IASC-ARS INTERIM CONFERENCE 2022

THE INTERPLAY BETWEEN STATISTICAL COMPUTING AND ARTIFICIAL INTELLIGENCE

12-13 DEC

Book of Abstracts

香港教育大學
The Education University of Hong Kong

Table of Contents

Welcoming Messages

Welcoming Message from the President of EdUHK

Prof. Stephen Yan-Leung Cheung

President
The Education University of Hong Kong

On behalf of The Education University of Hong Kong, I welcome you all to the IASC-ARS Interim Conference 2022.

The theme of the Conference is "The Interplay between Statistical Computing and AI". There is no doubt that statistical computing plays an important role in data science, machine learning, and artificial intelligence (AI). In recent years, AI has become popular in the education sector in Hong Kong. In his 2022 Policy Address, the Chief Executive remarked that more learning elements of I&T will be incorporated in the curriculum, with the aim of at least 75% of publicly-funded schools implementing enriched coding education at the upper primary level and introducing I&T elements such as AI in the junior secondary curriculum by the 2024/25 school year. In view of this, the University launched the first "Bachelor of Science (Honours) Degree in AI and Educational Technology" programme in Hong Kong this year in order to nurture more talents in AI and Educational Technology.

I would like to take this opportunity to express my sincere gratitude to the Asia Regional Section of the International Association for Statistical Computing (IASC-ARS) and Mr. and Mrs. Lam Kin Research Fund for AI in Educational and Financial Technologies. Without their full support and sponsorship, this Conference would not have been possible. Also, I would like to express my great thanks to the Honorary Advisors, Organising Committee, Scientific Programme Committee, staff, and student helpers for their hard work in organising this Conference. I wish all of you a memorable and rewarding experience at the IASC-ARS Interim Conference 2022. Thank you!

Welcoming Message from the Chairperson of IASC-ARS

Prof. Ray-Bing Chen

Chairperson of the Asia Regional Section of the International Association for Statistical Computing (IASC-ARS)

Professor
Department of Statistics
National Cheng Kung University

On behalf of the Asia Regional Section of the International Association for Statistical Computing (IASC-ARS), I am honoured and delighted to welcome you to the IASC-ARS Interim Conference 2022.

IASC-ARS established in 1993 and has grown bigger and stronger thanks to the efforts of all members of the ARS. Since its establishment, IASC-ARS has been undergoing continuous development to promote regional cooperation in organizing international or regional seminars, conferences, meetings on theoretical and practical aspects of statistical computing.

The conference would not be successful without the support from your participation. I would like to thank the Local Organising Committee who have been working tirelessly in planning and organizing the Conference. I would also like to express my appreciation to the organisers of the invited sessions, all the speakers and registered participants for their exciting academic contributions to the intellectual exchange in this conference.

I hope you will have a fruitful and enjoyable conference!

Welcoming Message from Conference General Chair

Prof. Philip Leung-Ho Yu

General Chair, IASC-ARS Interim Conference 2022

Head and Professor
Department of Mathematics and Information Technology
The Education University of Hong Kong

A warm welcome to the IASC-ARS Interim Conference 2022 hosted by the Department of Mathematics and Information Technology of the Education University of Hong Kong.

It is my hope that the IASC-ARS Interim Conference 2022 would be able to achieve its objective in providing an effective forum for academician, researchers, and practitioners to discuss and exchange research ideas, new concepts and recent methods in statistical computing and artificial intelligence.

It is pleasing to note that the programme of this conference covers a wide range of interesting topics related to all theoretical and practical aspects, but not limited to high dimensional data analysis, deep learning, complex time series analysis, dimension reduction, tensor decomposition, network analysis, advances in clustering, Bayesian computation, text analytics, and applications to education, COVID-19, government, health care and finance.

The IASC-ARS Interim Conference 2022 features two keynote speeches and 60 oral presentations involving 117 authors and participants from 10 countries or regions: Hong Kong, China, Taiwan, Japan, South Korea, Singapore, the Philippines, U.K., Australia, and the Netherlands. We appreciate the full support and contribution from presenters, members of the Scientific Programme Committee and Local Organizing Committee, supporting staff and student helpers, and honorary advisors to make this conference possible!

May God bless you all with good health to make this conference a successful and enjoyable one!

Organisation

Organizers:
Department of Mathematics and Information Technology, The Education University of Hong Kong, Hong Kong, China
The Asian Regional Section of the International Association for Statistical Computing (IASC-ARS)

Honorary Advisors:
Wai Keung LI, The Education University of Hong Kong, Hong Kong, China
Kin LAM, Mr. and Mrs. Lam Kin Research Fund for AI in Educational and Financial Technologies, Hong Kong, China

General Chair:
Philip Leung Ho YU, The Education University of Hong Kong, Hong Kong, China

Scientific Programme Committee:
Wai Keung LI (Chair) , The Education University of Hong Kong, Hong Kong, China
Erniel BARRIOS, University of Philippines, The Philippines
Chun-houh CHEN, Academia Sinica, Taiwan
Ray-Bing CHEN, National Cheng Kung University, Taiwan
Cathy W.S. CHEN, Feng Chia University, Taiwan
Ying CHEN, National University of Singapore, Singapore
Christophe CROUX, EDHEC Business School, France
Ke DENG, Tsinghua University, China
Donguk KIM, Sungkyunkwan University, Korea
Jaeyong Lee, Seoul National University, Korea
Guodong LI, The University of Hong Kong, Hong Kong, China
Yuichi MORI, Okayama University of Science, Japan
Junji NAKANO, The lnstitute of Statistical Mathematics, Japan

Local Organizing Committee:
Philip Leung Ho YU (Chair), The Education University of Hong Kong, Hong Kong, China
Gary Kwok Shing CHENG (Co-Chair), The Education University of Hong Kong, Hong Kong, China
Alpha Man Ho LING (Co-Chair), The Education University of Hong Kong, Hong Kong, China
Mike Ka Pui SO (Co-Chair), Hong Kong University of Science and Technology, Hong Kong, China

Finance & Registration:
Chung Man, LAM, The Education University of Hong Kong, Hong Kong, China

Technical Support:
Eric K. S. TANG, The Education University of Hong Kong, Hong Kong, China
Adrian H. Y. TAM, The Education University of Hong Kong, Hong Kong, China

Student Helper:
Zhixuan, SONG, The Education University of Hong Kong, Hong Kong, China
Lingyi, ZHU, The Education University of Hong Kong, Hong Kong, China
Chujie, WEN, The Education University of Hong Kong, Hong Kong, China

Programme Schedule

https://www.eduhk.hk/iasc-ars2022/program-schedule.html

- **Format**: Online via Zoom
- **Presentation time:** 20-minute oral presentation + 5-minutes Q&A

(All times are displayed in GMT + 8)

Day 1: 12 Dec 2022 (Mon)			
Time	**Event (Room A)**		
9:30 – 10:00	Opening Ceremony		
10:00 – 10:45	Keynote Speech #1		
10:45 – 11:00	Break		
	Room A	***Room B***	***Room C***
11:00 – 12:40	Invited Session #5 *Some Tools in Analyzing High Dimensional Data*	Invited Session #6 *Recent Developments in Statistical Computing and Data Analysis*	Invited Session #16 *Advances in Complex Time Series*
12:40 – 14:00	Lunch Break		
14:00 – 15:40	Invited Session #12 *High-dimensional and Complex Statistical Models*	Invited Session #13 *Recent Advance in Statistical Analysis that Affect Government and Health Policy*	Invited Session #11 *Analyzing the Structure of Scientific Articles*
15:40 – 16:00	Break		
16:00 – 17:40	Invited Session #9 *Advances and Applications in Clustering*	Invited Session #2 *Econometric Analysis during Covid-19*	Invited Session #15 *Recent Developments in Financial Time Series*
17:40 – 18:40	IASC-ARS BoD Meeting		
Day 2: 13 Dec 2022 (Tue)			
Time	**Events**		
	Room A	***Room B***	***Room C***
10:00 – 11:15	Invited Session #1 *New Development in Experimental Design Related Researches*	Invited Session #3 *Modeling for complex and high-dimensional data*	Contributed Session
11:15 – 11:30	Break		
11:30 – 12:45	Invited Session #4 *Modeling Strategies in Count Data*	Invited Session #7 *Deep Learning and High-Dimensional Data Analysis*	Invited Session #8 *Statistical Computing for Large-scale Data*
12:45 – 14:00	Lunch Break		
14:00 – 15:15	Invited Session #10 *From Reports to Knowledge for Patient Safety Improvement through Advancements in Artificial Intelligence*	Invited Session #17 *AI and Data Analytics for the Public Sector*	Invited Session #14 *Statistics and Its Application in Deep Learning*
15:15 – 15:30	Break		
	Event (Room A)		
15:30 – 16:15	Keynote Speech #2		
16:15 – 16:45	Closing Ceremony		

Day 1: 12 Dec 2022 (Mon)	
Time	**Event**
9:30 – 10:00	**Opening Ceremony** • Welcoming Speech by the President of EdUHK, Prof. Stephen Yan-Leung Cheung • Welcoming Speech by Chairperson of IASC-ARS, Prof. Ray-Bing Chen • Opening Speech by Conference General Chair, Prof. Philip Leung-Ho Yu • Group Photo Taking
10:00 – 10:45	**Keynote Speech #1** *Statistical Computing and Artificial Intelligence: A Smart Health Project as An Example* Prof. Chun-Houh Chen Research Fellow and Director Institute of Statistical Science Academia Sinica Taiwan
10:45 – 11:00	Break
11:00 – 12:40	**Invited Session #5** *Some Tools in Analyzing High Dimensional Data* Organiser: Erniel B. Barrios Chairperson: Erniel B. Barrios • *Evaluation of Clustering in High Dimensional Sparse Data* Joseph Ryan G. Lansangan • *A Composite Index based on Reduced Rank Matrix Autoregressive Model* Elaine Ling Xin, Xiaohang Wang • *A Moving-Window Bayesian Network Model for Assessing Systemic Risk in Financial Markets* Shun Hin Chan, Amanda Man Ying Chu, Mike Ka Pui So • *Multiple Stratification Variables in Multipurpose Surveys with Principal Curves: Simulation Evaluations using Predicted Values* Kier Jesse Ballar, Erniel B. Barrios
11:00 – 12:40	**Invited Session #6** *Recent Developments in Statistical Computing and Data Analysis* Organiser: Ruby Chiu-Hsing Weng Chairperson: Ruby Chiu-Hsing Weng • *Categorical Exploratory Data Analysis in Major League Baseball* Elizabeth P. Chou • *Spike-and-slab Priors for Differential Item Functioning Detection in a Multiple-group IRT Tree Model* Yu-Wei Chang, Cheng-Xin Yang

	• *Permutation Tests for Testing Variable Importance* Po-Hsien Huang, Li-Ping Yen • *A Test-based Knot Selection Algorithm for Regression Splines and Some Extensions* Tzee-Ming Huang
11:00 – 12:40	**Invited Session #16** *Advances in Complex Time Series* Organiser: Wai Keung Li Chairperson: Wai Keung Li • *Grouped Network Poisson Autoregressive Model* Yuxin Tao, Dong Li, Xiaoyue Niu • *An efficient tensor regression for high-dimensional data* Yuefeng Si, Yingying Zhang, Guodong Li • *Asset pricing via the conditional quantile variational autoencoder* Ke Zhu • *Bootstrapping white noise checks for functional time series* Yu Miao, Muyi Li
12:40 – 14:00	Lunch Break
14:00 – 15:40	**Invited Session #12** *High-dimensional and Complex Statistical Models* Organiser: Jaeyong Lee Chairperson: Jaeyong Lee • *Path Algorithms for Fused Lasso Signal Approximator with Application to COVID-19 Spread in Korea* Donghyeon Yu, Johan Lim • *Asymptotic Properties for Bayesian Neural Network in Besov Space* Kyeongwon Lee • *Scalable and optimal Bayesian inference for sparse covariance matrices via screened beta-mixture prior* Seongil Jo • *Bayesian Estimation of High-dimensional Mixed DAG using Sparse Cholesky Factors* Min Ho Cho, Kyoungjae Lee, Lizhen Lin

14:00 – 15:40	**Invited Session #13** *Recent Advance in Statistical Analysis that Affect Government and Health Policy* Organiser: Thomas Fung Chairperson: Thomas Fung • *The impact of the COVID-19, social distancing, and movement restrictions on crime in NSW, Australia* Joanna J.J. Wang, Thomas Fung, Donald Weatherburn • *Modelling COVID and crime in the US as hierarchical time series* Thomas Fung, Karlene Ormsby, Joanna J.J. Wang • *A New Network-Based Multimorbidity Index for Better Primary Care Assessment for Elders: The Australian Context* Ruihua Guo, Boris Choy, Simon Poon • *Bayesian analysis of multivariate returns and covariance time series* Louis Zheng Lyu Huang
14:00 – 15:40	**Invited Session #11** *Analyzing the Structure of Scientific Articles* Organiser: Junji Nakano Chairperson: Junji Nakano • *Article's scientific prestige: Measuring the impact of individual articles in the Web of Science* Ying Chen • *A Study on How Scientific Collaboration Differs across Subjects via the Leading Author Models* Frederick Kin Hing Phoa, Hohyun Jung, Ming-Yen Lai • *Analysis of citations among U.S. patents using a stochastic generative model* Yuichiro Yasui, Junji Nakano
15:40 – 16:00	Break
16:00 – 17:40	**Invited Session #9** *Advances and Applications in Clustering* Organiser: Mika Sato-Ilic, Yuichi Mori Chairperson: Mika Sato-Ilic, Yuichi Mori • *Acceleration of Computation in Fuzzy Clustering* Yuichi Mori, Tatsuki Akaki, Masahiro Kuroda • *Fuzzy Clustering based Support Vector Machine for Autocoding and Interpretation of Results* Yukako Toko, Mika Sato-Ilic • *Deep Learning Method for Imbalanced Image Data Considering Reduction of Dimensionality by Multidimensional Scaling* Shojiro Takamura, Mika Sato-Ilic

