marker derived from high-resolution CT (HRCT) scans—termed the Single Time Point (STP) score—as a predictive biomarker for progression-free survival (PFS) in IPF. Using a combination of quantum particle swarm optimization (QPSO) and random forest algorithms, we identified key radiomic features that predict IPF progression within 9 to 12 months. The radiomic features were derived from classic representative region of interests (ROIs) at baseline HRCT, which were reviewed from the paired HRCT scans of baseline and 9-to-12-month follow-up, contoured at baseline HRCT, and labelled as stable or progression by a thoracic radiologist. The STP score was computed using five steps: (1) denoise HRCT using total variation to account for heterogenous noise from multi-center studies, (2) sampling a voxel within a 4-by-4 grid, (3) calculate radiomic features optimally selected by QPSO, (4) classify the predictive progressive area, (5) quantify the score in whole lung. A cohort of 250 subjects with IPF were collected from two-centers from March 2004 and October 2019. The median STP score was 30%. After adjusting for clinical risk factors, an STP ≥ 30% was ignificantly associated with disease progression (Hazard Ratio=1.57; p=0.017; C-index=0.581). The radiomic STP score at initial HRCT was predictive of progression in IPF patients, outperforming baseline QLF when adjusted for sex, age, and pulmonary function status. This study offers validation of the STP marker, which may be useful for enriching cohorts in clinical trials and informing treatment decisions. This study validates the STP marker, which may be valuable for enriching cohorts in clinical trials and informing treatment decisions. Furthermore, a prospective study (IS-PPF, NCT06162884) will be initiated to further validate STP as a prognostic marker.

Keywords: Particle swarm optimization; Prediction; Machine Learning

SAM-1-3-3

MM Optimization Algorithms for Analyzing Big Biomedical Data

Hua Zhou, *University of California, Los Angeles*

The majorization-minimization (MM) principle is an extremely general framework for deriving optimization algorithms. It includes the expectation-maximization (EM) algorithm, proximal gradient algorithm, concave-convex procedure, quadratic lower bound algorithm, and proximal distance algorithm as special cases. Besides numerous applications in statistics, optimization, and imaging, the MM principle finds wide applications in large-scale machine learning problems such as matrix completion, discriminant analysis, and nonnegative matrix factorizations. This talk presents two applications of the MM principle in the big data setting. In the first application, we derive a parallel block least squares algorithm that allows parallel update of regression coefficients with a large feature matrix partitioned by columns. In the second application, we introduce a de-weighting technique for weighted least squares that dramatically accelerates the fitting of generalized linear models and quantile regression.

Keywords: Optimization, Majorization-Minimization, MM algorithm, Weighted Least Squares

SAM-1-4-1

On p-value combination of independent and frequent signals: asymptotic efficiency and Fisher ensemble

Chung Chang, *National Sun Yat-sen University*

Combining p-values to integrate multiple effects is of long-standing interest in social science and biomedical research. In this paper, we focus on revisiting a classical scenario closely related to meta-analysis, which combines a relatively small (finite and fixed) number of p-values while the sample size for generating each p-value is large (asymptotically goes to infinity). We evaluate a list of traditional and recently developed modified Fisher' s methods to investigate their asymptotic efficiencies and finite-sample numerical performance. The result concludes that Fisher and adaptively weighted Fisher method have top performance and complementary advantages across different proportions of true signals. Consequently, we propose an ensemble method, namely Fisher ensemble, to combine the two top performing Fisher-related methods using a robust truncated Cauchy ensemble approach. We show that Fisher ensemble achieves asymptotic Bahadur optimality and integrates the strengths of Fisher and adaptively weighted Fisher methods in simulations.

Keywords: p-value combination; ensemble method

SAM-1-4-2

Elucidating the Type 2 Diabetes Epigenetic Landscapes: Multi-Omics Analysis Unveils Novel CpG Sites and Their Association with Cardiometabolic Traits

Ren-Hua Chung, *National Health Research Institutes*

Type 2 Diabetes (T2D) is a complex, multifactorial disease that poses a significant global health challenge. Despite numerous genetic variants identified by Genome-Wide Association Studies (GWAS), the functional impacts of these variants, particularly those in non-coding regions, remain poorly understood. This study employs a multi-omics approach, integrating methylome-wide association studies (MWAS), Mendelian Randomization (MR), and functional analyses in human pancreatic cells and mouse models to explore the consequences of genetic variants on T2D. Utilizing large-scale GWAS summary statistics and a DNA methylation prediction model, we identified 87 significant CpGs associated with T2D risk, including 13 novel sites, across European populations, with replication in additional datasets. Our results indicate substantial overlap of these CpGs with cardiometabolic traits and confirm their global relevance through trans-ethnic effects in East Asians. Functional analyses in pancreatic alpha and beta cells highlighted the regulatory roles of these CpGs in genes critical for glucose metabolism, especially the PPP1R3B gene, which exhibited differential expression related to T2D in both cell types. Mouse models further validated PPP1R3B's involvement in glucose regulation. These findings uncover novel genetic susceptibilities and offer potential targets for T2D therapy.

SAM-1-4-3

Generalized Estimating Equations Boosting (GEEB) machine for correlated data

Chao-Yu Guo, *Division of Biostatistics and Data Science, Institute of Public Health, National Yang Ming Chiao Tung University*
Yuan-Wey Wang, *Division of Biostatistics and Data Science, Institute of Public Health, National Yang Ming Chiao Tung University*
Hsin-Chou Yang, *Institute of Statistical Science, Academia Sinica, Taipei, Taiwan*
i-Hau Chen, *Institute of Statistical Science, Academia Sinica, Taipei, Taiwan*

Rapid development in data science enables machine learning and artificial intelligence to be the most popular research tools across various disciplines. While numerous articles have shown decent predictive ability, little research has examined the impact of complex correlated data. We aim to develop a more accurate model under repeated measures or hierarchical data structures. Therefore, this study proposes a novel algorithm, the Generalized Estimating Equations Boosting (GEEB) machine, to integrate the gradient boosting technique into the benchmark statistical approach that deals with the correlated data, the generalized Estimating Equations (GEE). Unlike the previous gradient boosting utilizing all input features, we randomly select some input features when building the model to reduce predictive errors. The simulation study evaluates the predictive performance of the GEEB, GEE, eXtreme Gradient Boosting (XGBoost), and Support Vector Machine (SVM) across several hierarchical structures with different sample sizes. Results suggest that the new strategy GEEB outperforms the GEE and demonstrates superior predictive accuracy than the SVM and XGBoost in most situations. An application to a real-world dataset, the Forest Fire Data, also revealed that the GEEB reduced mean squared errors by 4.5% to 25% compared to GEE, XGBoost, and SVM. This research also provides a freely available R function that could implement the GEEB machine effortlessly for longitudinal or hierarchical data.

Keywords: Correlated data, Hierarchical data, Generalized Estimating Equations, Machine learning, Gradient boosting

SAM-1-4-4

Model-based predictive accuracy with a case study on Taiwan AIDS data

Yi-Kuan Tseng, *National Central University*

To connect with a survival model, time-dependent ROC curves have been derived for the Cox model with fixed covariates. These curves serve not only for dynamic prediction but also for evaluating the model based on the average time-dependent AUC (the area under the ROC curves). However, the Cox regression model relies on a proportional hazards assumption, which may fail in some medical studies. In such situations, we've developed an approach to replace the Cox model with both the accelerated failure time (AFT) model and the proportional odds (PO) model, enabling the derivation of time-dependent sensitivity and specificity. Since the weighted average time-dependent AUC can be proven to be a concordance measure (C-index), for prediction purposes, we can select a better model among the incorporated these hazards regression models by choosing the one with higher predictive accuracy. Simulation studies are conducted to evaluate the performance of the proposed approach. Additionally, a case study using Taiwanese HIV/AIDS cohort data is utilized to illustrate the usefulness of the proposed model-based time-dependent AUC and predictive accuracy.

Keywords: AFT model, Cox model, joint model, proportional odds model, timedependent AUC

SAM-1-5-1

Projective Resampling with the Minimum Average Variance Estimator: Sufficiently Reducing Dimension with Multivariate Response

Seoyoung Kim, *Ewha Womans University*
Jae Keun Yoo, *Ewha Womans University*

This paper addresses the challenge of dimension reduction in datasets where both the predictor and response variables are multivariate. We introduce a novel approach that integrates the Minimum Average Variance Estimator (MAVE) with the Projective Resampling (PR) method to render its applicability in multivariate response settings. By applying MAVE within the framework of Projective Resampling, we achieve an efficient dimension reduction, ensuring the exhaustive recovery of the central mean subspaces and eradicating strong assumptions on the distribution of predictors. The proposed methodology is validated through a series of simulations to evaluate its effectiveness and robustness in various multivariate contexts. Further refinements were made to improve the computational performance as well.

Keywords: Dimension Reduction , Multivariate nonlinear regression, Central mean subspace, Average derivative estimation

SAM-1-5-2

Comparative Analysis of Likelihood-Based Dimension Reduction Methods

Hyemin Yoo, *Ewha Womans University*
Jae Keun Yoo, *Ewha Womans University*

To overcome the limitations of widespread dimension reduction methods such as Sliced Inverse Regression (SIR) and Sliced Average Variance Estimates (SAVE), many alternative methods have been proposed. In this paper, we compare two such methods—likelihood acquired directions (LAD) and unstructured principal fitted components (UPFC)—both of which are likelihood-based sufficient dimension reduction techniques. While LAD and UPFC share foundational similarities in their approach to dimension reduction, they diverge significantly in the specifics of their assumptions and implementation about the underlying data structure.

SAM-1-5-3

Applications of Response Dimension Reduction in Large p-Small n Problems

Minjee Kim, Jae Keun Yoo, *Ewha Womans University*

The goal of this paper is to show how multivariate regression analysis with high-dimensional responses is facilitated by the response dimension reduction. Multivariate regression, characterized by multi-dimensional response variables, is increasingly prevalent across diverse fields such as repeated measures, longitudinal studies, and functional data analysis. One of the key challenges in analyzing such data is managing the response dimensions, which can complicate the analysis due to an exponential increase in the number of parameters. Although response dimension reduction methods are developed, there is no practically useful illustration for various types of data such as so-called large p-small n data. This paper aims to fill this gap by showcasing how response dimension reduction can enhance the analysis of high-dimensional response data, thereby providing significant assistance to statistical practitioners and contributing to advancements in multiple scientific domains.

Keywords: High-dimensional data analysis, Large p-small n data, Model-based reduction, Multivariate regression, Response dimension reduction

SAM-1-5-4

Sparse Reduced Rank Regression with Covariance Matrix Estimation for Missing Multivariate Data

Nobuyoshi Kito, *Graduate School of Culture and Information Science, Doshisha University*
Shintaro Yuki, *Graduate School of Culture and Information Science, Doshisha University*
Kensuke Tanioka, *Faculty of Life and Medical Sciences, Doshisha University*
Hiroshi Yadohisa, *Faculty of Culture and Information Science, Doshisha University*

Multivariate regressions model the relationships between explanatory and multiple response variables. However, in traditional multivariate regression, each response variable is independently regressed on the explanatory variables, leading to an inaccurate estimation of the regression coefficient matrix when there is a high correlation among the response variables, ultimately reducing prediction accuracy. Additionally, as the number of explanatory variables increases, irrelevant variables may be included, leading to overfitting and consequently decreasing the prediction accuracy. We propose sparse reduced-rank regression (SRRR). This method reduces the dimensionality of the regression coefficient matrix and introduces a regularization term that treats each row as a group, improving prediction accuracy by preventing overfitting. However, when the explanatory variables contain missing data, accurately estimating the regression coefficient matrix becomes challenging, leading to decreased prediction accuracy. However, we proposed a Lasso with a high missing rate (HMLasso) within the framework of univariate regression to mitigate the loss of prediction accuracy due to missing data. This method estimates a covariance matrix from explanatory variables with missing data. This matrix is used to calculate the regression coefficients, improving the accuracy of the coefficient estimates and the prediction accuracy for the response variable. However, this method assumes a single response variable and cannot represent the correlation structure between multiple response variables. This study proposes a new method that combines covariance matrix estimation from explanatory variables with missing data and the SRRR. Specifically, we estimate the covariance matrix from explanatory variables with missing data, calculate new explanatory and response variables from the estimated matrix and apply SRRR. This approach improves the prediction accuracy for multiple response variables even when the explanatory variables contain missing data.

Keywords: Missing Data; Reduced Rank Regression; Covariance Matrix; Sparse Estimation

PARALLEL SESSIONS 同步演講場次 SPM-1

13:30 - 15:00 Saturday, December 14

SPM-1-1	**Complex Data Analysis** Organizer: CSAT Committee Chair: Johan Lim	Socrates 蘇格拉底廳
SPM-1-2	**Recent Advances in Time Series and Spatial Statistics** Organizer: Junho Yang Chair: Junho Yang	Nietzsche 尼采廳
SPM-1-3	**Advancements in Metaheuristic Algorithms and their Statistical Applications** Organizer: Ray-Bing Chen Chair: Ping-Yang Chen	Michelangelo 米開朗基羅廳
SPM-1-4	**Advanced Methods for Complex Data Analytics in the AI Era** Organizer: Ying Chen Chair: Ying Chen	Raphael 拉斐爾廳
SPM-1-5	**Recent Advances in Statistical Methods for Analyzing Time Series and Spatial Data** Organizer: IACS-ARS Committee Chair: Ci-Ren Jiang	Alexander 亞歷山大廳
SPM-1-6	**Recent Advances in Causal Inference and Applications to Biomedical Studies** Organizer: Young-Geun Choi Chair: Young-Geun Choi	Davinci 達文西廳

SPM-1-1-1

Bayesian Robust Inference of Chain Graph Models

Min Jin Ha, *Graduate School of Public Health, Yonsei University*

Chain graphs characterize conditional dependence structures of such multi-level data where variables are naturally partitioned into multiple ordered layers, consisting of both directed and undirected edges. Existing literature mostly focus on Gaussian chain graphs, which are ill-suited for non-normal distributions with heavy-tailed marginals, potentially leading to inaccurate inferences.

We propose a Bayesian robust chain graph model (RCGM) based on random transformations of marginals using Gaussian scale mixtures to account for node-level non-normality in continuous multivariate data. This flexible modeling strategy facilitates identification of conditional sign dependencies among non-normal nodes while still being able to infer conditional dependencies among normal nodes. In simulations, we demonstrate that RCGM outperforms existing Gaussian chain graph inference methods in data generated from various non-normal mechanisms. We apply our method to pharmaco-omics data to understand underlying biological processes holistically for drug response and resistance in lung cancer.

SPM-1-1-2

Optimization of Generalized Fused Lasso in Generalized Linear Models

Mineaki Ohishi, *Tohoku University*

I consider generalized linear models for grouped data. An interesting point here is to estimate parameters expressing group effects. For example, if group effects for groups 1 and 2 are equal, it is desirable to estimate their parameters to be exactly equal. One strategy to achieve this is a penalized estimation with generalized fused Lasso (GFL). In GFL estimation, differences between two parameters are shrunk based on adjacent relationships or network structure among groups, and GFL can provide estimates with cluster structure where estimates in each cluster are exactly equal. However, since GFL estimates cannot be obtained in closed form, an algorithm is very important. In this presentation, I propose a coordinate descent method to obtain GFL estimates efficiently.

SPM-1-1-3

Distribution-free model selection for longitudinal zero-inflated count data with missing responses and covariates

Chung-Wei Shen, *Department of Mathematics, National Chung Cheng University*

In many medical and social science studies, count responses with excess zeros are very common and often the primary outcome of interest. Such count responses are usually generated under some clustered correlation structures due to longitudinal observations of subjects. To model such longitudinal count data with excess zeros, the zero-inflated binomial (ZIB) models for bounded outcomes, and the zero-inflated negative binomial (ZINB) and zero-inflated poisson (ZIP) models for unbounded outcomes all are popular methods. To alleviate the effects of deviations from model assumptions, a semiparametric (or, distribution-free) weighted generalized estimating equations has been proposed to estimate model parameters when data are subject to missingness. In this article, we further explore important covariates for the response variable. Without assumptions on the data distribution, a model selection criterion based on the expected weighted quadratic loss is proposed to select an appropriate subset of covariates, especially when count responses have excess zeros and data are subject to nonmonotone missingness in both responses and covariates. To understand the selection effects of the percentages of excess zeros and missingness, we design various scenarios for covariate selection in the mean model via simulation studies and a real data example regarding the study of cardiovascular diseaseis also presented for illustration.