16:00 – 17:40	**Invited Session #2** *Econometric Analysis during Covid-19* Organiser: Ying Chen Chairperson: Ying Chen • *Nonparametric comparison of epidemic time trends: the case of COVID-19* Marina Khismatullina, Michael Vogt • *Seeding Large-scale Public Health Inventions in Multilayer Social Networks with Heterogeneous Nodes* Xiaoyi Han • *Policy Effectiveness on the Global Covid-19 Pandemic and Unemployment Outcomes: A Large Mixed Frequency Spatial Approach* Yijiong Zhang, Xiaoyi Han, Yanli Zhu, and Ying Chen
16:00 – 17:40	**Invited Session #15** *Recent Developments in Financial Time Series* Organiser: Ke Zhu Chairperson: Ke Zhu • *High-dimensional Quantiled Conditional Moments with Hypergraph* Zhoufan Zhu, Ningning Zhang, Ke Zhu • *Quantild conditional variance, skewness, and kurtosis by Cornish-Fisher expansion* Ningning Zhang, Ke Zhu • *Inference for under-identified IV regression model* Kunyang Song • *Forecasting High Dimensional Long Memory Time Series based on Memory augmented Gated Recurrent Unit* Maolin Yang, Muyi Li, Guodong Li
17:40 – 18:40	IASC-ARS BoD Meeting

Day 2: 13 Dec 2022 (Tue)	
Time	**Event**
10:00 – 11:15	**Invited Session #1** *New Development in Experimental Design Related Researches* Organiser: Ray-Bing Chen Chairperson: Ray-Bing Chen • *Predictive Subdata Selection for Large-Scale Deterministic Computer Models* Ming-Chung Chang • *A Survey on Multi-step Lookahead Bayesian Optimization* Peng Liu • *Generalized Bayesian D-optimal supersaturated multistratum designs* Chang-Yun Lin
10:00 – 11:15	**Invited Session #3** *Modeling for complex and high-dimensional data* Organiser: Qianqian Zhu Chairperson: Qianqian Zhu • *Determination of the effective cointegration rank in high-dimensional time-series predictive regression* Puyi Fang, Zhaoxing Gao, Ruey S. Tsay • *Quantile index regression* Yingying Zhang, Yuefeng Si, Guodong Li and Chil-Ling Tsai • *Network Autoregression for Incomplete Matrix-Valued Time Series* Xuening Zhu, Feifei Wang, Zeng Li
10:00 – 11:15	**Contributed Session** Chairperson: Alpha Man Ho Ling • *A Semiparametric Approach to Empirical Bayes Estimation of Discrete False Discovery Rates Using Kernels* Dominic B. Dayta, Sean Oliver M. Escalante • *Implementation of SVIR Mathematical Model in Simulating Covid-19 Infection Dynamics under Vaccination Intervention* Eduardo C. Dulay Jr., Dr. Bernhard Egwolf., Adrian Roel P. Lapus, Ian D. Marco • *AI-based Scoring for Picture-cued Writing Assessment* Yipeng Zhuang, Ruibin Zhao, Zhiwei Xie, Philip Leung Ho Yu

11:15 – 11:30	Break
11:30 – 12:45	**Invited Session #4** *Modeling Strategies in Count Data* Organiser: Erniel B. Barrios Chairperson: Erniel B. Barrios • *Spatiotemporal Modeling of Sparse Geostatistical Epidemic Data* Shirlee R. Ocampo, Erniel B. Barrios • *Incorporating Structural Change into Modeling of COVID-19 in the Philippines: A Spatiotemporal Model* Regina M. Tresvalles, Erniel B. Barrios • *Nonparametric Density-based Procedure of Detecting Emerging Events from Social Media* Louise Adrian DC. Castillo, Erniel B. Barrios
11:30 – 12:45	**Invited Session #7** *Deep Learning and High-Dimensional Data Analysis* Organiser: Han-Ming Wu Chairperson: Han-Ming Wu • *Combining measures of signal complexity and deep learning in medical images for neurodegenerative disease screening* Yi-Ju Lee, Chun-houh Chen • *Variable selection and estimation for misclassified responses and high-dimensional error-prone predictors* Li-Pang Chen • *Improve the Performance of Deep Learning Algorithms by Supervised Dimension Reduction Methods* Han-Ming Wu
11:30 – 12:45	**Invited Session #8** *Statistical computing for large-scale data* Organiser: Yingying Zhang Chairperson: Yingying Zhang • *Fast and Robust Sparsity Learning over Networks: A Decentralized Median Regression Approach* Xiaojun Mao • *Robust Inference for Change Points in High Dimension* Feiyu Jiang • *A Partially Functional Linear Modeling Framework for Integrating Genetic, Imaging, and Clinical Data* Ting Li

12:45 – 14:00	Lunch Break
14:00 – 15:15	**Invited Session #10** *From Reports to Knowledge for Patient Safety Improvement through Advancements in Artificial Intelligence* Organiser: Zoie SY Wong Chairperson: Zoie SY Wong • *The Development of an Annotation Guidelines for Medication-related Incident Reports* Zoie SY Wong, Ryohei Sasano, Shin Ushiro, Neil Waters • *Active Learning Framework for Clinical Named Entity Recognition based on Transformers and Transfer Learning* Jiaxing LIU, Zoie SY Wong • *Development of Multitask Incident Reports-pretrained BERT Model to Empower Incident Reporting and Learning System* Zoie SY Wong, Dentsu Data Artist Mongol Team
14:00 – 15:15	**Invited Session #17** *AI and Data Analytics for the Public Sector* Organiser: Leo Yu Chairperson: Philip Leung-Ho Yu • *Exploratory Research on Use of New Data for Computation of Private Housing Rent Index* Helen Y.W. Lee, Benjamin C.W. Cheung, Daniel W.L. Chong, Natalie K.P. Chung, Kevin S.Y. Hsia • *Automatic Coding of Occupations in General Household Survey (GHS) of Hong Kong* Ronald C.H. Chan, Chris C.W. Leung, Sharon P.W. Ng, Frances C.W. Wong, Matthew T.L. Wong • *Anomaly Detection in Merchandise Trade Data: A Deep Learning Approach* Benjamin C.H. Chan, Natalie K.P. Chung, Ian Y.C. Ng
14:00 – 15:15	**Invited Session #14** *Statistics and its application in deep learning* Organiser: Guodong Li, Feiqing Huang • *HAR-Itˆo models and high-dimensional HAR modeling forhigh-frequency data* Huiling Yuan, Yifeng Guo, Kexin Lu, Guodong Li • *Cross-layer retrospective retrieving via layer attention* Yanwen Fang, Yuxi Cai, Guodong Li • *SARMA: A Computationally Scalable High-Dimensional Time Series Model* Kexin Lu, Feiqing Huang, Yao Zheng, Guodong Li

15:15 – 15:30	Break
15:30 – 16:15	**Keynote Speech #2** *Modelling Matrix Time Series via a Tensor CP-Decomposition* Prof. Qiwei Yao Professor of Statistics Department of Statistics London School of Economics United Kingdom
16:15 – 16:45	**Closing Ceremony** • Closing Remark by Conference General Chair, Prof. Philip Leung-Ho Yu • Closing Speech by Chairperson of IASC-ARS, Prof. Ray-Bing Chen • Introduction to IASC-ARS 2023 by Conference Chair, Prof. Thomas Fung

Keynotes

Keynote 1: Statistical Computing and Artificial Intelligence: A Smart Health Project as An Example

Professor Chun-houh CHEN

Research Fellow and Director
Institute of Statistical Science
Academia Sinica
Taiwan

Abstract:

The Academia Sinica funded Smart Health project (http://sites.stat.sinica.edu.tw/SH/) is composed of researchers from three institutes (Biomedical Science, Information Science, and Statistical Science) with two primary data sources, the Taiwan Biobank (TWB, https://www.twbiobank.org.tw/index.php) and the Taiwan Precision Medicine Initiative (TPMI, https://tpmi.ibms.sinica.edu.tw/www/en/). TWB is collecting blood/urine/genetic data with surveyed phenotypes and medical images of 200,000 community-based participants across Taiwan. In comparison, TPMI will have genetic profiles and clinical information of 1,000,000 participants from 33 hospitals in Taiwan. It is possible to link the TWB participants to the National Health Insurance Research Database (NHIRD), Taiwan, with almost all 23 million population. On the other hand, with more than 30 working groups, each with a specific combination of diseases and treatments, the TPMI aims to optimize the clinical practice of precision medicine and identify genetic risks for diseases in Taiwan. Besides TWB and TPMI, the Smart Health project is also establishing a network with the major medical centers in Taiwan for collaborative research and developing real-world products in smart health. In this talk, we shall report how statistical computing and artificial intelligence interact in analyzing the vast and complex data from TWB and TPMI and developing related real-world medical informatics products for the collaborating hospitals in the Smart Health project. We shall also try to touch the following additional interplays: between academia and medicine, between data scales, between projects and expertise, between academia and industry, and between knowledge and real-world applications.

Keynote 2: Modelling Matrix Time Series via a Tensor CP-Decomposition

Professor Qiwei YAO

Professor of Statistics
Department of Statistics
London School of Economics
United Kingdom

Abstract:

We consider to model matrix time series based on a tensor CP-decomposition. Instead of using an iterative algorithm which is the standard practice for estimating CP-decompositions, we propose a new and one-pass estimation procedure based on a generalized eigenanalysis constructed from the serial dependence structure of the underlying process. To overcome the intricacy of solving a rank-reduced generalized eigenequation, we propose a further refined approach which projects it into a lower-dimensional full-ranked eigenequation. This refined method significantly improves the finite-sample performance of the estimation. The asymptotic theory has been established under a general setting without the stationarity. It shows, for example, that all the component coefficient vectors in the CP-decomposition are estimated consistently with certain convergence rates. The proposed model and the estimation method are also illustrated with both simulated and real data, showing effective dimension-reduction in modelling, and forecasting matrix time series.

Paper Presentations

Evaluation of clustering in high dimensional sparse data

Joseph Ryan G. Lansangan[1]

[1]*University of the Philippines, jglansangan@up.edu.ph*

Abstract

Different metrics are used to evaluate the goodness of a cluster solution and to recommend the right number of clusters. A popular evaluation measure is the Silhouette Coefficient (SC) (Rousseeuw, 1987), which evaluates cluster cohesion or whether observations are clustered well, and cluster separation or whether neighboring clusters are separated well. However, the cluster solution may be sensitive to the choice of distance measure, and evaluation using SC becomes more difficult when working with high dimensional (i.e., a large number of features) or sparse (e.g., subpopulations inherent only for a few observations on some features) data. In this study, we propose an alternative method to assess clustering efficiency using the generalized SC (Lengyel and Botta-Dukat, 2019) in a reduced dimension space of the data. Specifically, the generalized SC will be implemented in feature selection and as a performance measure. We examine how the proposed approach performs on simulated data sets with varying clustering characteristics.

Keywords: Clustering efficiency; High dimensional data; Silhouette.

A Composite Index based on Reduced Rank Matrix Autoregressive Model

Elaine Ling Xin[1], Xiaohang Wang[2]

[1]*BNU-HKBU United International College, China, elainexin@uic.edu.cn*
[2]*Shenzhen Technology University, China, anita.xhwang@yahoo.com*

Abstract

R&D expenses and patent citations are two of the most commonly used indicators for describing companies' innovation activities, while R&D expenses is an input indicator, and patent citation number is an output indicator. In each year, the observations of firms' innovation activities naturally have two classifications, either can be classified by firms or by indicators. Hence, a matrix can be formed with each column for a firm and each row for an indicator. The observations collected over time form a matrix-valued time series. The Matrix-valued autoregressive model have a column structure to capture the interplay among different firms and row structure to capture the interplay among different innovation indicators. In order to derive a composite index, we propose to impose low rank constrain to the left coefficient matrix to achieve the dimension reduction of multiple indicators. To achieve further dimension reduction due to limited data, we also impose the diagonal constrain to the right coefficient matrix to ignore the possible spillovers among the different firms. Our problem of interest is then formulated as estimation and statistical inference on the constrained MAR model. We develop the estimation procedure by iterated least squares method and make the inference by bootstrapping method. The simulation results verified the performance of our estimation procedure. In real data analysis, we apply the model to estimate the innovation index of a sample of Chinese listed firms from five different industries. The results show that the composite index is stable and reliable in selecting the most innovative enterprises. And the levels of persistency in the time series also reveals useful information of firm's innovation activities.

Keywords: Matrix Time Series; Composite Index; Reduced Rank; Firm Innovation.