Keywords: generalized estimating equations, missing at random, two-component mixture models, variable selection, zero-inflation

SPM-1-2-1

Clustering of Mountain Hiking GPS-Trajectory Data

Seungyeon Back, *Hanyang University*
Man Hwi Han, *Hanyang University*
Seoncheol Park, *Hanyang University*

In this presentation, we present a clustering method to analyze the characteristics of mountain tracking patterns. The GPS-Trajectory data was obtained from the exercise app. Since the inaccuracy of GPS measurements, the data have outliers, unexpected velocities, and missing values. Therefore, we first suggest an automatic data-cleaning method for the analysis. To reflect complex spatio-temporal patterns of GPS-Trajectories, the proposed clustering method consists of a weighted average of geographical, temporal, and velocity similarities. We suggest a data-adaptive weight selection with appropriate constraints. Real data analysis with clustering results will be provided.

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2024-00339064).

Keywords: GPS-trajectory data, Spatio-temporal data, Density estimation, Clustering

SPM-1-2-2

Spectral Analysis for Continuous-Time Stationary Processes Having Scaling Properties

Tetsuya Takabatake, *Osaka University*

The (local) self-similarity property of stochastic processes is important to capture several empirically observed phenomena appearing in financial data. In this talk, I will discuss a parametric estimation problem for continuous-time stationary processes having scaling properties like the long-memory and local self-similarity properties and present our recent results related to asymptotic properties of estimators based on some divergences in the frequency domain using discretely observed data of such continuous-time stationary processes.

Keywords: Self-similarity property, Fractional Brownian motion, Fractional Ornstein-Uhlenbeck process

SPM-1-2-3

Extreme Value Analysis Using Semiparametric Spatial Zero-Inflated Models

Chun-Shu Chen, *Graduate Institute of Statistics, National Central University, Taiwan*
Chung-Wei Shen, *Department of Mathematics, National Chung Cheng University, Taiwan*
Bu-Ren Hsu, *Graduate Institute of Statistics, National Central University, Taiwan*

Spatial two-component mixture models are effective for analyzing spatially correlated data with zero inflation. To avoid biases from assuming a specific distribution for response variables, we utilize a semiparametric spatial zero-inflated model. This model presents significant computational challenges, especially with large datasets, due to the high dimensionality of latent spatial variables, complex matrix operations, and slow estimation convergence. To address these issues, we introduce a projection-based method that reduces dimensionality by projecting latent spatial variables onto a lower-dimensional space using selected basis functions. An efficient iterative algorithm is developed for parameter estimation within a generalized estimating equation framework. We determine the optimal number of basis functions via Akaike's information criterion and assess the stability of parameter estimates using the block jackknife method. This approach is validated through simulation studies and applied to Taiwan's 2016 daily rainfall data, demonstrating its practical effectiveness.

Keywords: Akaike's information criterion, Generalized estimating equations, Parameter estimation, Thin-plate splines, Zero-inflation

SPM-1-2-4

Pseudo-spectra of multivariate inhomogeneous spatial point processes

Junho Yang, *Academia Sinica*
Qi-Wen Ding, *Academia Sinica*
Joonho Shin, *Sungshin Women's Univeristy*

In this presentation, we propose a new spectral method for multivariate inhomogeneous spatial point processes. A key idea is utilizing the asymptotic behavior of the periodogram. The periodogram is an asymptotically unbiased estimator of the spectrum of a second-order stationary point process. By extending this property to the inhomogeneous case, we show that the expectation of the periodogram converges to a matrix-valued function that is Hermitian and positive definite. We call this function the pseudo-spectrum of a multivariate inhomogeneous point process. We show that the pseudo-spectrum can be interpreted in terms of the integration of the local spectrum. We derive a consistent estimator of the pseudo-periodogram through kernel smoothing and propose a data-driven bandwidth selection method. In a simulation, we show that our estimator has satisfactory finite sample properties.

Keywords: bandwidth selection, inhomogeneous spatial point processes, multivariate spatial point processes, pseudo-spectrum

SPM-1-3-1

Discrete Consensus-Based Optimization

Joong-Ho Won, *Seoul National University*

We propose Discrete Consensus-Based Optimization (DCBO), a fully discrete version of the Consensus-Based Optimization (CBO) framework. DCBO is a multi-agent method for the global optimization of possibly non-convex and non-differentiable functions. It aligns with the CBO paradigm, which promotes a consensus among agents towards a global optimum through simple stochastic dynamics amenable to rigorous mathematical analysis. Despite the promises, there has been a gap between the analysis of CBO and the actual behavior of the agents from its time-discrete implementation, as the former has focused on the system of continuous stochastic differential equations defining the model or its mean-field approximation. In particular, direct analysis of CBO-type algorithms with heterogeneous stochasticity is very challenging. DCBO distinguishes itself from these approaches in the sense that it has no continuous counterpart, thanks to the replacement of the "softmin" operator with the "hardmin" one, which is inherently discrete. Yet, it maintains the operational principles of CBO and allows for rich mathematical analysis. We present conditions, independent of the number of agents, for achieving a consensus or convergence and study the circumstances under which global optimization occurs. We test DCBO on a large number of benchmark functions to show its merits. We also demonstrate that DCBO is applicable to a diverse range of real-world problems, including neural network training, compressed sensing, and portfolio optimization, with competitive performance. Applications to optimal design of experiments are discussed. This is a joint work with Junhyeok Byeon (SNU) and Seung-Yeal Ha (SNU).

Keywords: consensus-based optimization, global optimization, interacting particle systems, stochastic particle methods, optimal design of experiments

SPM-1-3-2

Efficient Multi-Stage Design Generator for Phase II Clinical Trials Using Swarm Intelligence Optimization Techniques

Ping-Yang Chen, *Department of Statistics, Nation Taipei University*

In Phase II clinical trials, multi-stage designs offer a strategic advantage by enabling interim analyses that allow for the early termination of trials if the treatment is deemed ineffective. However, with the increasing complexity of modern clinical trials, efficiently generating designs with more stages is challenging. Traditional exhaustive search methods are computationally expensive and lack scalability beyond two or three stages. In this talk, we propose a novel approach utilizing the Particle Swarm Optimization (PSO) algorithm to generate multi-stage designs for Phase II trials. To illustrate the performance of the proposed approach, we consider two commonly used types of designs: the optimal design, which minimizes the expected sample size, and the minimax design, which minimizes the total sample size, both under pre-specified constraints on Type I and Type II errors. Our numerical results demonstrate the effectiveness of PSO in generating both two-stage and three-stage designs. Notably, PSO has discovered some optimal and minimax three-stage designs that outperform those generated by traditional exhaustive search methods. Additionally, for designs involving more than three stages, we discuss PSO parameter configurations that enable the efficient identification of optimal four-stage and five-stage designs.

Keywords: Multi-stage Design, Phase II Clinical Trial, Particle Swarm Optimization

SPM-1-3-3

Nature-inspired Metaheuristic Optimization Algorithms: Some Challenges and Proposed Solutions

Kwok Pui Choi, *Department of Statistics & Data Science/National University of Singapore*

Nature-inspired swarm-based optimization algorithms, such as Particle Swarm Optimization algorithm, are general purpose metaheuristic optimization algorithms. They are easy to implement; and require very mild assumptions on the objective functions for their application. Consequently, they are increasingly applied to tackle high-dimensional and complex optimization problems across many disciplines. Two common challenges are encountered: One is the tendency of getting stuck in local optima, thereby not able to converge to a global optimum. The other is that their performance in an optimization problem depends on a suitable choice of tuning parameters. In this talk, we present our proposed solutions to address these challenges.

SPM-1-4-1

Time series forecasting: Exploring hybrid strategies with singular spectrum analysis

Paulo Canas Rodrigues, *Federal University of Bahia*

Time series forecasting plays a key role in areas such as energy, environment, economy, and finances. Hybrid methodologies, combining the results of statistical, mathematical, and machine learning methods, have become popular for time series analysis and forecasting, as they allow researchers to compensate for the limitations of one approach with the strengths of the other and combine them into new frameworks while improving forecasting accuracy. In this class of methods, algorithms for time series forecasting are applied sequentially, i.e., the second algorithm is applied to the residuals that were not captured by the first one. In this talk, I will discuss several hybrid strategies for time series forecasting that use singular spectrum analysis, classical time series models, and recurrent neural networks, with application to several areas of research.

Keywords: Time series forecasting, Hybrid mothods, Ringular spectrum analysis, Recurrent neural networks

SPM-1-4-2

Asymmetric Multidimensional Scaling by Using Asymmetric L1 and L2 Norm

Jun Tsuchida, *Kyoto Women's University*
Hiroshi Yadohisa, *Doshisha University*

Asymmetric multidimensional scaling (AMDS) is one of the methods for visualizing the data, considering the asymmetric relationships between objects. Many AMDSs are developed by the modification of the distance, such as Euclidian distance, as the weighted distance by using different weighted terms from object i to j and from object j to i. However, it is difficult to interpretation of the distance used by these methods. In this presentation, we proposed a novel AMDS by using asymmetric distance based on a quantile norm known as an asymmetric norm. Adopting the quantile norm is expected to make the interpretation of the coordinate vector of objects easy.

Keywords: Asymmetric norm; data visualization; expectile; quantile

SPM-1-4-3

Enhancing Financial Data Management Efficiency: Case-Based Reasoning Approach

Rujira Ouncharoen, *International College of Digital Innovation, Chiang Mai University\Research Group in Mathematics and Applied Mathematics Department of Mathematics, Science, Chiang Mai University Thailand*
Thacha Lawanna, *International College of Digital Innovation, Chiang Mai University*

The comparison of various data management techniques—Data Cleansing and Preprocessing, Data Warehousing and ETL Processes, Master Data Management (MDM), and Case-Based Reasoning (CBR)—reveals insights into their effectiveness in addressing the pain point of Data Quality and Integration within the financial sector. While traditional methods excel in improving data accuracy and centralization, they often lack adaptability in dynamic environments. However, employing a Case-Based Reasoning (CBR) approach can enhance their performance significantly. CBR, with its three distinct phases—Case Retrieval for Identifying Best Practices, Adaptation of Successful Integration Techniques, and Continuous Learning and Improvement—offers superior adaptability and iterative improvements compared to static methods. Through Performance Metrics, User Satisfaction Surveys, and Comparative Analysis, it becomes evident that CBR consistently outperforms traditional methods by approximately 10-30% in efficiency. This underscores the significance of CBR in enhancing data management practices, particularly in addressing the challenges of Data Quality and Integration in finance. By incorporating CBR into existing methodologies, financial institutions can realize significant improvements in data management efficiency and effectiveness, ultimately leading to more informed decision-making and better business outcomes.

Keywords: Case-Based Reasoning, CBR, Financial , Data , Management

SPM-1-4-4

Adaptive Machine Learning Approaches Applied to Near Infrared Spectroscopic Data for the Prediction of Chemical Attributes on Intact Mangoes

Hizir Sofyan, *Universitas Syiah Kuala*
Agus A. Munawar, *Universitas Syiah Kuala*

This paper presents a comprehensive study on the predictive modeling of acidity, soluble solids content (SSC), and vitamin C in intact mangoes using Near Infrared Spectroscopy (NIRS) coupled with advanced machine learning algorithms. The primary objective of this research is to assess the prediction performances of Support Vector Machines (SVM), XG-Boost regression, and General Regression Neural Networks (GRNN) in analyzing a complex of NIRS data to predict key quality indicators in mangoes non-destructively. The study involved collecting NIRS spectra from a diverse sample of mangoes under controlled conditions, followed by the calibration and validation of models using the aforementioned machine learning techniques. Model performances were evaluated based on their accuracy, precision, and robustness, gauged through metrics such as R^2, RMSE, and RPD indexed, with achieved R2 values ranging from 0.91 to 0.98 and Ratio Prediction to Deviation (RPD) between 4.85 and 7.42. Results indicated that each algorithm has unique strengths in dealing with the spectral data, where SVM and XG-Boost showed strong potential for SSC and acidity predictions, respectively, while GRNN was notably effective in estimating vitamin C content. This research not only enhances the understanding of NIRS as a rapid, efficient technology for fruit quality assessment but also provides a basis for future applications of machine learning in the agro-food industry, aiming at quality control and assurance.

SPM-1-5-1

Additive partial linear models with autoregressive symmetric errors and its application to the hospitalizations for respiratory diseases

Shu Wei Chou-Chen, *School of Statistics, University of Costa Rica, Costa Rica\Center for Research in Pure and Applied Mathematics, University of Costa Rica, Costa Rica.*
Rodrigo A. Oliveira, *Tribunal Regional do Trabalho da 18ª Região, Goiás, Brazil.*
Irina Raicher, *Clinics Hospital of the University of São Paulo Medical School, São Paulo, Brazil\Hospital Israelita Albert Einstein, São Paulo, Brazil.*
Gilberto A. Paula, *Institute of Mathematics and Statistics, University of São Paulo, Brazil.*

Additive partial linear models with symmetric autoregressive errors of order p are proposed in this paper for modeling time series data. Specifically, we apply this model class to explain the weekly hospitalization for respiratory diseases in Sorocaba, São Paulo, Brazil, by incorporating climate and pollution as covariates, trend and seasonality. The main feature of this model class is its capability of considering a set of explanatory variables with linear and nonlinear structures, which allows, for example, to model jointly trend and seasonality of a time series with additive functions for the nonlinear explanatory variables and a predictor to accommodate discrete and linear explanatory variables. Additionally, the conditional symmetric errors allow the possibility of fitting data with high correlation order, as well as error distributions with heavier or lighter tails than the normal ones. We present the model class, and a novel iterative process is derived by combining a P-GAM type algorithm with a quasi-Newton procedure for the parameter estimation. The inferential results, diagnostic procedures, including conditional quantile residual analysis and local influence analysis for sensitivity, are discussed. Simulation studies are performed to assess finite sample properties of parametric and nonparametric estimators. Finally, the data set analysis and concluding remarks are given.

Keywords: cubic splines; cyclic splines; robust estimation; penalized likelihood; climate, hospitalization

SPM-1-5-2

Buffered Threshold Modeling in Nonlinear Time Series Analysis

Philip Leung Ho Yu, *The Education University of Hong Kong*
Wai Keung Li, *The Education University of Hong Kong*

There has been increasing interest in extending the traditional threshold time series models by incorporating a buffer zone for regime transitions. Typically, the traditional threshold models assume that structural shifts occur when a variable crosses a specific threshold. However, regime switching can often be delayed – a phenomenon known as hysteresis – across various fields such as economics, engineering, and biology. This indicates that the regime-switching mechanism may benefit from a buffer zone, allowing for delayed transitions when the variable is within that zone. Zhu, Yu and Li (2014) and Li, Guan, Li and Yu (2015) were the first to introduce the buffering concept into the regime-switching framework of threshold models. Since their work, multiple models have been developed in this area. In this talk, I will review recent advancements in buffered time series models and explore potential opportunities and challenges for future research.

Keywords: Buffer zone; Threshold models; Nonlinear Time Series; Regime switching; Hysteresis

SPM-1-5-3

Dynamic variable selection based on Kalman filter

Wataru Yoshida, *Joint Graduate School of Mathematics for Innovation, Kyushu University*
Kei Hirose, *Institute of Mathematics for Industry, Kyushu University*

The linear-Gaussian state space model is an effective model for time series analysis. For example, the dynamic regression model, one of the linear-Gaussian state space models, is a regression model that allows the coefficients to vary over time; it achieves high accuracy in electricity demand forecasting in our experiments by expressing seasonal variations. However, when considering the time variation of the model, there are numerous combinations in variable selection, and it is difficult to verify all of them. To alleviate this issue, we propose a new dynamic variable selection method based on the Kalman filter, an efficient coefficient estimation algorithm. Our method performs sparse estimation allowing non-zero coefficients to vary over time. A significant feature of this method is that the regularization parameter is optimized at each time point during the estimation. This approach makes it possible to adjust the strength of the coefficient shrinkage in accordance with the time variation of the model sparsity. Another advantage is computation speed; the computational speed of our method is comparable to that of the Kalman filter. The actual estimation accuracy and computation time will be verified via numerical simulations.