A Moving-Window Bayesian Network Model for Assessing Systemic Risk in Financial Markets

Shun Hin Chan[1], Amanda Man Ying Chu[2], Mike Ka Pui So[3]

[1]*Department of Information Systems, Business Statistics and Operations Management, The Hong Kong University of Science and Technology, shchanai@connect.ust.hk*
[2]*Department of Social Sciences, The Education University of Hong Kong, amandachu@eduhk.hk*
[3]*Department of Information Systems, Business Statistics and Operations Management, The Hong Kong University of Science and Technology, immkpso@ust.hk*

Abstract

Systemic risk refers to the uncertainty that arises due to the breakdown of a financial system. The concept of "too connected to fail" suggests that network connectedness plays an important role in measuring systemic risk. The failure of a small part can trigger the failure of the other parts, which eventually leads to the failure of the whole system. In this paper, we first recover a time series of Bayesian networks for stock returns, which allow the direction of links among stock returns to be formed with Markov properties in directed graphs, as opposed to the usual undirected network approaches in the literature. We rank the stocks in the time series of Bayesian networks based on the topological orders of the stocks in the learned Bayesian networks, in order to quantify the topological features of the stocks over time. A topological order is a linear ordering of the nodes (which are the stocks in our study) in a Bayesian network. We then propose an order distance, a new measure with which to assess the changes in the topological orders of the stocks. The order distance is the L1-norm of the difference between the topological orders on two consecutive trading days, where we normalized the topological orders to mimic the idea of portfolio weights. In an empirical study using stock data from the Hang Seng Index in Hong Kong from 2008 to 2021, we use the order distance to predict the extreme absolute return, which is a proxy of extreme market risks, or a signal of systemic risks, using the LASSO regression model. Our results indicate that the network statistics of the time series of Bayesian networks and the order distance substantially improve the predictability of extreme absolute returns. We illustrate the use of the rolling-window Bayesian network modeling with the topological order in predicting extreme market changes and provide insights into the assessment of systemic risk. We also apply the proposed methodology to the Dow Jones Industrial Index. The prediction performances of extreme absolute returns are also improved when we include the order distance as a predictor. These results supports that the application of the methodology could be applied to different stock markets. We also note that the proposed order distance is not only restricted to use in the LASSO regression model, but is also usable as a predictor in any volatility

prediction models for assessing systemic risk, which can provide additional information to enhance the prediction performance and financial risk management.

Keywords: Bayesian Network; Time Series Prediction; Financial Market; Financial Network Connectedness; Topological Order; Systemic Risk; Volatility Model; Moving-Window Model.

Multiple Stratification Variables in Multipurpose Surveys with Principal Curves: Simulation Evaluations using Predicted Values

Kier Jesse Ballar[1], Erniel B. Barrios, PhD[2]

[1]*School of Statistics, University of the Philippines Diliman, kdballar@up.edu.ph*
[2]*School of Statistics, University of the Philippines Diliman, ebbarrios@up.edu.ph*

Abstract

In stratified sampling, constructing strata which are homogenous within and heterogenous across is critical in obtaining precise estimates. Further, most existing strata construction methods, such as the cumulative root and Sethi's (1963) method, rely on a univariate stratification variable that is correlated to the target variable. However, most surveys are multipurpose with multiple target variables. A dilemma now arises with the existing strata construction methods for multipurpose surveys, which stratification variable to choose among the multiple potential stratification variables. We approach this dilemma by proposing a new strata construction method which utilizes all the multiple stratification variables instead of choosing a univariate stratification variable, through dimension reduction methods such as the Principal Curve. Further, a new approach in evaluating stratified sampling estimates is adopted — instead of using the stratification variables to evaluate the precision of the stratified sampling design estimates, we use predicted values of the target variables using information from the stratification variables. Evaluation of the stratified sampling estimates show that the proposed strata construction method, rooted specifically in the Local Principal Curve, is beneficial in increasing the precision of estimates for multipurpose surveys with multiple target variables.

Keywords: Multipurpose Surveys; Multivariate Stratification; Principal Curves; Stratified Sampling; Principal Components Analysis.

Categorical Exploratory Data Analysis in Major League Baseball

[1]Elizabeth P. Chou

[1]National Chengchi University, eptchou@nccu.edu.tw

Abstract

Many-System Problem studies how heterogeneity constraints and hides structural mechanisms and how to uncover and reveal hidden major factors from homogeneous parts. We developed a computational protocol for identifying various collections of significant factors of various orders underlying response-vs-covariate dynamics. Categorical Exploratory Data Analysis (CEDA) on PITCHf/x database for the information content of Major League Baseball's (MLB) pitching dynamics will be used as an example. In the first part of the talk, I will discuss an indirect-distance-measure-based label embedding tree that leads to discovering the asymmetry of mixing geometries among labels' point-clouds. In the second part, the CEDA-based major factor selection protocol will be presented.

Keywords: Categorical Exploratory Data Analysis; PITCHf/x; Major League Baseball; Feature Selection.

Spike-and-Slab Priors for Differential Item Functioning Detection in an IRT Tree Model

Yu-Wei Chang[1], Cheng-Xin Yang[2]

[1] *Department of Statistics, National Chengchi University, ychang@nccu.edu.tw*
[2] *Department of Statistics, National Chengchi University*

Abstract

Group difference has particular implications in analyzing questionnaire or item response data. For example, whether two persons from different demographic groups, such as gender or race, with same shopping preference have different shopping habits on one aspect helps store manager better design their display. The shopping habits and shopping preference could be measured, respectively, by items and some latent factor in a questionnaire, and the different shopping habits observed on an item are called differential item functioning (DIF). In the current study, we propose a multiple-group Item Response Theory tree model to take the group difference and missing data in questionnaire or item response data into consideration. Different from most of present DIF studies where one has to select anchor items and then detect DIF items, we achieve DIF detection and parameter estimation simultaneously through applying some spike-and-slab priors (Ishwaran and Rao 2005; Rockova and George 2018) in Bayesian estimation. We will present the suggested estimation procedure for the MG-IRTree model in this talk. Simulation studies are conducted to illustrate the validation of the proposed estimation procedure and the efficiency of DIF detection. The proposed method is further applied to a real data set for illustration.

Keywords: Bayesian estimation; differential item functioning; Item Response Theory tree model; missing data; spike-and-slab priors.

Permutation Tests for Testing Variable Importance

Po-Hsien Huang[1], Li-Ping Yen[2]

[1]*National Chengchi University, psyphh@nccu.edu.tw*
[2] *National Chengchi University, 111752001@nccu.edu.tw*

Abstract

Traditional statistics highly relies on linear models to test scientific hypotheses. Recently, machine learning algorithms provide alternative ways to discover the relationships among variables. To assist interpretating the results made by these "black-box" algorithms, variable importance indices were developed to quantify the effects of features. However, these indices are only descriptive and cannot be used for statistical hypothesis testing. In this study, we proposed a permutation test for variable importance indices. The permutation test calculates p-values and can be used to determine the statistical significances of features. When the learning algorithms were linear regression and random forest, a simulation study showed that the proposed method controls empirical type I error well and exhibit reasonable power. Theoretical properties of the permutation test are waiting for development.

Keywords: Permutation Test, Variable Importance, Machine Learning

A Test-based Knot Selection Algorithm for Regression Splines and Some Extensions

Tzee-Ming Huang[1]

[1]*National Chengchi University, tnhuang@nccu.edu.tw*

Abstract

Spines are known to have good approximation power to smooth functions and are often used in nonparametric function estimation. The choice of the knots of a space of splines is crucial to the approximation power of splines. A test based on knot selection algorithm has been proposed in Huang (2019), The test-based method has been extended to the logistic regression framework and to the multivariate case for knot detection. Recently, I derived an extension of this method to the density estimation case and found that it worked in some simulation experiments. In this talk, I will review the approach in Huang (2019) and present the new extension to the density estimation case.

Keywords: knot detection; splines; regression; density estimation.

Grouped Network Poisson Autoregressive Model

Yuxin Tao[1], Dong Li2, Xiaoyue Niu3

[1]*Tsinghua University, taoyx19@mails.tsinghua.edu.cn*
[2]*Tsinghua University, malidong@tsinghua.edu.cn*
[3]*The Pennsylvania State University, xiaoyue@psu.edu*

Abstract

Multivariate Poisson autoregressive models are common ways to fit count time series data, while the statistical inference is quite challenging. The network Poisson autoregressive model (NPAR) reduces the inference complexity by incorporating network information into the dependence structure, where the response of each individual can be explained by its lagged values and the average effect of its neighbors. However, NPAR makes one strong assumption that all individuals are homogeneous and they share a common autoregressive coefficient. Here we propose a grouped network Poisson autoregressive model (GNPAR), where the individuals are classified into different groups with group-specific parameters to describe heterogeneous nodal behaviors. We present the stationarity and ergodicity of the GNPAR model and study the asymptotic properties of the maximum likelihood estimation. We develop an EM algorithm to estimate the unknown group labels and investigate the finite-sample performance of our estimation procedure using simulations. We analyze the Chicago Police Investigatory Stop Report data and find distinct dependence patterns in different neighborhoods of Chicago that could be potentially helpful for future crime prevention.

Keywords: EM algorithm; Individual heterogeneity; Maximum likelihood estimation; Multivariate Poisson autoregression; Network data.

An efficient tensor regression for high-dimensional data

Yuefeng Si[1], Yingying Zhang[2], Guodong Li[1]

[1]University of Hong Kong, [2]East China Normal University

Abstract

Most currently used tensor regression models for high-dimensional data are based on Tucker decomposition, which has good properties but loses its efficiency in compressing tensors very quickly as the order of tensors increases, say greater than four or five. However, for the simplest tensor autoregression in handling time series data, its coefficient tensor already has the order of six. This paper revises a newly proposed tensor train (TT) decomposition and then applies it to tensor regression such that a nice statistical interpretation can be obtained. The new tensor regression can well match the data with hierarchical structures, and it even can lead to a better interpretation for the data with factorial structures, which are supposed to be better fitted by models with Tucker decomposition. More importantly, the new tensor regression can be easily applied to the case with higher order tensors since TT decomposition can compress the coefficient tensors much more efficiently. The methodology is also extended to tensor autoregression for time series data, and non-asymptotic properties are derived for the ordinary least squares estimations of both tensor regression and autoregression. A new algorithm is introduced to search for estimators, and its theoretical justification is also discussed. Theoretical and computational properties of the proposed methodology are verified by simulation studies, and the advantages over existing methods are illustrated by two real examples.

Keywords: high-dimensional time series; nonasymptotic properties; projected gradient descent; tensor decomposition; tensor regression.

Asset Pricing via the Conditional Quantile Variational Autoencoder

Ke Zhu[1]

[1]*Department of Statistics & Actuarial Science, The University of Hong Kong, mazhuke@hku.hk*

Abstract

We propose a new asset pricing model that is applicable to the big panel of return data. Our model aims to explain the conditional mean of the return from the conditional distribution of the return, which is approximated by a step distribution function constructed from conditional quantiles of the return. To study conditional quantiles of the return, we propose a new conditional quantile variational autoencoder (CQVAE) network. The CQVAE network species a factor structure for conditional quantiles with latent factors learned from a VAE network and nonlinear factor loadings learned from a "multi-head" network. Under the CQVAE network, we allow the observed covariates such as asset characteristics to guide the structure of latent factors and factor loadings. Furthermore, we provide a two-step estimation procedure for the CQVAE network. Finally, we apply our CQVAE asset pricing model to analyze a large 60-year US equity return data set. Compared with the benchmark conditional autoencoder model, the CQVAE model not only delivers much larger values of out-of-sample total and predictive R^2s, but also earns at least 30.9% higher values of Sharpe ratios for both long-short and long-only portfolios.

Keywords: Conditional asset pricing model; Dynamic loadings; Machine learning; Neural networks; Nonlinear quantile factor model; Variational autoencoder

Bootstrapping white noise checks for functional time series

Yu Miao[1], Muyi Li[2]

[1] *Department of Statistics and Data Science, School of Economics, Xiamen University, China, 474166746@qq.com*
[2] *Department of Statistics and Data Science, School of Economics, Xiamen University, China, limuyi@xmu.edu.cn*

Abstract

We consider white noise testing for stationary functional time series observations. Our procedure is based on the sum of the $\boldsymbol{L^2}$-norms of the empirical autocovariance operators. The limiting distributions of the proposed test statistic are established under both the null and alternative hypotheses in Hilbert space. Due to the unknown dependent structure of the observations, the test statistics is non-pivotal hence we employ the block bootstrap procedure to obtain the critical values and justify its first-order consistency. Compared to the existing methods, our test allows for the possible nonlinearity in the functional time series and does not involve the selection of functional principal components. Monte Carlo simulations under various scenarios suggest the effectiveness of the bootstrap approach and a real example is analyzed.

Keywords: Functional time series; White noise checking; Block bootstrap; Hilbert space.

Path Algorithms for Fused Lasso Signal Approximator with Application to COVID-19 Spread in Korea

Won Son[1], Johan Lim[2], Donghyeon Yu[3]

[1]*Department of Statistics, Dankook University, son.won@dankook.ac.kr*
[2]*Department of Statistics, Seoul National University, johanlim@stats.snu.ac.kr*
[3]*Department of Statistics, Inha University, dyu@inha.ac.kr*

Abstract

The fused lasso signal approximator (FLSA) is a smoothing procedure for noisy observations that uses fused lasso penalty on unobserved mean levels to find sparse signal blocks. Several path algorithms have been developed to obtain the whole solution path of the FLSA. However, it is known that the FLSA has model selection inconsistency when the underlying signals have a stair-case block, where three consecutive signal blocks are either strictly increasing or decreasing. Modified path algorithms for the FLSA, such as Qian and Jia (2016) and Son and Lim (2019), have been proposed to guarantee model selection consistency regardless of the stair-case block. In this paper, we provide a comprehensive review of the path algorithms for the FLSA and prove the properties of the recently modified path algorithms' hitting times. Specifically, we reinterpret the modified path algorithm by Son and Lim (2019) as the path algorithm for local FLSA problems and reveal the condition that the hitting time for the fusion of the modified path algorithm is not monotone in a tuning parameter. To recover the monotonicity of the solution path, we propose a pathwise adaptive FLSA having monotonicity with similar performance as the modified solution path algorithm. Finally, we apply the proposed method to the number of daily-confirmed cases of COVID-19 in Korea to identify the change points of its spread.