Keywords: time series; Kalman filter; sparse estimation; dynamic variable selection

SPM-1-5-4

Spatiotemporal Modeling in the Presence of Structural Change

Regina Macarangal Tresvalles, *De La Salle University*

This paper intends to use a robust estimation procedure for a spatiotemporal model with structural change. This estimation procedure was originally proposed by Bastero and Barrios in 2011. The result uses a spatiotemporal model that incorporates the forward search algorithm and maximum likelihood estimation into the backfitting framework. The forward search algorithm is used to filter the effect of short-term structural change in the estimation of covariate and spatial parameters. The model is used on the spread of the coronavirus disease 2019 (COVID-19), showing its suitability in the pandemic experience in the Philippines. The result shows that the method that uses the effect of structural change offers a good model fit especially in the COVID-19 experience in the Philippines where structural change is encountered at different spaces and times. Simulation studies further illustrate the potential of this model to produce robust estimates, especially in the presence of structural change.

Keywords: spatio-temporal model, structural change, forward search, COVID-19

SPM-1-6-1

Robust and Flexible High-Dimensional Causal Mediation Model for DNA Methylation Studies

An-Shun Tai, *Department of Statistics, National Cheng Kung University*

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

Keywords: Causal inference, High-dimensional mediation analysis, Multiple robustness, Spike-and-slab priors, DNA methylation

SPM-1-6-2

Hypothesis test for causal mediation of semicompeting risks under copula, frailty and multistate models

Jih-Chang Yu, *Department of Statistics, National Taipei University*
Yen-Tsung Huang, *Institute of Statistical Science, Academia Sinica*

Semicompeting risks problem concerns two time-to-event outcomes where the intermediate outcome may be censored by the terminal outcome but not vice versa. Huang' s 2010 has shown that the semicompeting risks can be formulated as a mediation model where a direct effect (DE), the effect of an exposure on the terminal outcome not through the intermediate outcome, and an indirect effect (IE), the effect on the terminal outcome mediated by the intermediate outcome, are studied. In this article, we propose testing procedures to evaluate the DE and IE under three classic semicompeting risks models: Clayton copula model, gamma frailty model, and multistate model. We study the correspondence of the DE and IE with the model parameters and establish testing rules for the two effects under the three models. We use the U-statistic approach for the Clayton copula model and nonparametric maximum likelihood estimation for the multistate and gamma frailty models for statistical inference. The simulation study shows that among the three models, the Clayton copula model attains the best statistical power if the model assumption holds, but has the potential bias caused by model misspecification; the gamma frailty model is the most robust model by sacrificing the efficiency; the multistate model balances the efficiency and robustness. We apply the proposed method to a hepatitis study, and the aforementioned models unanimously suggest that both hepatitis B and C lead to a higher incidence of liver cancer by increasing liver cirrhosis incidence.

Keywords: copula model, frailty model, multistate model, causal mediation analysis, semicompeting risks

SPM-1-6-3

Detecting differential outlying methylation regions

Shih-Kai Chu, *National Taipei University*

A rare DNA methylation variant exhibits dramatically different methylation levels in a small group of individuals compared to the rest of the studied population. Recent studies show that rare DNA methylation variants are linked to several human diseases, such as cancer and congenital anomalies. Existing methods to compare rare DNA methylation variants between two groups of samples were mostly single-variant association tests. In this study, we introduced a procedure for multi-variant association analysis. The regional outlying methylation difference between two groups was measured by aggregating the group differences for each variant in that region. The performance of the proposed method was supported by a simulation study, and we have applied the method to multiple real datasets.

Keywords: Epigenetics, Region-based association analysis, Outlying DNA methylation variation

PARALLEL SESSIONS 同步演講場次 SPM-2

15:20 - 16:50 Saturday, December 14

SPM-2-1	**Novel Approaches in Bayesian and Empirical Bayes Methods** Organizer: Ray-Bing Chen Chair: Ray-Bing Chen	Socrates 蘇格拉底廳
SPM-2-2	**Causal Inference and Its Application on Statistics** Organizer: Sheng-Hsuan Lin Chair: Sheng-Hsuan Lin	Nietzsche 尼采廳
SPM-2-3	**Innovative Statistical and Machine Learning Techniques for Enhanced Prediction and Optimization in Insurance, Finance, and Criminology** Organizer: Joanna Wang Chair: Joanna Wang	Michelangelo 米開朗基羅廳
SPM-2-4	**Statistical and Mathematical Approaches in High Throughput Health Data Analytics** Organizer: Tzu-Pin Lu Chair: Yen-Chen Anne Feng	Raphael 拉斐爾廳
SPM-2-5	**Causal Mediation analysis and Statistical Inference for Latent Class Models** Organizer: IACS-ARS Committee Chair: An-Shun Tai	Alexander 亞歷山大廳
SPM-2-6	**Recent Advances in Symbolic Data Analysis and Computational Algorithms for Statistical Inference** Organizer: IACS-ARS Committee Chair: Chen-An Tsai	Davinci 達文西廳

SPM-2-1-1

A Bayesian Method for Multinomial Probit Model

Keunbaik Lee, *Sungkyunkwan University*
Chanmain Kim, *Sungkyunkwan University*

The independence of irrelevant alternatives (IIA) property states that the ratio of any two choice probabilities in a set of alternatives is independent of the presence or absence of other alternatives. In the modeling of multinomial data, the IIA is not feasible. In this situation, the multinomial probit (MNP) model is a type of discrete choice model that is commonly used. Due to the identifiability problem and the positive-definiteness constraint, modeling the covariance matrix in the MNP is difficult. All existing methods use unidentifiable parameters in the covariance matrix to solve the unidentifiability problem and improve the rate of convergence of a data augmentation algorithm. These methods also use the inverse Wishart distribution, which is frequently insufficient \citep{barnard2000modeling}. We employed variance-correlation decomposition to decompose the identifiable covariance matrix into standard deviations and a correlation matrix instead of using the unidentifiable covariance matrix. Hypersphere decomposition was also used to decompose the correlation matrix. Thus, the estimated covariance matrix satisfied the positive definiteness constraint. The performance of our proposed model was illustrated using a detergent dataset from market research.

Keywords: covariance matrix, identifiability, independence of irrelevant alternatives (IIA), positivedefiniteness

SPM-2-1-2

Bayesian structure selection based on mixed-effects model with categorical responses

Chi-Hsiang Chu, *Institute of Statistics, National University of Kaohsiung*
Ray-Bing Chen, *Department of Statistics and Institute of Data Science, National Cheng Kung University*

In this work, we study the Bayesian structure selection approaches for mixed-effect models with categorical responses, where these kinds of data structure appeared frequently in many medical studies. As the mixed-effect models for categorical responses are commonly used in the medical data analysis, it is important to be able to select the significant explanatory variables correctly and efficiently. To deal with the categorical response variable, a multinomial probit model is adopted by introducing the proper latent variables, and we have the row sparsity assumption for the coefficient matrix, because each row corresponds to the effect of one variable. Then an indicator-based Bayesian variable selection approach is considered and the corresponding MCMC algorithm is proposed for generating the posterior samples for future inference. Here simulation studies and a medical data set will be used for demonstrating the performance of the proposed Bayesian selection approach.

Keywords: Component-wise algorithm, group structure, Markov chain Monte Carlo, mixed-effect model, multi-task learning

SPM-2-1-3

Post Empirical Bayes Regression

Yu-Chang Chen, *National Taiwan University*
Shuo-Chieh Huang, *Rutgers University*
Shen-Hsun Liao, *National Taiwan University*
Sheng-Kai Chang, *National Taiwan University*

Empirical Bayes (EB) methods are widely utilized in economics for estimating individual and group-level fixed effects across diverse contexts, including teacher value-added, hospital qualities, and neighborhood effects. While estimates generated by EB are often incorporated into downstream statistical analyses like regression models, the econometric properties of post-EB regression have not been justified. This paper addresses this issue through two key contributions. First, we introduce a unified framework for two-step EB methods that applies to both linear and non-linear models, offering frequentist properties and assessing their robustness against model misspecification. Second, we undertake a critical evaluation of the commonly used two-step EB methods in existing empirical research. Our analysis demonstrates that existing post-EB regression implementation, without proper adjustments, can introduce systematic bias, particularly in non-linear models.