Keywords: Change points; fused lasso signal approximator; modified path algorithm; pathwise adaptive weight; solution path.

Asymptotic Properties for Bayesian Neural Network in Besov Space

Kyeongwon Lee[1]

[1]*Seoul National University, South Korea, lkw1718@snu.ac.kr*

Abstract

Neural networks have shown great predictive power when dealing with various unstructured data such as images and natural languages. The Bayesian neural network captures the uncertainty of prediction by putting a prior distribution for the parameter of the model and computing the posterior distribution. In this paper, we show that the Bayesian neural network using spike-and-slab prior has consistency with nearly minimax convergence rate when the true regression function is in the Besov space. Even when the smoothness of the regression function is unknown the same posterior convergence rate holds and thus the spike and slab prior is adaptive to the smoothness of the regression function. We also consider the shrinkage prior and show that it has the same convergence rate. In other words, we propose a practical Bayesian neural network with guaranteed asymptotic properties.

Scalable and optimal Bayesian inference for sparse covariance matrices via screened beta-mixture prior

Seongil Jo[1]

[1]*Inha University, South Korea, joseongil@gmail.com*

Abstract

In this paper, we propose a scalable Bayesian method for sparse covariance matrix estimation by incorporating a continuous shrinkage prior with a screening procedure. In the first step of the procedure, the off-diagonal elements with small correlations are screened based on their sample correlations. In the second step, the posterior of the covariance with the screened elements fixed at 0 is computed with the beta-mixture prior. The screened elements of the covariance significantly increase the efficiency of the posterior computation. The simulation studies and real data applications show that the proposed method can be used for the high-dimensional problem with the 'large p, small n'. In some examples in this paper, the proposed method can be computed in a reasonable amount of time, while no other existing Bayesian methods work for the same problems. The proposed method has also sound theoretical properties. The screening procedure has the sure screening property and the selection consistency, and the posterior has the optimal minimax or nearly minimax convergence rate under the Frobeninus norm.

Keywords: Sure screening property; selection consistency; minimax convergence rate.

Bayesian Estimation of High-dimensional Mixed DAG using Sparse Cholesky Factors

Min Ho Cho[1], Kyoungjae Lee[2], Lizhen Lin[3]

[1]*Department of Statistics, Inha University, South Korea, mcho@inha.ac.kr*
[2]*Department of Statistics, Sungkyunkwan University, South Korea, leekjstat@gmail.com*
[3]*Department of Applied and Computational Mathematics and Statistics, The University of Notre Dame, U.S.A., llin4@nd.edu*

Abstract

There have been some recent works on modeling a mixed directed acyclic graph (DAG) in biomedical studies which includes both directed and undirected edges, subject to the restriction of no directed cycles in the graph. In a mixed DAG, some of the directed edges representing the causal relationships between variables are often masked by undirected edges induced by unobserved confounding factors. We study the high-dimensional sparse mixed DAG models applying a factor analysis model to effectively identify and remove undirected edges. The proposed approach is based on the empirical sparse Cholesky prior for the directed edges and the Gaussian prior on the latent factors, leading to a simpler and more interpretable causal network. In addition, we derive the posterior contraction rates for precision matrices and Cholesky factors with respect to various matrix norms, and obtain the model selection consistency under general conditions. Our proposed method is evaluated and compared to existing methods through both numerical simulation studies and real data examples.

Keywords: Cholesky factor; Factor analysis model; Mixed DAG model; Model selection consistency; Posterior contraction rate.

The impact of the COVID-19, social distancing, and movement restrictions on crime in NSW, Australia

Joanna J.J. Wang[1], Thomas Fung[2], Donald Weatherburn[3]

[1]*School of Mathematical & Physical Sciences, University of Technology, Sydney, Australia, joanna.wang@uts.edu.au*
[2]*School of Mathematical & Physical Sciences, Macquarie University, Sydney, Australia, thomas.fung@mq.edu.au*
[3]*National Drug and Alcohol Research Centre, University of New South Wales, Sydney, Australia, d.weatherburn@unsw.edu.au*

Abstract

The spread of COVID-19 has prompted Governments around the world to impose draconian restrictions on business activity, public transport, and public freedom of movement. The effect of these restrictions appears to vary from country to country and, in some cases, from one area to another within a country. In this talk, we examines the impact of the COVID-19 restrictions imposed in New South Wales (NSW) by the State Government. We examine week-to-week changes in 13 categories of crime (and four aggregated categories) from 2 January 2017 to 28 June 2020. Rather than using the pre-intervention data to make a forecast and then comparing that with what is actually observed, we use a Box–Jenkins (ARIMA) approach to model the entire time series. Our results are broadly in accord with those of other studies, but we find no effect of the lockdown (upward or downward) on domestic assault.

Keywords: COVID-19; Lockdown; Assault; Criminal opportunity; Domestic violence.

Modelling COVID and crime in the US as hierarchical time series

Thomas Fung[1], Karlene Ormsby[2], Joanna J.J. Wang[3]

[1]*School of Mathematical & Physical Sciences, Macquarie University, Sydney, Australia, thomas.fung@mq.edu.au*
[2]*School of Mathematical & Physical Sciences, Macquarie University, Sydney, Australia, karlene.ormsby@mq.edu.au*
[3]*School of Mathematical & Physical Sciences, University of Technology, Sydney, Australia, joanna.wang@uts.edu.au*

Abstract

Crime time series data can often be naturally disaggregated by various attributes of interest, either by their crime type or geographical location. When modelling this type of data, the current recommended 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 are often only interested at the big picture, which requires researchers to either simply summing the fitted value series up or model the aggregated series independently. This leads to poor forecasting performance at the higher levels of aggregation in practice as the most disaggregated series often have a high degree of volatility while the most aggregated time series is usually smooth and less noisy. Intuition also requires the forecasts to add up in the same way as the data, but one can't guarantee that would be the case when series are modelled independently. In this talk, we will explain why the hierarchical and grouped time series method of Wickramasuriga et al. (2019) should be considered as the default technique for modelling this kind of data. US COVID and crime data of Abrams (2020) will be used as an example.

Keywords: hierarchical time series; grouped time series; fable; US crime; COVID.

A New Network-Based Multimorbidity Index for Better Primary Care Assessment for Elders: The Australian Context

Ruihua GUO[1], Boris Choy[2], Simon Poon[3]

[1]*Australian Institute of Health and Welfare, Ruihua.Guo@aihw.gov.au*
[2]*The University of Sydney, boris.choy@sydney.edu.au*
[3]*The University of Sydney, simon.ppn@sydney.edu.au*

Abstract

Multimorbidity is a complex problem that has received increasing attention in recent decades. The static quantitative measurements and small-scale conditions focus may obscure the interaction amongst conditions. Network science can inherently capture these interactions via extracting topological features in the network. This study developed a dynamic index to assess primary care for the elderly with multimorbidity. 1924 Australians with long-term multimorbidity, aged 65 and over, were extracted from the Australian National Health Survey 2014-2015. Twenty-eight health conditions were weighted via gender-stratified network eigenvector centralities. The correlation between the network index for individuals and corresponding general practitioner (GP) utilization was compared with the Charlson Comorbidity Index (CCI). The network approach can achieve a 114% increase in this correlation and a 53% increase in R^2 after risk-factor adjustment compared to the CCI. Therefore, network approach may offer a new and dynamic approach to assess GP utilization for Australian elderly patients with mental and physical multimorbidity. A high correlation between CCI and network index suggests that combining with CCI the new index may promote the assessment of primary care utilization. However, the generalizability of the new approach requires further validation tests.

Keywords: Weighted indices; Network analysis; Multimorbidity; Primary care; Population aging.

Bayesian analysis of multivariate returns and covariance time series

Zheng Lyu HUANG[1]

[1]*The University of Sydney, zhua8113@uni.sydney.edu.au*

Abstract

This paper proposes to model the observed time series of covariance matrices using some matrix distributions such as Wishart distribution. We propose two ways, modelling directly the covariance matrix or modelling the variances and correlation matrix, to capture the dynamics for the mean process of the distribution which is also the covariance process of returns. Then we apply the multivariate Variance Gamma distribution to model the returns adopting the fitted covariance matrices subjected to scale adjustment. The whole model is called the multivariate two-stage volatility-return model. We test the accuracy of our models using simulated data with various levels of long-term correlation and dispersion of Wishart distributions. We also compare the two ways of specifying the mean process of Wishart distribution. We implement our models using Rtan and apply the models to analyse some cryptocurrencies.

Article's scientific prestige: Measuring the impact of individual articles in the Web of Science

Ying Chen[1]

[1]*National University of Singapore, Singapore, matcheny@nus.edu.sg*

Abstract

We performed a citation analysis on the Web of Science publications consisting of more than 63 million articles and 1.45 billion citations on 254 subjects from 1981 to 2020. We proposed the Article's Scientific Prestige (ASP) metric and compared this metric to number of citations (#Cit) and journal grade in measuring the scientific impact of individual articles in the large-scale hierarchical and multi-disciplined citation network. In contrast to #Cit, ASP, that is computed based on the eigenvector centrality, considers both direct and indirect citations, and provides steady-state evaluation cross different disciplines. We found that ASP and #Cit are not aligned for most articles, with a growing mismatch amongst the less cited articles. While both metrics are reliable for evaluating the prestige of articles such as Nobel Prize winning articles, ASP tends to provide more persuasive rankings than #Cit when the articles are not highly cited. The journal grade, that is eventually determined by a few highly cited articles, is unable to properly reflect the scientific impact of individual articles. The number of references and coauthors are less relevant to scientific impact, but subjects do make a difference. This is a joint work with Thorsten Koch, Nazgul Zakiyeva, Kailiang Liu, Zhitong Xu, Chun-houh Chen, Junji Nakano, Keisuke Honda, https://arxiv.org/abs/2202.08695

A Study on How Scientific Collaboration Differs across Subjects via the Leading Author Models

Frederick Kin Hing Phoa[1], Hohyun Jung[2], Ming-Yen Lai[3]

[1]*Institute of Statistical Science, Academia Sinica, Taiwan.*
fredphoa@stat.sinica.edu.tw
[2]*Department of Statistics, Sungshin Women's University, South Korea.*
hhjung@sungshin.ac.kr
[3]*Data Science Degree Program, National Taiwan University, Taiwan.*
0102322@gmail.com

Abstract

The pattern of collaboration is obviously different across different subjects in scientific researches, but a quantitative measurement on such difference is lack. In specific, due to the variations in the development of research fields, the popularity effects, which refers to the advantages of popular authors in making more publications, of the scientific collaboration process differ. In this work, we first focus on the popularity effect of the scientific collaboration process that popular authors have an advantage in making more publications. Standard network analysis has been used to analyze the scientific collaboration network, but it is limited in explaining the scientific output by binary co-authorship relationships as papers have various numbers of authors. We apply a leading author model to understand the popularity effect mechanism while avoiding the use of the standard network structure. For each subject found in the Web of Science, the estimation algorithm helps to analyze the size of the popularity effect, and we list influential authors through the estimated genius levels of authors from the popularity effect. We apply the proposed model to the real scientific collaboration data, and the results show positive popularity effects in all the collaborative systems. Finally, we classify all subjects into several categories with similar properties of the leading author models.

Keywords: Scientific Collaboration Network, Leading Author Model, Popularity Effects, Subject Classification.

Analysis of citations among U.S. patents using a stochastic generative model

Yuichiro Yasui[1], Co. Junji Nakano[2]

[1]*The Graduate University for Advanced Studies, SOKENDAI, Japan, y-yasu@ism.ac.jp*
[2]*Chuo University, Japan, nakanoj@tamacc.chuo-u.ac.jp*

Abstract

The U.S. patent bibliographic database published by the NBER (National Bureau of Economic Research) contains 3,774,768 articles and 16,518,948 citations from 1975 to 1999. We extracted articles with metadata such as categories and subcategories and constructed a citation network consisting of 2,075,770 nodes (=articles) and 10,557,536 edges (=citations). It is easily found that the number of citations within a category is clearly larger than the number of citations between categories, and the same structure applies to subcategories, forming a hierarchical cluster structure consisting of categories and subcategories. We have already proposed a generative model to explain citation networks of articles in some academic fields. The model assumes that citations occur with a probability based on the citation type, the importance of the cited article, and the difference between the publication times of the two articles. Citation is expressed by a directed edge, generated by the Preferential attachment mechanism using above defined information, or by the Triad formation (TF) mechanism, which constructs a triad citation structure with probability p. This model assumes a "single field" in academic papers and is not suitable for a whole hierarchal network of patent citations. Therefore we focus on the citation network consisting of 81,457 nodes and 252,503 edges related to subcategory #41 (Electrical Devices). We found that our model has some problems, especially with the distribution of the number of triangles on each node. We found that the fit cannot be improved by changing the execution probability p of the TF mechanism. Then, we modify the model so that the execution probability of the TF mechanism can be set flexibly. This model is useful for explaining the citation relationships in subcategory #41. We can find that the modified model also improves the fit of the academic citation networks.

Keywords: citation network; U.S. patents; stochastic generative model; preferential attachment; triad formation.

Acceleration of Computation in Fuzzy Clustering

Yuichi Mori[1], Tatsuki Akaki[2], Masahiro Kuroda[3]

[1] *Faculty of Management, Okayama University of Science, Japan,*
yuichi-mori@ous.ac.jp
[2] *Graduate School of Management, Okayama University of Science, Japan,*
m22mm39tf@ous.jp
[3] *Faculty of Management, Okayama University of Science, Japan,*
kuroda@ous.ac.jp

Abstract

The fuzzy c-means clustering (FCM) is one of useful soft clustering methods to search a reasonable form of clustering in which each data point belongs to multiple clusters. In FCM, alternating least squares (ALS) type computation is performed to estimate two parameters, the membership matrix and the cluster centroid matrix, alternatingly. Therefore, high computational cost is sometimes required due to the iterative convergence.