Keywords: Empirical Bayes, Measurement Errors, Deconvolution

SPM-2-1-4

Natural Gradient Variational Bayes without Fisher Matrix Analytic Calculation and Its Inversion

Minh Ngoc Tran, *The University of Sydney, Australia*
A. Godichon-Baggioni, *Sorbonne University, France*
Duy Nguyen, *Marist College, United States*

This paper introduces a method for efficiently approximating the inverse of the Fisher information matrix, a crucial step in achieving effective variational Bayes inference. A notable aspect of our approach is the avoidance of analytically computing the Fisher information matrix and its explicit inversion. Instead, we introduce an iterative procedure for generating a sequence of matrices that converge to the inverse of Fisher information. The natural gradient variational Bayes algorithm without analytic expression of the Fisher matrix and its inversion is provably convergent and achieves a convergence rate of order O(log s/s), with s the number of iterations. We also obtain a central limit theorem for the iterates. Implementation of our method does not require storage of large matrices, and achieves a linear complexity in the number of variational parameters. Our algorithm exhibits versatility, making it applicable across a diverse array of variational Bayes domains, including Gaussian approximation and normalizing flow Variational Bayes. We offer a range of numerical examples to demonstrate the efficiency and reliability of the proposed variational Bayes method.

Keywords: Bayesian computation, Stochastic gradient descent, Bayesian neural network, Fisher information

SPM-2-2-1

Causal Mechanisms in Biomedical Science

Etsuji Suzuki, *Okayama University*

For several decades, the counterfactual model and the sufficient cause model have shaped our understanding of causation in biomedical science and, more recently, the link between these two models has enabled us to obtain a deeper understanding of causality. In this presentation, I provide a brief overview of these fundamental causal models using a simple example. The counterfactual model focuses on one particular cause or intervention and gives an account of the various effects of that cause. By contrast, the sufficient cause model considers sets of actions, events or states of nature which together inevitably bring about the outcome under consideration. In other words, the counterfactual framework addresses the question "what if?", while the sufficient cause framework addresses the question "why does it happen?" Although these two models are distinct and address different causal questions, they are closely related and used to elucidate the same cause-effect relationships. Importantly, the sufficient cause model makes clear that causation is a multifactorial phenomenon, and it is a "finer" model than the counterfactual model; an individual is of one and only one response type in the counterfactual framework, whereas an individual may be at risk of none, one, or several sufficient causes. By taking into account the potential completion time of each sufficient cause, I discuss the two types of etiologic fraction: the accelerating etiologic proportion and the total etiologic proportion. Although the differences between them might be subtle, they are closely related to the definition of causality. Therefore, it is important to clarify which measures are used on each occasion. Understanding the link between the two causal models can provide greater insight into causality and can facilitate the use of each model in appropriate contexts, highlighting their respective strengths.

Keywords: causal inference; counterfactual model, etiologic fraction; mechanisms; sufficient cause model

SPM-2-2-2

Nonparametric Bayesian Adjustment of Unmeasured Confounders in Cox Proportional Hazards Models

Shunichiro Orihara, *Tokyo Medical University*
Shonosuke Sugasawa, *Keio University*
Tomohiro Ohigashi, *Tokyo University of Science*
Tomoyuki Nakagawa, *Meisei University*
Masataka Taguri, *Tokyo Medical University*

In observational studies, unmeasured confounders present a crucial challenge in accurately estimating desired causal effects. To calculate the hazard ratio (HR) in Cox proportional hazard models for time-to-event outcomes, two-stage residual inclusion and limited information maximum likelihood are typically employed. However, these methods are known to entail difficulty in terms of potential bias of HR estimates and parameter identification. This study introduces a novel nonparametric Bayesian method designed to estimate an unbiased HR, addressing concerns that previous research methods have had. Our proposed method consists of two phases: 1) detecting clusters based on the likelihood of the exposure and outcome variables, and 2) estimating the hazard ratio within each cluster. Although it is implicitly assumed that unmeasured confounders affect outcomes through cluster effects, our algorithm is well-suited for such data structures. The proposed Bayesian estimator has good performance compared with some competitors.

Keywords: general Bayes, invalid instrumental variable, Mendelian randomization, UK Biobank, weak instrumental variable

SPM-2-2-3

Prognostic covariate adjustment with treatment interaction in the presence of historical data for multi-arm randomized controlled trials

Kenichi Hayashi, *Keio University*
Yuka Miyake, *Keio University*

Clinical trials are experimental studies conducted on humans to evaluate the effectiveness and safety of new treatments, medications, and medical devices. In randomized clinical trials, patients are randomly allocated into two or more arms. In this paper, we consider how to estimate the effect of a new treatment in a randomized clinical trial.

In recent years, there has been growing interest in the use of historical data as a means of reducing the sample size required for trials, leading to reduced costs. Prognostic covariate adjustment uses a model trained on historical data (called a prognostic model) to output a covariate (referred to as a prognostic score) that is added to the explanatory variables in a regression model (Schuler et al., 2021). There are basically three models of prognostic covariate adjustment. The first is the PROCOVATM model, which includes only the treatment and the prognostic score as explanatory variables. The second is a natural extension in which the interaction between the treatment and the prognostic score is also considered as an explanatory variable. The third is regarded as a hybrid of the ANCOVA (analysis of covariance) II and the PROCOVA models. The asymptotic variance of the treatment effect estimators has previously been derived using semiparametric theory for the estimation of treatment effects in two groups. However, for the case with two arms, models including interaction terms are likely to be less efficient.

The novel contributions of our study are twofold. First, we derive the properties of prognostic covariate adjustment in multi-arm trials using semiparametric theory. When conducting regression analysis using a linear model, there is a possibility of misspecification of the conditional model. However, even in cases of model misspecification, it is possible to construct a consistent estimator by applying semiparametric theory. Second, through simulation experiments, we compare the models with prognostic covariate adjustment, raw covariate adjustment, and without any adjustment.

Keywords: Clinical Trials; Covariate adjustment; Historical data; Prognostic score; Treatment effects

SPM-2-2-4

A stratified longitudinal targeted maximum likelihood estimator for time-varying treatment effects

Masataka Taguri, *Department of Health Data Science, Tokyo Medical University*
Toru Shirakawa, *Division of Public Health, Department of Social Medicine, Osaka University Graduate School of Medicine\Center for Targeted Machine Learning and Causal Inference, University of California, Berkeley*
Kazuharu Harada, *Department of Health Data Science, Tokyo Medical University*

In longitudinal studies involving time-varying treatments, time-dependent confounding can occur if the time-dependent covariates are influenced by past treatments. Multiply robust estimators, which are based on augmented inverse probability weighting (AIPW) estimating equations, have been proposed to protect against model misspecification and to attain asymptotic efficiency (Bang and Robins, 2005; Rotnitzky et al., 2017; Tran et al., 2019). However, these estimators may have large variances when there is significant variability in inverse probability weights of treatments. To address this issue, we propose a stratified targeted maximum likelihood estimator based on the stratification of inverse probability weights. For the point treatment, our proposed method with parametric models for the initial nuisance estimates is closely related to the regression adjustment estimator after the propensity score stratification (Lunceford and Davidian, 2004). We show that our proposed estimator is equivalent to an AIPW estimator using the strata-specific coarsened weights. We will present the results of simulations comparing the proposed estimator with existing estimators.

Keywords: AIPW estimator, causal inference, propensity score stratification, targeted maximum likelihood estimator, time-varying treatment

SPM-2-3-1

Claim prediction and premium pricing for telematics auto-insurance data using Poisson regression with lasso regularisation

Jennifer So-kuen Chan, *University of Sydney*

We leverage telematics data on driving behavior variables to assess driver risk and predict future insurance claims in a case study utilising a representative telematics sample. In the study, we aim to categorise drivers according to their driving habits and establish premiums that accurately reflect their driving risk. To accomplish our goal, we employ the two-stage Poisson model, the Poisson mixture model, and the Zero-Inflated Poisson model to analyse the telematics data. These models are further enhanced by incorporating regularisation techniques such as lasso, adaptive lasso, elastic net, and adaptive elastic net. Our empirical findings demonstrate that the Poisson mixture model with the adaptive lasso regularisation outperforms other models. Based on predicted claim frequencies and drivers' risk groups, we introduce a novel usage-based experience rating premium pricing method. This method enables more frequent premium updates based on recent driving behaviour, providing instant rewards and incentivising responsible driving practices. Consequently, it helps to alleviate cross-subsidization among risky drivers and improves the accuracy of loss reserving for auto insurance companies.