To reduce such computational cost, a general procedure to accelerate the iterative computation in ALS, called vector epsilon algorithm, can be used. This algorithm generates a new accelerated convergent sequence based on the original convergent sequence in estimating two or more parameters alternatingly, where the original sequence converges linearly. Since FCM generates a linearly convergent sequence for the membership and cluster centroid matrices, applying the vector epsilon algorithm to FCM provides faster computational results than the original FCM in terms of the number of iterations and CPU time.

The performance of the vector epsilon accelerated FCM is evaluated in simulations under various conditions such as data size, the number of variables, the number of original clusters, and the number of expected clusters. It is also evaluated with a couple of real data. These numerical experiments demonstrates that the vector epsilon accelerated FCM accelerates the computation almost twice as faster as the original one.

(A part of this study was conducted by Mr. Takatsugu Yoshioka in 2018, who was a graduate student of Informatics, Okayama University of Science.)

Keywords: soft clustering; alternating least squares; vector epsilon algorithm; iterative convergence; the number of iterations; CPU time.

Fuzzy Clustering based Support Vector Machine for Autocoding and Interpretation of Results

Yukako Toko[1], Mika Sato-Ilic[2]

[1]*National Statistics Centre, Japan, ytoko@nstac.go.jp*
[2]*Faculty of Engineering, Information and Systems, University of Tsukuba, Japan, mika@risk.tsukuba.ac.jp*

Abstract

This paper presents a new autocoding method for the coding task of governmental surveys. In addition, we propose a method to obtain the interpretation of the result of the proposed method.

In official statistics, text response fields are often found in survey forms. Coding tasks, translating text descriptions into corresponding classes, are usually performed to those respondents' text descriptions for efficient data processing. Coding tasks are originally performed manually, whereas the studies of automated coding have made progress with the improvement of computer technology in recent years. In particular, the data related to the Family Income and Consumption survey included text descriptions extracted from digital receipts, which have been getting larger and more complex in recent years.

Therefore, we have developed a classification method for autocoding. The purpose of this method is to obtain stable results of discrimination as coding with high generalization performance dealing with cognitive uncertainty for text description data. This method combines multi-class Support Vector Machine (SVM) by fuzzy c-means and the previously developed reliability score-based classification method. This method utilizes both SVM, a machine learning method known for high generalization performance, and the fuzzy c-means method in the area of computational intelligence, which is a part of Artificial Intelligence (AI) known for high performance in dealing with cognitive uncertainty. Also, the proposed method utilizes the previously developed classification method based on reliability scores, including the idea of fuzzy logic. This method shows better performance when compared with the previously proposed method for considering the complexity and large amount of data. However, in this method, the explainability of the results commonly occurring in artificial intelligence techniques has been an open question.

Therefore, we have developed a new method to explain the results by using the probability scores obtained by SVM using larger classification categories compared with the targeted categories of the classes. By using this, we can obtain the mixed properties for the single large category (or single class) with other properties of other large categories (or classes). This obtained feature can assist in the main result of the autocoding.

We will present several results of numerical examples using governmental survey data during our presentation.

Keywords: Autocoding; Computational Intelligence; Machine Learning; Support Vector Machine; Fuzzy c-Means.

Deep Learning Method for Imbalanced Image Data Considering Reduction of Dimensionality by Multidimensional Scaling

Shojiro Takamura[1], Mika Sato-Ilic[2]

[1]*Graduate School of Science and Technology, University of Tsukuba, Japan, s2120538@s.tsukuba.ac.jp*
[2]*Faculty of Engineering, Information and Systems, University of Tsukuba, Japan, mika@risk.tsukuba.ac.jp*

Abstract

The purpose of this study is to propose a new classification method for imbalanced image data. In recent years, the development of information technology has made it possible to obtain a large amount of data. One such data is image data which is effectively used in various fields. Deep learning methods such as convolutional neural networks (CNN) are known to perform high accuracy in the classification of such images. However, in general, image data is large in volume, and so the computational load for data processing is also large. Therefore, image classification using this method is computationally expensive to create a classifier that achieves reasonable accuracy. To overcome this problem, a method that incorporates Multidimensional Scaling (MDS) into deep learning has been proposed. This method uses MDS to obtain the value of coordinate in a low-dimensional space based on the classification results from a CNN, which is then used to re-predict. Although efficient for the reduction of computational loads, this method targets only data that are balanced for given classes. That is, the amount of data in the classes is almost the same, and its efficiency has not been verified for imbalanced data. On the other hand, in real-world data, the number of data belonging to each class is often not uniform. In addition, the problem of the bias for the classification of such imbalanced data is well-known, such as predictions for majority classes have a large impact on the results, while predictions for minority class have a small impact, so it is not possible to obtain valid results.

Therefore, the purpose of this study is to propose an image classification method that can be applied to real-world data while keeping the computational load low by introducing methods that can be applied to imbalanced data to the previously proposed method of incorporating MDS into deep learning. The SMOTE method can be applied to imbalanced data and is a well-known method to equalize the amount of data per class by synthesizing data for minority classes. Also used, IBLM ResNet is a recently developed deep learning model for imbalanced data are used. We also applied Safe-Level-SMOTE and MWMOTE, which were developed as improved SMOTE methods, and discussed the results. We also evaluate the results of several different optimizers that affect the efficiency and accuracy of deep learning loss function minimization. From several numerical examples, we can see that the proposed method is valid, and the proposed method significantly reduces the

number of dimensions and increase classification accuracy and data processing speed.

Keywords: Deep Learning; Multidimensional Scaling; Imbalanced Data; Image Data; Convolutional Neural Networks.

Nonparametric comparison of epidemic time trends: the case of COVID-19

Marina Khismatullina[1], Michael Vogt[2]

[1]*Erasmus School of Economics, Erasmus University Rotterdam*
[2]*Department of Mathematics and Economics, Ulm University*

Abstract

The COVID-19 pandemic has been one of the most pressing issues for the past two years. A question which was particularly important for governments and policymakers is the following: Does the virus spread in the same way in different countries? Or are there significant differences in the development of the epidemic? We devise a new inference method for detecting differences in the development of the epidemic time trends across countries. Specifically, it allows making simultaneous confidence statements about the regions where the trends differ. In the theoretical part, we prove that the method controls the familywise error rate, that is, the probability of wrongly rejecting at least one null hypothesis. In our empirical study, we use the method to compare the outbreak patterns of the epidemic in a number of European countries.

Seeding Large-scale Public Health Inventions in Multilayer Social Networks with Heterogeneous Nodes

Xiaoyi Han[1]

[1]*The Wang Yanan Institute for Studies in Economics, Xiamen University*

Abstract

Vaccination campaign is an important public health intervention to suppress the spread of infectious diseases. This paper studies the seeding strategies to select targeted states to boost vaccination rates to achieve the maximum reduction in COVID-19 confirmed cases in the United States. We consider two forms of spillovers across states: the peer effect of vaccination working through the Facebook friendship network and the diffusion of confirmed cases working through the travel flow network. We propose a threshold spatial dynamic panel data (TSDPD) method to model the temporal development and spatial dependence of confirmed cases, allowing multiple channels of vaccination effects on confirmed cases as well as heterogeneities based on population size. We compare the case reduction effects of four seeding strategies. Two network-informed strategies select the key players in the friendship and travel flow networks respectively as targets, while two heterogeneity-informed strategies select the states with the largest population and those with slowest vaccination respectively as targets. Different seeding strategies lead to different amounts and different spatial distributions of case reductions. The cost-benefit analysis reveals important tradeoffs by accounting for state differences in the vaccine shots needed for the campaign and per-case medical expenses.

Policy Effectiveness on the Global Covid-19 Pandemic and Unemployment Outcomes: A Large Mixed Frequency Spatial Approach

Yijiong Zhang[1], Xiaoyi Han[2], Yanli Zhu [3], and Ying Chen[4]

[1]Finance and Financial Risk Management Centre, NUS (Chongqing) Research Institute, China. Email: zhang.yijiong@u.nus.edu
[2]The Wang Yanan Institute for Studies in Economics (WISE), Xiamen University, China. Email: xiaoyihan@xmu.edu.cn
[3]Business School, Hohai University, China. Email: zhuyanli_921@126.com
[3]Department of Mathematics & Risk Management Institute, National University of Singapore, Singapore. Email: matcheny@nus.edu.sg. Corresponding author.

Abstract

We measure the effectiveness of policies in a mixed frequency spatial VAR (MF-SVAR) modeling framework, conditional on the spillover and diffusion effects of the global pandemic and unemployment. Specifically, we examine two aspects of policy effectiveness from a spatio-temporal perspective, namely the impact of policy start dates and policy timeliness on Covid-19 weekly new case growth and monthly changes in unemployment rates for 68 countries across six continents from January 2020 to August 2021. We find that government policies have a significant impact on the growth of new cases, but only a marginal effect on the change in unemployment rates. A policy's start date is critical for its effectiveness. In terms of both immediate impact on the near term and total impact over the following four weeks, starting a policy in the 4th week of a month is most effective at reducing the growth of new cases. At the same time, starting in the 2nd or 3rd week can be counterproductive. In addition, our estimates suggest that the spillover and diffusion effects are much stronger than a country's temporal effect during a global pandemic, both for new case growth and changes in unemployment. We also find that new case growth influences changes in unemployment, but not vice versa. Counterfactual experiments provide further evidence of policy effectiveness in various scenarios and also reveal the main risk-vulnerable and risk-spillover countries during pandemic.

Keywords: Policy effectiveness; Spatio-temporal model; Mixed frequency; Spatial panel data.

High-dimensional Quantiled Conditional Moments with Hypergraph

Zhoufan Zhu[1], Ningning Zhang[2], Ke Zhu[3]

[1]*Shanghai University of Finance and Economics, tylerzzf@163.sufe.edu.cn*
[2]*University of Hong Kong, xcxks1@connect.hku.hk*
[3]*University of Hong Kong, mazhuke@hku.hk*

Abstract

Modern portfolio selection requires a reliable estimate of the conditional mean, variance, skewness, and kurtosis of returns. When the number of assets is larger than the sample size, this constitutes a difficult estimation problem, especially for the higher-order conditional moments (variance, skewness, and kurtosis). In this paper, we propose a novel graph-based quantiled conditional moments (GQCM) model to estimate the conditional moments. In the first place, we inject the domain knowledge into a hypergraph, and apply the temporal graph convolution network to estimate the conditional mean and quantiles for asset returns. Next, we employ the quantiled conditional moments (QCM) method to estimate the higher-order conditional moments from conditional mean and quantiles. The proposed GQCM model takes the interdependence between assets into consideration by utilizing the hypergraph, and obtains the high-dimensional conditional moments without assuming any parametric forms. An application in NASDAQ and NYSE stock markets shows that the GQCM model performs much better than some benchmark models.

Keywords: Quantiled conditional moments; Graph Neural Network; High-dimensional time series.

Quantiled conditional variance, skewness, and kurtosis by Cornish-Fisher expansion

Ningning Zhang[1] , Ke Zhu[2]

[1]University of Hong Kong, xcxks1@hku.hk
[2]University of Hong Kong, mazhuke@hku.hk

Abstract

The conditional variance, skewness, and kurtosis play a central role in time series analysis. These three conditional moments (CMs) are often studied by some parametric models but with two big issues: the risk of model mis-specification and the instability of model estimation. To avoid the above two issues, this paper proposes a novel method to estimate these three CMs by the so-called quantiled CMs (QCMs). The QCM method first adopts the idea of Cornish-Fisher expansion to construct a linear regression model, based on n different estimated conditional quantiles that can be obtained without assuming any parametric forms of the CMs. Next, it computes the QCMs simply and simultaneously at each fixed timepoint by using the ordinary least squares estimator of this regression model. Under regular conditions, the QCMs are shown to be consistent with the convergence rate $n^{-1/2}$. Simulation studies indicate that the QCMs perform well under different scenarios of estimated conditional quantiles. In the application, the study of QCMs for eight major stock indexes demonstrates the effectiveness of financial rescue plans during the COVID19 pandemic outbreak, and unveils a new "non-zero kink" phenomenon in the "news impact curve" function for the conditional kurtosis.

Keywords: Conditional moments; Cornish-Fisher expansion; "News impact curve"; Quantile time series estimation; Quantiled conditional moments.

Inference for under-identified IV regression model

Kunyang Song[1]

[1]*The University of Hong Kong, skyhku@connect.hku.hk*

Abstract

Finding valid instrumental variables (IVs) is important but hard to deal with endogeneity in the linear regression model. In many cases, researchers may not have enough valid IVs to implement 2-stage least square estimator or may not even know whether the chosen IVs are valid. This paper first proposes a new martingale difference divergence (MDD) estimator, which is applicable to the under-identified IV regression model. Next, a new MDD-based test is constructed to examine whether all chosen IVs are valid. Moreover, a new MDD-based Hausman test is given to examine endogeneity, and it works for under-identified IV regression model. Under regular conditions that allow for dependent data, the asymptotics of all proposed estimator and tests are established. As an extension, this paper also studies the above inferential methods for the general nonlinear regression model. Finally, the importance of all of the proposed methods is illustrated by simulations and real examples.

Keywords: causal inference; instrumental variables; time series.