SPM-2-3-2

Temporal Multivariate Density Network for Portfolio Optimization

Fong Lam, *Discipline of Business Analytics, The University of Sydney*

This study delves into recent developments in deep learning and statistical analysis, introducing a new framework based on a Multivariate Density Network (DN) model for predicting asset price percentage returns and their associated uncertainties. Additionally, it assesses the performance of deep neural networks, with a focus on Attention-based Long Short-Term Memory (ALSTM) models, in forecasting the returns and variance of financial time series. The methodology incorporates the predicted mean returns and the variance-covariance matrix into the classical Markowitz mean-variance (MV) optimization framework for portfolio construction. A comparative analysis with portfolios built using sample mean returns and sample variance-covariance matrices shows that portfolios constructed with predicted values generally achieve better Sharpe scores during the testing period, highlighting the superiority of this approach over traditional methods.

Keywords: Density Network, Attention-based Long Short-Term Memory, Portfolio optimization , Sharpe scores

SPM-2-3-3

Coherent Estimation And Criminal Justice Program Evaluation In Hierarchical Time Series

Thomas Fung, *Macquarie University*
Joanna Wang, *University of Technology, Sydney*

Crime time series data can often be naturally disaggregated based on various attributes of interest, such as crime type or geographical location. When modelling this type of data, the current practice in crime science is to model each series at the most disaggregated level, as it helps to identify more subtle changes. However, authorities and stakeholders often focus on the bigger picture, leading researchers to either simply sum the fitted value series or model the aggregated series independently. This practice often leads to poorer performance at the higher levels of aggregation as the most disaggregated series typically exhibit a high degree of volatility, while the most aggregated series tends to be smoother and less noisy. In this presentation, we will demonstrate how the hierarchical and grouped time series structure can be utilised to provide "coherent" estimates for all disaggregate and aggregate series while also "reconciling" them to enhance forecast and criminal justice program evaluation by using all the available information. We will utilise US crime data alongside the COVID lockdown as the intervention effect for illustrative purposes.

Keywords: Time Series Analysis, Crime Statistics, Hierarchical Time Series, Estimation, COVID

SPM-2-3-4

Exploring model performance with varied intervention dates in an interrupted time series analysis of crime data

Joanna Wang, *University of Technology Sydney*
Yazhen Zhu, *University of Technology Sydney*
Thomas Fung, *Macquarie University*

The objective of this research was to identify a more flexible modelling approach for evaluating interventions, addressing the sensitivity of results to the placement of the intervention. We applied both linear and Generalized Additive Models (GAM) to interrupted time series data, revealing that parameter estimates are highly sensitive regardless of the chosen modeling technique. Additionally, we explored various measures, including cumulative effects, to assess the robustness of our findings. Through a combination of simulation studies and empirical applications, we demonstrate that the optimal placement of an intervention is critical and should be determined on a case-by-case basis. Our study provides recommendations for practitioners on carefully considering the placement of interventions to ensure reliable and valid outcomes.

Keywords: GAM, ITS, evaluation

SPM-2-4-1

Multimodal approaches to understand genetic predisposition and functional consequences of complex diseases

Amrita Chattopadhyay, *Institute of Epidemiology and Preventive Medicine, National Taiwan University*

Complex diseases are characterized by a complex etiology and pathophysiology, which makes studying and treating them challenging. Techniques such as genome wide association study (GWAS) coupled with genotype imputation, to enhance the chip marker density, has proven to be effective tools in identifying genetic variants that are associated with complex diseases. Furthermore, polygenic risk scores (PRS) allows quantification of individuals' genetic predisposition to disease by utilizing significant associations from GWAS, and are used for risk stratification and clinically useful applications. In recent decades, the field of biobanking has undergone significant development comprising of human biological specimens, combined with clinical and demographic data, thereby playing an important role in the advances of health research. In this study, utilizing UK Biobank data, a multimodal approach was applied to identify candidate pathways and biomarkers for predicting frailty syndrome in individuals. Frailty, a prevalent clinical syndrome in aging adults, is characterized by poor health outcomes, represented via a standardized frailty-phenotype (FP), and Frailty Index (FI). Genotype, clinical and demographic data of subjects (aged 60-73 years) from UK Biobank were utilized. FP was defined by Fried's criteria. FI was calculated using electronic-health-records. Genome-wide-association-studies (GWAS) were conducted and polygenic-risk-scores (PRS) were calculated for both FP and FI. Functional analysis provided interpretations of underlying biology. Finally, machine-learning (ML) models were trained using clinical, demographic and PRS towards identifying frail from non-frail individuals. Several significantly associated loci were identified accounting for 12% heritability, some of which were known associations for body-mass-index, coronary diseases, cholesterol-levels, and longevity, while the rest were novel. The findings suggest frailty as a highly polygenic-trait, enriched in cholesterol-remodeling and metabolism and to be genetically associated with cognitive abilities. ML models utilizing FP and FI + PRS were established that identified frailty-syndrome patients with high accuracy.

Keywords: Genome wide association study, UK Biobank, Frailty syndrome, Polygenic risk score, Machine learning prediction model

SPM-2-4-2

Exploring Genomic Predictions with Common and Rare Variants for Complex Diseases Using Deep Learning: A Simulation Study

Sing-Wen Chen, *Division of Biostatistics and Data Science, Institute of Epidemiology and Preventive Medicine, College of Public Health, National Taiwan University*
Yen-Chen Anne Feng, *Institute of Health Data Analytics and Statistics, College of Public Health, National Taiwan University*

In recent years, Genome-Wide Association Studies (GWAS) have uncovered numerous genetic risk loci implicated in complex traits and disorders, with a majority of these studies focusing on common variants. Although rare variants have a lower minor allele frequency in the human genome, they are enriched in coding regions and have been shown to carry significant pathogenic potential in certain diseases. Polygenic Risk Score (PRS) is one of the most commonly used metrics for assessing individual genetic liability to specific phenotypes. However, PRS predominantly evaluates linear additive effects. Evidence from some studies suggests that complex diseases may be associated with non-linear or interaction effects between genetic variants, relationships that are challenging to elucidate using traditional GWAS and PRS methodologies. Deep learning is increasingly being applied to genetic research, with neural network models offering advantages over linear models in uncovering the non-linear relationships between genetic variation and complex diseases, potentially leading to improved predictive capabilities. In this study, we designed simulations for various genetic architectures of complex traits, particularly those involving gene-gene interactions. We incorporated the effects of both common variants and deleterious rare exonic variants for prediction modeling with neural networks. Finally, we compared the predictive performance of neural networks against traditional PRS approaches.

Keywords: genomic prediction; complex diseases; deep learning; rare variants; non-linear effect

SPM-2-4-3

Uncovering Health Insights: Using Clustering Analysis to Identify Digital Biomarkers from Wearable Device Data

Charlotte Wang, *Institute of Health Data Analytics and Statistics, College of Public Health, National Taiwan University, Taipei, Taiwan (ROC)\Master of Public Health Program, College of Public Health, National Taiwan University, Taipei, Taiwan (ROC)*
Ya-Ting Liang, *Institute of Epidemiology and Preventive Medicine, College of Public Health, National Taiwan University, Taipei, Taiwan (ROC)*

Numerous studies have investigated the association between physical activity and health outcomes. While technological advances have made collecting physical activity data through wearable devices increasingly convenient and prevalent, analyzing such data remains challenging due to variability in daily activity patterns and measurement errors. In this context, the development of meaningful digital biomarkers from wearable device data for health-related studies presents an intriguing and novel research area. Previous studies have primarily used summary statistics of activity counts as digital biomarkers, which offer limited insights into specific activity patterns. To address this limitation, we propose an elastic-based motif clustering algorithm to identify specific activity patterns (motifs) in free-living physical activity data. We utilize functional principal component analysis (FPCA) to construct digital biomarkers from these motifs. Our method, applied to two mental health studies, demonstrates the potential to discover motifs in free-living physical activity data. The digital biomarkers derived from these motifs are associated with diseases, indicating a significant relationship between specific activity patterns and disease outcomes. Moreover, these biomarkers serve as reliable features for predictive models, enhancing the sensitivity of patient classification. In conclusion, our method effectively identifies motifs in free-living physical activity data. The subsequent application of digital biomarkers derived from these motifs can advance personalized health assessment and disease detection, offering a promising future for healthcare.