Forecasting High Dimensional Long Memory Time Series based on Memory augmented Gated Recurrent Unit (MGRU)

Maolin Yang[1], Muyi Li[1,2], Guodong Li[3]

[1]Wang Yanan Institute for Studies in Economics (WISE), Xiamen University, China , merlynyang546@gmail.com
[2]Department of Statistics and Data Science, School of Economics, Xiamen University, China, limuyi@xmu.edu.cn
[3]Department of Statistics and Actuarial Science, The University of Hong Kong, Hong Kong, China, gdli@hku.hk

Abstract

Modeling long-range dependency is one of the biggest challenges in time series analysis. Statistical models like the ARFIMA and HAR models can capture the long memory effect in time series, but they often suffer from the curse of dimensionality. In the meantime, recurrent neural networks like the Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) models are two popular models to approximate nonlinear structures in high dimensional data, and the latter one is more succinct. In this paper, we propose a new model by adding a fractionally integrated filter into the GRU's structure and call it the memory-augmented GRU (MGRU). We justify the existence of the long memory effect in the MGRU model. We demonstrate the outperformance of the proposed MGRU in real examples of language modeling and realized volatilities in stock markets, both of which are high dimensional time series with long memory properties.

Keywords: Long memory; Gated Recurrent Unit; Forecasting.

Predictive Subdata Selection for Large-Scale Deterministic Computer Models

Ming-Chung Chang[1]

[1]*Institute of Statistical Science, Academia Sinica, mcchang0131@gmail.com*

Abstract

Computer models are implementations of complex mathematical models using computer codes. Tremendous amounts of data generated from computer models are becoming ubiquitous owing to advanced technology. Such data richness, however, may yield an inability to conduct statistical analysis in terms of the time cost. Recently, increased attention has focused on solving this data reduction problem. In this talk, I will introduce a new subdata selection method for large-scale deterministic computer models. In addition to the geometry of the input space, the proposed method takes advantage of the information of the output values and adaptively updates the current subdata with affordable computational cost. Simulated examples and real data analyses are provided.

Keywords: Computer experiment; Subsampling; Data reduction.

A Survey on Multi-step Lookahead Bayesian Optimization

Peng Liu[1]

[1]*Singapore Management University, liupeng@smu.edu.sg*

Abstract

In recent years, Bayesian Optimization (BO) has received increasing attention due to its high sample efficiency in the global optimization of expensive-to-evaluate functions. It consists of two major components: a surrogate model to approximate the unknown objective function and provide posterior estimates, and an acquisition function to guide the sampling strategy by quantifying the expected marginal gain in the utility of the collected observations. Among multiple choices of acquisition function, a particular research direction focuses on the multi-step lookahead policy that incorporates multiple future evolutions into the decision-making at the current time step. Compared with common one-step acquisition functions such as expected improvement and knowledge gradient, the multi-step lookahead acquisition function enjoys the nice property of being nonmyopic, in the sense that any decision taken at the current step bears an impact on all future decisions. Following Bellman's principle of optimality, the multi-step lookahead policy assumes a Dynamic Programming (DP) formulation that consists of nested maximization and expectation operations in order to deliver the best average-case return in the long run. In this paper, we introduce the fundamentals of modeling the multi-step lookahead acquisition function as a Markov decision process in the BO setting, provide a comprehensive review of the current state of this less well-studied but promising research direction, and visit several main computational heuristics to approximate the solution of the intractable DP problem.

Keywords: Bayesian optimization; dynamic programming; multi-step lookahead.

Generalized Bayesian D-optimal supersaturated multistratum designs

Chang-Yun Lin[1]

[1]*Department of Applied Mathematics and Institute of Statistics, National Chung Hsing University, Taichung, Taiwan, 40227, chlin6@nchu.edu.tws*

Abstract

Supersaturated designs are useful in the initial stage of experiments to identify important factors from many of interest with a small number of runs. Traditional supersaturated designs were mainly constructed for completely randomized experiments, which have single-stratum structures. They cannot be used for experiments that have multistratum structures, such as the split-plot, strip-plot, and staggered-level experiments. How to construct supersaturated multistratum designs for complex experiments has gained much attention recently. In this paper, we consider the situation in which the experimenters have prior knowledge of which factors are more likely to be important (called the primary factors) than the others (called the potential factors). By taking primary and potential factors into account, we propose an approach using the generalized Bayesian D (GBD) criterion to construct a new class of supersaturated multistratum designs. The GBD-optimal supersaturated multistratum designs provide guidelines on how to assign factors to the designs, which enhances efficiency on identifying active factors. A case study shows that the proposed supersaturated design (32 runs with 19 factors) is as effective as the full 26 factorial design (64 runs with 6 factors) to identify important factors in a battery cell experiment.

Keywords: Potential terms; Primary terms; Split-plot; Staggered-level; Strip-plot.

Determination of the effective cointegration rank in high-dimensional time-series predictive regression

Puyi Fang[1], Zhaoxing Gao[2], Ruey S. Tsay[3]

1School of Economics, Zhejiang University, puyifang@zju.edu.cn
2Center for Data Science, Zhejiang University, zhaoxing gao@zju.edu.cn
3Booth School of Business, University of Chicago, ruey.tsay@chicagobooth.edu

Abstract

This paper proposes a new approach to identifying the effective cointegration rank in high-dimensional unit-root (HDUR) time series from a prediction perspective using reduced-rank regression. For a HDUR process x and a stationary series $y \in Rp$ of interest, our goal is to predict future values of y using the cointegrated variables of x and lagged values of yt. The proposed framework consists of a two-step estimation procedure. First, the well-known method of Principal Component Analysis (PCA) is used to identify all cointegrating vectors of x. Second, the transformed stationary series obtained via the cointegrating vectors are used as regressors, together with some lagged variables of y, in predicting y. The estimated reduced-rank is then defined as the effective coitegration rank of x. Under the scenario that the autoregressive coefficient matrices are sparse (or of low-rank), we apply the Least Absolute Shrinkage and Selection Operator (LASSO) (or the reduced-rank techniques) to estimate the autoregressive coefficients when the dimension is high. Theoretical properties of the estimators are established under the assumptions that the dimensions p and N and the sample size $T \to \infty$. Both simulated and real data examples are used to illustrate the proposed framework, and the empirical application suggests that our procedure provides satisfactory performance in predicting stock returns.

Keywords: Cointegration; Factor model; Reduced-rank; High-dimension; LASSO.

Quantile index regression

Yingying Zhang[1], Yuefeng Si[2] , Guodong Li[2] and Chil-Ling Tsai[3]

[1]East China Normal University, China
[2]University of Hong Kong, Hong Kong, China
[3]University of California at Davis, USA

Abstract

Estimating the structures at high or low quantiles has become an important subject and attracted increasing attention across numerous fields. However, due to data sparsity at tails, it usually is a challenging task to obtain reliable estimation, especially for high-dimensional data. This paper suggests a flexible parametric structure to tails, and this enables us to conduct the estimation at quantile levels with rich observations and then to extrapolate the fitted structures to far tails. The proposed model depends on some quantile indices and hence is called the quantile index regression. Moreover, the composite quantile regression method is employed to obtain non-crossing quantile estimators, and this paper further establishes their theoretical properties, including asymptotic normality for the case with low-dimensional covariates and non-asymptotic error bounds for that with highdimensional covariates. Simulation studies and an empirical example are presented to illustrate the usefulness of the new model.

Keywords: Asymptotic normality; High-dimensional analysis; Non-asymptotic property; Partially parametric model; Quantile regression.

Network Autoregression for Incomplete Matrix-Valued Time Series

Xuening Zhu[1], Feifei Wang[2], Zeng Li[3], Yanyuan Ma[4]

[1]*Fudan University, xueningzhu@fudan.edu.cn*
[2]*Renmin University of China, feifei.wang@ruc.edu.cn*
[3]*Southern University of Science and Technology, lizeng124@gmail.com*
[4]*The Pennsylvania State University, yanyuanma@yahoo.com*

Abstract

We study the dynamics of matrix-valued time series with observed network structures by proposing a matrix network autoregression model with row and column networks of the subjects. We incorporate covariate information and a low rank intercept matrix. We allow incomplete observations in the matrices and the missing mechanism can be covariate dependent. To estimate the model, a two-step estimation procedure is proposed. The first step aims to estimate the network autoregression coefficients, and the second step aims to estimate the regression parameters, which are matrices themselves. Theoretically, we first separately establish the asymptotic properties of the autoregression coefficients and the error bounds of the regression parameters. Subsequently, a bias reduction procedure is proposed to reduce the asymptotic bias and the theoretical property of the debiased estimator is studied. Lastly, we illustrate the usefulness of the proposed method through a number of numerical studies and an analysis of a Yelp data set.

Keywords: Bias reduction; Incomplete matrix observations; Matrix-valued time series; Network autoregression.

A Semiparametric Approach to Empirical Bayes Estimation of Discrete False Discovery Rates Using Kernels

DAYTA, Dominic B.[1], ESCALANTE, Sean Oliver M.[2]

[1]*University of the Philippines, dbdayta@up.edu.ph*
[2]*Asian Institute of Management, SEscalante.MSDS2023@aim.edu*

Abstract

Advances in computing technology have given birth to more complex inferential problems, chief among them the problem of testing m hypotheses simultaneously, with m normally running to the thousands, especially in genomics. In this paper, we estimate the local false discovery rate via a modified version of the semiparametric kernel density estimator for the mixture distribution. Our method is built specifically to handle tests with discrete distributions. Via a series of simulation studies, we show that the method effectively avoids the loss of power that have been demonstrated to be a common result of using methods designed with the continuous paradigm in mind. Measured against Efron's Empirical Bayes method, the proposed semiparametric kernel-based estimator, despite performing slightly worse in terms of estimation error, exhibits significant boosts in terms of rejection power, while keeping overall false discovery rates (FDR) controlled. An application on breast cancer data confirms this result.

Keywords: Multiple Hypothesis Testing; Large-Scale Inference; False Discovery Rates; Empirical Bayes; Kernel Density Estimation.

Implementation of SVIR Mathematical Model in Simulating Covid-19 Infection Dynamics under Vaccination Intervention

Eduardo C. Dulay Jr.[1], Dr. Bernhard Egwolf.[2], Adrian Roel P. Lapus[3], Ian D. Marco[4]

[1]*Deparment of Mathematics and Physics, University of Sto. Tomas, ecdulay@ust.edu.ph*
[2]*Deparment of Mathematics and Physics, University of Sto. Tomas, begwolf@ust.edu.ph*
[3]*Deparment of Mathematics and Physics, University of Sto. Tomas, adrianroel.lapus.sci@ust.edu.ph*
[4]*Deparment of Mathematics and Physics, University of Sto. Tomas, ian.marco.sci@ust.edu.ph*

Abstract

The emergence of the Covid-19 pandemic has brought many challenges to the Philippines especially to its medical sector. Mathematical models have historically played a critical role in epidemiology in preventing pandemics from further running rampant. The SVIR (Susceptible, Vaccinated, Infectious, and Removed) model is a chamber-type mathematical model that simulates the daily changes in its four categories for a specified period of time. In this study, the SVIR model, using the Python programming language, was used to forecast the Susceptible, Vaccinated, Infectious, and Removed categories based on Philippine data from August to September 2021. The minimize function from Python's SciPy module was used to obtain the optimal rates to be used in forecasting the course of the pandemic. When compared to actual data, results indicate that the SVIR model was able to simulate the Vaccinated and Removed category within the 10% mean absolute percentage error (MAPE) acceptability threshold, while the Infectious category slightly exceeds this margin with an obtained MAPE of 19.59%. Overall, the SVIR model is a feasible model in forecasting the nature of the Covid-19 pandemic under vaccination intervention; however it is not ideal to use it in forecasting longer time periods.

Keywords: Covid-19; SVIR Model; Forecasting; Optimization; Vaccination

AI-based Scoring for Picture-cued Writing Assessment

Yipeng Zhuang[1], Ruibin Zhao[2], Zhiwei Xie[3], Philip L.H. Yu[4]

[1]*The Education University of Hong Kong, s1136220@s.eduhk.hk*
[2]*The Education University of Hong Kong, zhaor@eduhk.hk*
[3]*The Education University of Hong Kong, xiez@eduhk.hk*
[4]*The Education University of Hong Kong, plhyu@eduhk.hk*

Abstract

Grading assignments is inherently subjective and time-consuming, in response to an ongoing need for innovative assessment in language learning, researchers and educators are increasingly turning to computer-based approaches, especially automated writing assessment. These automatic scoring tools can greatly reduce teacher workloads and shorten the time needed for providing feedback to learners. We propose an AI-based method for automatically scoring student responses to picture-cued writing tasks. As a popular paradigm for language learning and assessment, a picture-cued writing task is typically to ask students to describe a picture or pictures, which means our automatic scoring method must be able to measure the link(s) between visual pictures and their textual descriptions. For this purpose, we designed a picture-cued writing test in our study and collected nearly 4k responses by recruiting 250 k12 students to participate in the writing test. We then developed our AI scoring model by incorporating the visual tokens (image features) and word tokens to construct a transformer-based multimodal encoder. Finally, the performance of the model was evaluated carefully with a small mean absolute error of 0.39 and a high adjacent-agreement rate of 92.6%, demonstrating an accurate scoring result. We believe this method could provide timely, effectively and accurate feedback to students, reduce the subjective elements inherent in human-centred grading and save teachers' time from the mundane task of grading.

Keywords: AI-based scoring; Deep learning; Language assessment.