Keywords: cluster analysis, elastic distance, free-living physical activity, motif discovery, wearable device

SPM-2-4-4

On the Asymptotic Properties of Product Principal Component Analysis under the High Dimension Settings

Chi Chun Yeh, *Institute of Health Data Analytics and Statistics, National Taiwan University*
Hung Hung, *Institute of Health Data Analytics and Statistics, National Taiwan University*

The eigenvalue ordering is an important issue in Principal component analysis (PCA). In Hung and Huang, Product-PCA (PPCA) was proposed and shown to have the robustness of the eigenvalue ordering without suffering from asymptotic loss when n $\to\infty$ and p remain fixed. However, it is well-known that the eigenvalues of PCA may have bias under high-dimensional settings, where n$\to\infty$, p$\to\infty$ and c$\to p/n$. In this article, we investigate the asymptotic behavior of the eigenvalue of PPCA under high-dimensional settings. We also prove the robustness of PPCA under Johnstone's simple spike model.This gives the picture that PPCA would be a more appropriate method than PCA when the dimension is large enough compared to the sample size.

Keywords: PCA, leading eigenvalue, simple spiked model, robustness, random matrix theory

SPM-2-5-1

An Analytic Approach to Causal Mediation Analysis for Binary Outcomes with Asymmetric Binary Models

Yuji Tsubota, *Graduate School of Human Sciences, Osaka University*
Michio Yamamoto, *Graduate School of Human Sciences, Osaka University\RIKEN AIP , Data Science and AI Innovation Research Promotion Center, Shiga University*

Mediation analysis is a popular tool for researchers who try to investigate the mechanism of the effect of a treatment on an outcome of interest. Due to the recent expansion of causal mediation analysis literature, a wide range of methods are available for such researchers. Among those existing methods, the so-called analytic approaches to causal mediation analyses are popular for their simplicity. For binary outcomes, the existing literature on analytic approaches to causal mediation analyses often utilizes logistic and probit analyses, which implicitly assume the symmetric shape of the success probability curves. This symmetry assumption is not always plausible in real data analysis, and in such cases, symmetric models may lead to biased results. To address this issue and extend the existing literature, we propose a causal mediation analysis framework in which we utilize the complementary log-log model, a binary generalized linear model with asymmetric success probability curves. We define the total effect, the controlled direct effect, the natural direct effect, and the natural indirect effect in our framework with the complementary log-log model. We discuss the conditions in which these causal effects can be estimated with observed data. The closed-form analytic expressions of our causal effects can be derived under appropriate parametric modeling assumptions, allowing us to estimate the effects by simple regression analyses and leading to a more straightforward interpretation of the effects. To illustrate our proposed methodology, we apply it to a real-world dataset for which the complementary log-log model showed a better fit than logistic and probit models. We also assess the estimation accuracy of our causal effects through numerical simulations.

Keywords: Mediation Analysis, Natural Effects, Binary Outcomes, Complementary log-log models, Causal Inference

SPM-2-5-2

On the Limiting Spectral Distributions of Stochastic Block Models

May-Ru Chen, *National Sun Yat-sen University*
Giap Van Su, *National Sun Yat-sen University and Thai Nguyen University of Education*

The stochastic block model (SBM) is an extension of the Erdős-Rényi graph by dividing nodes into K subsets, known as blocks or communities. Let $\tilde{A}_N$ be an N × N normalized adjacency matrix of an SBM with K blocks of any sizes, and let $\mu_{\tilde{A}_N}$ be the empirical spectral density of $\tilde{A}_N$.
In this talk, we first showed that if the connecting probabilities between nodes of different blocks are zero, then $\lim N\to\infty \mu_{\tilde{A}_N} = \mu$ exists almost surely, and we gave the explicit formulas for μ and its Stieltjes transform, respectively. Second, we showed under a suitable conditions $\mu_{\tilde{A}_N}$ converges both in probability and expectation.

Keywords: Erdős–Rényi graph, empirical spectral density, semicircle law, stochastic block models

SPM-2-5-3

Enhancing solution feasibility in conditionally dependent latent class models using alternative GEE methods

Zong-Lin Lin, *National Yang Ming Chiao Tung University*

Latent class models (LCM) often assume conditional independence among item responses; however, this assumption is often violated in practice, leading to estimation bias and poor model fit. To address this issue, we apply methodologies for repeated measures data using generalized estimating equations (GEE) to handle conditional dependencies in LCM. Our approach uses two quasiscore functions: one for estimating the model parameters and another for estimating pairwise covariances between item responses given the latent class. These covariances, treated as nuisance parameters, are estimated using conditional odds ratios. While our method is similar to that of Reboussin et al. (2008), we provide additional details on working covariance, parameter estimation, and asymptotic distribution. A significant improvement is our strategy for resolving convergence issues with the Fisher scoring algorithm by using the generalized method of moments (GMM). This alternative method effectively addresses the convergence problems and yields stable estimates.

Keywords: latent class models, conditional dependence, correlated binary data, generalized estimating equations, generalized method of moments

SPM-2-5-4

Weighting in Unit-level Model to Balance Covariates in Causal Inference for RCT

Kawsar Ahmed, *School of Mathematics and Statistics, Central South University*
Hong Wang, *School of Mathematics and Statistics, Central South University*

Achieving covariate balance has become increasingly important in observational study. Nonetheless, estimation techniques continue to appear as challenges in randomized controlled trials (RCTs). This situation primarily occurs due to the presence of non-accessible confounding variables. We aid in solving a related problem through our unit-level method to determine causal effects. Additionally, the transformation of OLS is explored in a unique model to optimize its broad application for our multiunit settings. In the random allocation, this study adopts a design for defining the average treatment effect, which is challenging owing to the complexities of adjustment. Our research first presents a unit-level imputed model that precisely aligns and provides substantial adjustments for regression fitting in potential outcomes. We discuss the phases of propensity score design to identify suitable covariate balancing strategies for the random treatment arm (RTA). Our investigation offers an asymptotic property for both single-unit and multi-unit levels. We deploy weights for both large and modest sample datasets to strengthen the association between regression modeling and causal consequences. Simulation experiments demonstrate the enhanced efficacy of the suggested models and facilitate the comparison of optimal balancing technique selection. Our unit-level model findings finally resolve the possibility of choosing the best strategy for covariate matching and balancing in the treatment and control arm. This study outlines the covariate balance for RTC, offering the statistical learning community and causal inference audience to earn potential methodological advancements.

SPM-2-6-2

A visualization of aggregated symbolic data by scoring categorical variables

Nobuo Shimizu, *The Institute of Statistical Mathematics*
Junji Nakano, *Chuo University*

We are often interested in comparing meaningful groups of individuals, where each individual is described by observations of continuous and categorical variables. To summarize each group, we use the number of individuals and the first and second order moments of continuous variables and dummy variables for categorical variables. We call such statistics as aggregated symbolic data (ASD).
As we simplify ASD more for intuitive understanding and visualization, we hope to treat continuous and categorical variables equally in the simplification by defining appropriate scores for categorical values. We use the method of multiple correspondence analysis to determine scores for categorical values. An example about the real estate data in Japan is analyzed and visualized by the proposed method.

Keywords: categorical data analysis, multiple correspondence analysis, data visualization

SPM-2-6-3

Reinforcement Learning for Adaptive MCMC

Wilson Chen, *The University of Sydney*
Congye Wang, *Newcastle University*
Heishiro Kanagawa, *Newcastle University*
Chris Oates, *Newcastle University*

An informal observation, made by several authors, is that the adaptive design of a Markov transition kernel has the flavour of a reinforcement learning task. Yet, to-date it has remained unclear how to exploit modern reinforcement learning technologies for adaptive MCMC. The aim of this work is to set out a general framework, called Reinforcement Learning Metropolis-Hastings, that is theoretically supported and empirically validated. Our principal focus is on learning fast-mixing Metropolis-Hastings transition kernels, which we cast as deterministic policies and optimise via a policy gradient. Control of the learning rate provably ensures conditions for ergodicity are satisfied. The methodology is used to construct a gradient-free sampler that outperforms a popular gradient-free adaptive Metropolis-Hastings algorithm on approximately 90% of tasks in the PosteriorDB benchmark.

Keywords: Metropolis-Hastings, Reinforcement Learning, MCMC, Bayesian, Policy Gradient

STATISTICAL COMPUTING AND METHODS FOR COMPLEX DATA

國立臺灣大學

National Taiwan University