Spatiotemporal Modeling of Sparse Geostatistical Epidemic Data

Ocampo, Shirlee R.[1], Barrios, Erniel B.[2]

[1] *De La Salle University, Manila, Philippines, shirlee.ocampo@dlsu.edu.ph*
[2]*School of Statistics, University of the Philippines Diliman, ebbarrios@up.edu.ph*

Abstract

Many spatiotemporal processes such as epidemic data, signal spectrum, and channel frequencies, weather and environmental occurences are often sparse. COVID-19 data in the Philippines are messy and sparse due to many gaps of underreported and unreported cases. Modeling sparse data has been a challenge since the sparse features increase the space and time complexity of models. The study postulated spatial autoregression (SAR) and spatiotemporal epidemic models (STEM) linking sparse geostatistical COVID-19 incidence and mortality rates to the healthcare system, demographic and economic indicators, disease prevalence, vaccination, urbanity, and environmental factors is proposed. It further integrated sparsity regularization terms in the spatiotemporal models. Sparse spatial weight matrices were applied for SAR model estimation, while the STEM parameters were estimated using a hybrid of Cochranne-Orcutt procedure and backfitting algorithm. Moreover, hurdle models were explored. The predictive abilities of the models were compared using mean square error and mean absolute error. The postulated models were applied to daily COVID-19 cases and deaths across provinces and cities in the National Capital Region (NCR) from April 1, 2020 to September 15, 2021 in the Philippines. Results show significant association of COVID-19 prevalence rate with number of health workers, revenue of the local government unit (LGU) which are provinces or cities in NCR, and prevalence of tuberculosis, and COVID-19 mortality rates with number of health workers, seniors, males, number of licensed COVID-19 testing laboratories, number of cities in an LGU, revenue of the LGU, prevalence rates of cancer, tuberculosis, cardiovascular disease, and diabetes. The models emphasize the importance of resources available in the local government that can boost the capabilities of the health care system. Pre-existing health conditions (comorbidities) of the communities also determine prevalence and mortality rates of COVID-19 in the Philippines.

Keywords: sparse data, spatiotemporal model, backfitting algorithm, COVID-19, spatial autoregression.

Incorporating Structural Change into Modeling of COVID-19 in the Philippines: A Spatiotemporal Model

Tresvalles, Regina M.[1], Barrios, Erniel B.[2]

[1]*De La Salle University, regina.tresvalles@dlsu.edu.ph*
[2]*School of Statistics, University of the Philippines Diliman, ebbarrios@up.edu.ph*

Abstract

This paper proposes the use of the spatiotemporal model with structural change in modelling the prevalence rate of the coronavirus disease 2019 (COVID-19), showing its suitability in the pandemic experience in the Philippines. The model that 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 temporary surge in the estimation of covariate and spatial parameters. The first two methods in this paper made use of the original procedure proposed by Bastero and Barrios in 2011, one without the structural change and the other one with structural change. The last method incorporated the use of the infection rate from the susceptible to the infectious state in the compartmental model Susceptible-Infectious-Quarantined-Hospitalized (SIQH) type in integrating the term for structural change. The method that use the effect of structural change offer good model fit especially in COVID-19 experience in the Philippines where structural change is encountered in different regions at different timepoints.

Keywords: spatiotemporal model, COVID-19, forward search algorithm.

Nonparametric Density-based Procedure of Detecting Emerging Events from Social Media

Louise Adrian DC. Castillo[1] , Erniel B. Barrios[2]

[1]*University of the Philippines Diliman, louiseadrian.castillo@gmail.com*
[2]*School of Statistics, University of the Philippines Diliman, ebbarrios@up.edu.ph*

Abstract

Detecting emerging events early enough is key to mitigation, prevention and containment of events especially those related to disease outbreaks, natural disasters, public safety, among others. Most tests for goodness of fit focuses on the central location of the density function, missing to detect changes that usually occur at the tails of the distribution. Oftentimes, these methods are prone to false negative results. We proposed two methods of discovering emerging events based on nonparametric density estimates from a data-generating process reflecting those in social media text data. The first method compares the percentile values between the baseline and speculated distribution while the second algorithm modifies the Kolmogorov Smirnov test to focus on the upper tails of the distribution. Simulation studies suggest that both algorithms can detect emerging events specially in cases with large number of data points.

Keywords: emerging events; nonparametric methods; test for goodness of fit; density estimation; tail events.

Combining measures of signal complexity and deep learning in medical images for neurodegenerative disease screening

Yi-Ju Lee[1], Chun-houh Chen[2]

[1]*Institute of Statistical Science, Academia Sinica, yijulee@stat.sinica.edu.tw*
[2] *Institute of Statistical Science, Academia Sinica, cchen@stat.sinica.edu.tw*

Abstract

Introduction
Advanced analytics in complexity science has been established to efficiently extract the crucial health-related information from high dimension data, which provides novel interdisciplinary methods to understand brain physiology. The well-validated power law scaling in physics has been used to understand the complexity of brain signals at temporal scales, which represents the dynamical properties. Deep learning models such as the VGG network has shown potential in medical image analysis for region detection and classification. This research aims to integrate and evaluate the cross-scale property of power law in fMRI as a data transformation step to advance the current AI models in identifying Alzheimer's Disease (AD).

Participant
The participants with MRI images, demographic and clinical data were selected from Alzheimer's Disease Neuroimage Initiative (ADNI) cohort. Functional and structural brain imaging data of 100 age and sex-matched, right-handed AD patients (age mean = 67.30 ± 16.64; male = 49.5%) and 100 cognitively normal elderly subjects (age mean = 67.22±13.41) were retrieved. The research was approved and followed the ADNI data use agreement.

Data Preprocessing
Functional images were then corrected for slice timing, realigned and co-registered to anatomical images. To extract the brain signal complexity, Fourier transform was used to quantify the power spectral density (PSD) of fMRI signal We estimated power law scaling by determining the slope of regression line fitting to the log-log plot of PSD. The power law scaling of brain activity was transformed into a heatmap for model training.

Statistical Method
The images were split randomly in the ratio of 8:1:1 for the training, testing and validation data set. The classification models were conducted with 3D VGG-16 using ReLU as the activation function. The experiments were conducted with Python 3.8 and run on two NVIDIA A100 (40G) GPU.

Results
The power law transformed image data shows significantly decreased training time, averagely from 168 mins to 31 mins. The average validation accuracy and sensitivity have shown no difference. The validation sensitivity was increased from 0.67 to 0.76. with Bonferroni correction ($p < 0.0005$) for multiple comparisons, both

the superior medial frontal gyrus and left precuneus were found as the region showing abnormal brain signals.

Conclusions

This research suggests evidence of advancing deep learning in neuroscience by applying the complexity method as a regular image data down-sampling step. Adequately used, the model with power law transformation performs significantly reduced computation time with similar classification results. From biological prospect, our result suggests the altered pathological hemodynamics of AD representing in the frequency domain, losing brain signal complexity, plays a role in deep learning-based extracted features. From the statistical prospect, incorporating deep learning and complexity can be practical and improve efficiency in clinical diagnosis. The future work would focus on applying complexity index in cooperating classification and segmentation models to further identify the key brain regions to classify disease.

Keywords: Power Law Scaling; Alzheimer's Disease; Deep Learning

Variable selection and estimation for misclassified responses and high-dimensional error-prone predictors

Li-Pang Chen[1]

[1]*Department of Statistics, National Chengchi University, Taipei, Taiwan (ROC)., lchen723@nccu.edu.tw*

Abstract

Binary classification has been an attractive topic in statistical analysis or supervised learning. To model a binary response and predictors, logistic regression models or probit models are perhaps commonly used approaches. However, because of the rapid growth of the dimension of the data as well as the non-ignorability of measurement error in responses and/or predictors, data analysis becomes challenging and conventional methods are invalid. To address those concerns, we propose a valid inferential method to deal with measurement error and handle variable selection simultaneously. Specifically, we primarily consider logistic regression models or probit models, and propose corrected estimating functions by incorporating erroreliminated responses and predictors. After that, we develop the boosting procedure with corrected estimating functions accommodated to do variable selection and estimation. Through numerical studies, we find that the proposed method accurately retains informative predictors as well as gives precise estimators, and its performance is generally better than that without measurement error correction.

Keywords: Binary data; boosting; error elimination; measurement error; regression calibration.

Improve the Performance of Deep Learning Algorithms by Supervised Dimension Reduction Methods

Han-Ming Wu[1]

[1]*Department of Statistics, National Chengchi University, wuhm@g.nccu.edu.tw*

Abstract

There is wide research and application interest in deep learning algorithms such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Generative Adversarial Networks (GANs), and autoencoders. Various approaches have been proposed to improve the performance of these algorithms, including dimension reduction (DR) methods. One of them, Principal Component Analysis (PCA), is a commonly used unsupervised method which acts as a feature extractor. In this study, we investigate the sliced inverse regression (SIR) technique, a supervised DR approach, to assist DL algorithms by incorporating class information into their modeling processes. Through the application of SIR+DL to real-world datasets such as images, we conduct several comparison studies for classification problems. Performance is measured by accuracy, F1 score, and AUC. According to the results, the proposed method can be useful in DL under certain circumstances.

Keywords: Convolutional neural networks, Generative adversarial networks, Recurrent neural networks, Sliced inverse regression.

Fast and Robust Sparsity Learning over Networks: A Decentralized Median Regression Approach

Xiaojun Mao[1]

[1]*Shanghai Jiao Tong University, China*

Abstract

Decentralized sparsity learning has attracted a significant amount of attention recently due to its rapidly growing applications. To obtain the robust and sparse estimators, a natural idea is to adopt the non-smooth median loss combined with a ℓ_1 sparsity regularizer. However, most of the existing methods suffer from slow convergence performance caused by the {\em double} non-smooth objective. To accelerate the computation, in this paper, we proposed a decentralized surrogate median regression (deSMR) method for efficiently solving the decentralized sparsity learning problem. We show that our proposed algorithm enjoys a linear convergence rate with a simple implementation. We also investigate the statistical guarantee, and it shows that our proposed estimator achieves a near-oracle convergence rate without any restriction on the number of network nodes. Moreover, we establish the theoretical results for sparse support recovery. Thorough numerical experiments and real data study are provided to demonstrate the effectiveness of our method.

Robust Inference for Change Points in High Dimension

Feiyu Jiang[1]

[1]*Fudan University, China*

Abstract

This paper proposes a new test for a change point in the mean of high-dimensional data based on the spatial sign and self-normalization. The test is easy to implement with no tuning parameters, robust to heavy-tailedness and theoretically justified with both fixed-n and sequential asymptotics under both null and alternatives, where n is the sample size. We demonstrate that the fixed-n asymptotics provide a better approximation to the finite sample distribution and thus should be preferred in both testing and testing-based estimation. To estimate the number and locations when multiple change-points are present, we propose to combine the p-value under the fixed-n asymptotics with the seeded binary segmentation (SBS) algorithm. Through numerical experiments, we show that the spatial sign based procedures are robust with respect to the heavy-tailedness and strong coordinate-wise dependence, whereas their non-robust counterparts proposed in Wang et al. (2022) appear to under-perform. A real data example is also provided to illustrate the robustness and broad applicability of the proposed test and its corresponding estimation algorithm.

A Partially Functional Linear Modeling Framework for Integrating Genetic, Imaging, and Clinical Data

Ting Li[1]

[1]*Shanghai University of Finance & Economics, China*

Abstract

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

The Development of an Annotation Guidelines for Medication-related Incident Reports

Zoie S.Y. Wilkins-Wong[1], Ryohei Sasano[2], Shin Ushiro[3], Neil Waters[4]

[1] *Graduate School of Public Health, St. Luke's International University, zoiesywong@gmail.com*
[2] *Nagoya University, sasano@i.nagoya-u.ac.jp*
[3] *Japan Council for Quality Health Care (JQ), ushiro.shin.161@m.kyushu-u.ac.jp*
[4] *Graduate School of Public Health, St. Luke's International University, nawaters3@gmail.com*

Abstract

Background: Medication errors are a frequent cause of patient harm, so much so that the World Health Organization has declared 'Medication Without Harm' as the third Global Patient Safety Challenge. Although reporting systems have collected millions of incident reports, synthesising these narrative, free-text reports into actionable data that can inform best practice and reduce medication errors remains challenging. Natural language processing and AI offer a solution: models, once trained, can autonomously capture key information from reports, providing a standardised synthesis of data that can guide learning and reduce medication errors. Incident reports that have been accurately and systematically annotated, i.e., gold standards, are needed to train these models.
Materials and Methods: Here, we present a set of guidelines for the annotation of incident reports of medication errors. By demonstrating our annotation ontology to others, we hope to spur the creation of more gold-standard datasets and further the development of data extraction models.
Results: The guidelines first detail the basis for our annotation framework: an extensive review of medication errors, classification schemes and annotation methodologies. Second, the various named entities to be extracted are described: 'drug', 'form', 'strength', 'duration', 'timing', 'frequency', 'date', 'dosage', 'route' and 'wrong patient'. Third, attributes are introduced, which provide key information specific to certain named entities. Attributes further categorise the data, e.g., how named entities relate to each other. The last section covers how the annotations from the previous sections are systematically interpreted to categorise incidents by type. Throughout the guidelines, real-life incident reports are used as illustrative examples. Training materials, aimed at providing an accessible introduction to our annotation framework, are also described.
Conclusion: Directions for future work are detailed, namely the expansion of our annotation framework to include two key tasks. The detection of negation, e.g., 'antibiotics were not administered', is needed to correctly determine an entity's presence or absence. The ability to infer temporal order, e.g., 'patient A reported symptoms prior to the administration of drug A', has clear clinical implications, such as determining whether the administration of a drug was an error or in response to an error. We aim to design framework to be usable in other languages, including

English. This is to further promote the creation of datasets of manually annotated incident reports. Ultimately, our aim is to revolutionise the way we learn from past medication errors and improve patient safety.

Keywords: annotation guidelines; ontology for mediation-related incident reports; gold standard; incident report; patient safety.

Active Learning Framework for Clinical Named Entity Recognition based on Transformers and Transfer Learning

Jiaxing LIU[1], Zoie SY Wong[2]

[1]*School of Statistics and Mathematics, Zhongnan University of Economics and Law, jxliu@zuel.edu.cn*
[2]*Graduate School of Public Health, St. Luke's International University, zoiesywong@gmail.com*

Abstract

Objectives: Clinical named entities recognition (CNER) is one of the fundamental tasks to understand and extract knowledge from clinical texts. Transformers, which have achieved prominent performance in this task, usually require large annotated training datasets. However, annotation efficiency and costs have been identified as a problem especially for clinical text. Active learning (AL) selects a batch of informative unannotated samples to query annotators and train the CNER system iteratively with newly annotated datasets to reduce the annotation cost. We aim to investigate the effectiveness and annotation costs of active learning strategies with transformers for CNER.

Materials and Methods: We conducted experiments on the 2018 n2c2 Track 2 dataset under the AL framework based on ClinicalBERT. We examined three types of AL selection strategies including: uncertainty-based methods (LC, MTP), diversity-based methods (CLUSTER) and hybrid methods (HLC, HMTP) and compared them with random selection strategy. The experiments were initialized with 1% data as the annotated set. Each AL selection strategy was then employed to screen the free text of the remaining data. In each iteration, the selected additional 2% of the dataset was then added into the annotated set. ClinicalBERT was finetuned on the entire annotated set in each run. The experiments were performed with 25 AL iterations and were repeated for five times. The overall performances on a separated test set were compared using mean micro-averaged F1 measures. Mean tokens reviewed and corrections made by annotators were used to evaluate the annotation costs.

Results: Our experiment showed that all AL strategies outperformed random selection in terms of micro-averaged F1 measures after 25 iterations (MTP: 0.9315; HLC: 0.9312; HMTP: 0.9310; LC: 9296; CLUSTER: 0.9278; random: 0.9278) and can achieve 99% of the baseline performance that trained from the entire dataset within 7 iterations using 15% of the entire dataset. In order to achieve 99% baseline performance, LC and MTP required 14.4% of all tokens reviewed by annotators, while random selection required 22.8%. However, random selection required only 4,800 token-level corrections to achieve the 99% performance, while all other AL strategies required more corrections (CLUSTER: 4,898; LC: 22,040; MTP: 14,818; HLC: 18,764 and HMTP: 13,629).

Conclusion: In the experiments, the recommended AL selection strategy varies depending on the way of annotation costs evaluation and performance expectation.

To minimize the tokens reviewed by annotators, LC and MTP are recommended. To minimize the corrections made by annotators, random and CLUSTER are recommended. To reach a better performance when the annotation budget is relatively adequate, MTP and HLC are recommended.

Keywords: Active Learning; Named Entity Recognition; Clinical Text; Medication Extraction; Artificial Intelligence.

Development of Multitask Incident Reports-pretrained BERT Model to Empower Incident Reporting and Learning System

Zoie SY Wong[1], Dentsu Data Artist Mongol Team[2]*

[1]St. Luke's International University, zoiesywong@gmail.com
*[2]Dentsu Data Artist Mongo*Team members: Batmunkh Batsaikhan, Telmuun Enkhbold, Khatanbold Batzorig, Od Ganzorig, od@mn.data-artist.com.*

Abstract

Background: The unstructured nature of incident reports and lack of a systematic approach to register incidents pose tremendous challenges in respect to incident report learning. Natural language processing shows promise in clinical decision-making and can potentially assist in overcoming this research gap. Our study advances a BERT model in Japanese language to automatically structure free text incident reports and enable subsequent incident learning.

Methods:
Using the developed annotation methods for medication-related incident reports, we investigated a hybrid rule-based and transformer-based approach to enhance incident report named entity recognition and relation extraction. We developed NLP rules to identify named entities of drug name, form, route, strength, frequency, date, dosage, time, and duration using the Japanese drug name lexicon, gold standard data unique list, and regular expression using RegEx. Based on a pre-trained BERT on Japanese Wikipedia & Twitter data, we further pre-trained the model using a large corpus of unlabelled incident reports (>120,000 reports) and carried out two phases of finetuning for named entity and relation identification using the rule-based annotated data and gold standard data.

Results: Ultimately, one single BERT model consisting of three densely layered models was developed. The best 3-layer BERT model identifies NEs with F-1 score of 0.97 and 0.84 in the validation and testing sets respectively. The model identified whether an NE isn't error indication (i.e. condition of intended and actual) at F-1 score of 0.87. The developed model increased F1 score by 15% (comparing with rule-based model). We also experimented with a content-based recommender engine using SpaCy cosine similarity score to match learning resources for incident reports and deployed the models into digital platform.

Conclusion: Our BERT model has shown promising results in NER and relation extraction tasks using incident reports. We plan to assess the suitability of the clinical and generic language models (such as ClinicalBERT, GPT-J large language models (LLMs), and Google Pathways system), and rescale candidate language models to enhance drug NER and information extraction capabilities. In the future, the team plans to expand explore vector-based recommender engine and graph learning based methods to enable library information retrieval, and enhance the

decision support modules thorough integrating dashboard and graphical visualizations, in order to report spatial and temporal trends of incident occurrence and summarise key statistics.

Keywords: Incident Reports; BERT; rule-based; Patient Safety; Learning Health System.

Exploratory Research on Use of New Data for Computation of Private Housing Rent Index

Helen Y.W. Lee[1], Benjamin C.W. Cheung[2], Daniel W.L. Chong[3], Natalie K.P. Chung[4], Kevin S.Y. Hsia[5]

[1]*Census and Statistics Department, Hong Kong, China, hywlee@censtatd.gov.hk*
[2]*Census and Statistics Department, Hong Kong, China, bcwcheung@censtatd.gov.hk*
[3]*Census and Statistics Department, Hong Kong, China, dwlchong@censtatd.gov.hk*
[4]*Census and Statistics Department, Hong Kong, China, nkpchung@censtatd.gov.hk*
[5]*Census and Statistics Department, Hong Kong, China, ksyhsia@censtatd.gov.hk*

Abstract

The private housing rent index, a component index of the Consumer Price Index (CPI) of Hong Kong, China, measures the price changes of private housing rent over time. Under the existing practice, it is compiled based on data from two main sources, viz. (i) the Rent Survey (RS) conducted along with the monthly General Household Survey, and (ii) the estimated rents obtained from Rating and Valuation Department for newly-let cases identified in RS.

In the past few years, the normal conduct of household surveys had somewhat been disrupted by the COVID-19 pandemic, creating uncertainty in the availability of timely survey data. In view of the above, this research examined the feasibility of using a new data source, viz. rental transaction records available on the websites of two major property leasing agents in Hong Kong, and data analytics models to compile an alternative private housing rent index.

Keywords: Private housing rent index; Consumer Price Index; Data analytics, Web scraping; Household survey.

Automatic Coding of Occupations in General Household Survey (GHS) of Hong Kong

Ronald C.H. Chan[1], Chris C.W. Leung[2], Sharon P.W. Ng[3],
Frances C.W. Wong[4], Matthew T.L. Wong[5]

[1]*Census and Statistics Department, Hong Kong, China, rchchan@censtatd.gov.hk*
[2]*Census and Statistics Department, Hong Kong, China, cwleung@censtatd.gov.hk*
[3]*Census and Statistics Department, Hong Kong, China, spwng@censtatd.gov.hk*
[4]*Census and Statistics Department, Hong Kong, China: cwwong3@censtatd.gov.hk*
[5]*Census and Statistics Department, Hong Kong, China, mtlwong@censtatd.gov.hk*

Abstract

The General Household Survey (GHS) of Hong Kong is a monthly household survey for compiling employment and unemployment statistics. Traditionally, occupation coding in GHS is mostly done manually by a team of coders, who convert the word descriptions (e.g. job title) to an occupation code according to a set of occupation classification. Since manual coding is expensive and time-consuming, it is more desirable to code the textual data automatically. In this study, different data analytics algorithms were developed for the task using the manually coded occupation data available from GHS in 2017-2021 for model training. By evaluating the performance and quality of the models, the feasibility of adopting such models occupation coding in practical environment has been assessed.

Keywords: Automatic coding; Occupation coding; Data analytics; Labour force survey; Household survey.

Anomaly Detection in Merchandise Trade Data: A Deep Learning Approach

Benjamin C.H. Chan[1], Natalie K.P. Chung[2], Ian Y.C. Ng[3]

[1]*Census and Statistics Department, Hong Kong, China, bchchan@censtatd.gov.hk*
[2]*Census and Statistics Department, Hong Kong, China, nkpchung@censtatd.gov.hk*
[3]*Census and Statistics Department, Hong Kong, China, iycng@censtatd.gov.hk*

Abstract

Merchandise trade statistics of Hong Kong, China are compiled by the Census and Statistics Department (C&SD) based on trade declarations lodged by importers/exporters within 14 days after the shipment. While the post-shipment declaration arrangement facilitates trade activities in general, the clarity and accuracy of data declared in relevant trade documents might not be as high as that of similar documents in other places serving customs clearance purposes. To safeguard the quality of trade statistics, C&SD has been adopting for many years a rule-based computer validation system based mainly on numeric data to identify trade declarations that are most susceptible to having reporting errors. This talk presents the results of C&SD's exploration in the application of big data analytics in this quality checking mechanism. It has been shown that the use of deep learning models based on both unstructured textual information and numeric data could bring about impressive improvements in terms of both accuracy and efficiency.

Keywords: Merchandise trade statistics; Data quality checking; Deep learning; Text analytics; Anomaly detection.

HAR-It^o models and high-dimensional HAR modeling for high-frequency data

Huiling Yuan[1], Yifeng Guo[2], Kexin Lu[3], Guodong Li[4]

[1]*The University of Hong Kong, huilyuan@hku.hk*
[2]*The University of Hong Kong, gyf9712@connect.hku.hk*
[3]*The University of Hong Kong, neithen@connect.hku.hk*
[4]*The University of Hong Kong, gdli@hku.hk*

Abstract

It is an important task to model realized volatilities for high-frequency data in finance and economics and, as arguably the most popular model, the heterogeneous autoregression (HAR) model has dominated the applications in this area. However, this model suffers from three drawbacks: (i.) its heterogeneous volatility components are linear combinations of daily realized volatilities with fixed weights, which limit its flexibility for different types of assets, (ii.) it is still unknown what is the high-frequency probabilistic structure for this model, as well as many other HAR-type models in the literature, and (iii.) there is no high-dimensional inference tool for HAR modeling although it is common to encounter many assets in real applications. To overcome these drawbacks, this paper proposes a multilinear low-rank HAR model by using tensor techniques, where a data-driven method is adopted to automatically select the heterogeneous components. In addition, HAR-It^o models are introduced to interpret the corresponding high-frequency dynamics, as well as those of other HAR-type models. Moreover, non-asymptotic properties of the high-dimensional HAR modeling are established, and a projected gradient descent algorithm with theoretical justifications is suggested to search for estimates. Theoretical and computational properties of the proposed method are verified by simulation studies, and the advantages over existing methods are illustrated in real analysis.

Keywords: Diffusion process; Heterogenous autoregressive model; High-dimensional analysis; High-frequency data; Non-asymptotic property; Tensor technique.

Cross-layer retrospective retrieving via layer attention

Yanwen Fang[1], Yuxi Cai[1], Guodong Li[1]

[1]*Department of Statistics & Actuarial Science, The University of Hong Kong*

Abstract

More and more evidence has shown that strengthening layer interactions can enhance the representation power of a deep neural network, while self-attention excels at learning interdependencies by retrieving query-activated information. Motivated by this, we devise a cross-layer attention mechanism, called multi-head recurrent layer attention (MRLA), that sends a query representation of the current layer to all previous layers to retrieve query-related information from different levels of receptive fields. A light-weighted version of MRLA is also proposed to reduce the quadratic computation cost. The proposed layer attention mechanism can enrich the representation power of many state-of-the-art vision networks, including CNNs and vision transformers. Its effectiveness has been extensively evaluated in image classification, object detection and instance segmentation tasks, where improvements can be consistently observed. For example, our MRLA can improve 1.6% Top-1 accuracy on ResNet-50, while only introducing 0.16M parameters and 0.07B FLOPs. Surprisingly, it can boost the performances by a large margin of 3-4% box AP and mask AP in dense prediction tasks.

SARMA: A Computationally Scalable High-Dimensional Time Series Model

Lu Kexin[1], Huang Feiqing[2], Zheng Yao[3], Li Guodong[4]

[1]*University of Hong Kong, neithen@hku.hk*
[2]*University of Hong Kong, amieehuang@hku.hk*
[3]*University of Connecticut, yao.zheng@uconn.edu*
[4]*University of Hong Kong, gdli@hku.hk*

Abstract

This paper introduces a novel parametric infinite-order vector autoregressive model. As a variant of the vector autoregressive moving average (ARMA) model, it not only inherits desirable properties such as parsimony and rich temporal dependence structures, but also avoids two well-known drawbacks of the former: (i) non-identifiability and (ii) computational intractability even for moderate-dimensional data. Moreover, its parameter estimation is scalable with respect to the complexity of temporal dependence, namely the number of decay patterns constituting the autoregressive structure; hence it is called the scalable ARMA (SARMA) model. In the high-dimensional setup, we further impose a low-Tucker-rank assumption on the coefficient tensor of the proposed model. The resulting model has the form of a regression with embedded dynamic factors and hence can be especially suited for financial and economic data. Non-asymptotic error bounds for the proposed estimator are derived, and a tractable alternating least squares algorithm is developed. Theoretical and computational properties of the proposed method are verified by simulation studies, and the advantages over existing methods are illustrated in real applications.

Keywords: High-dimensional time series; Identifiability; Reduced-rank regression; Scalability; Tensor decomposition; Vector AR(∞); Vector ARMA.