2019 ISI ISM ISSAS Joint Conference

Indian Statistical Institute, India(ISI)

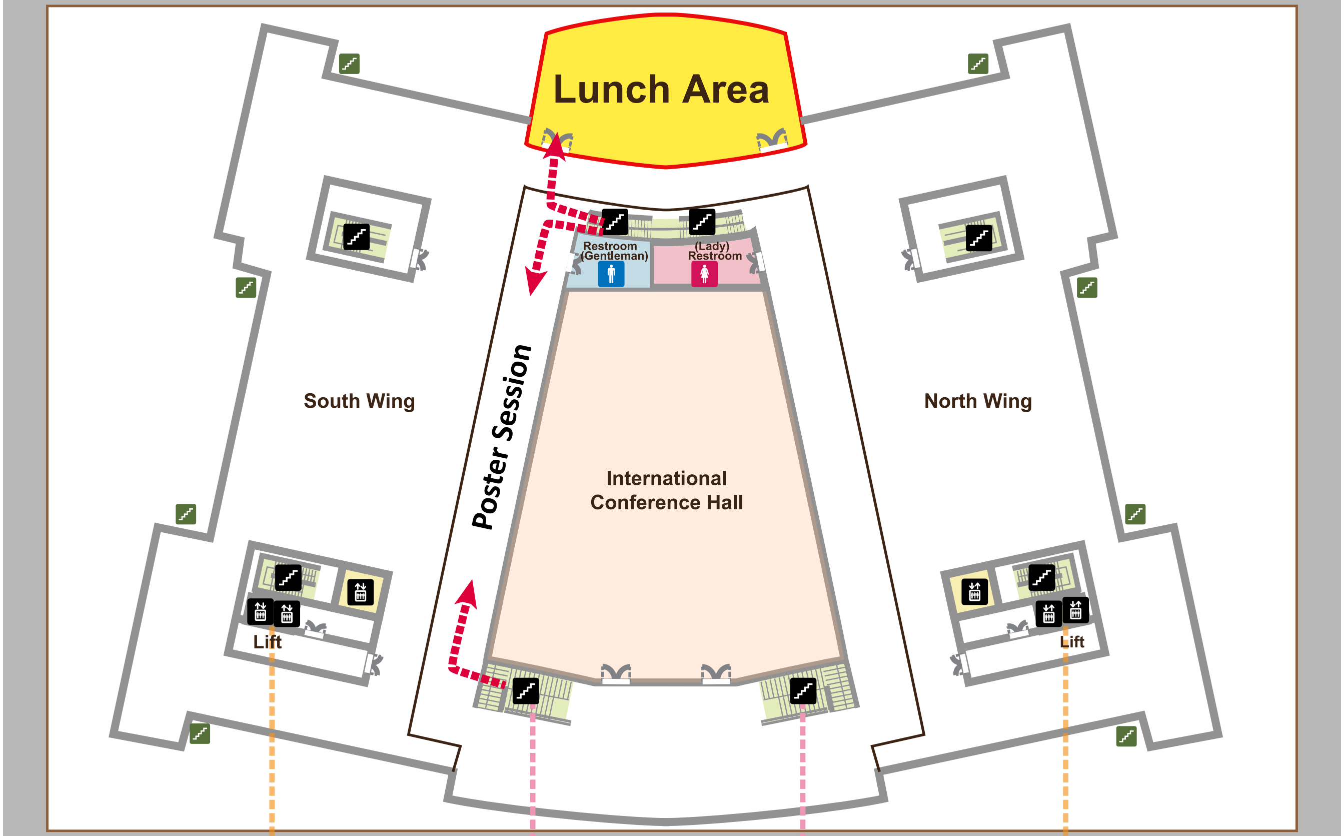
Institute of Statistical Mathematics, Japan(ISM)

Institute of Statistical Science, Academia Sinica, Taiwan(ISSAS)

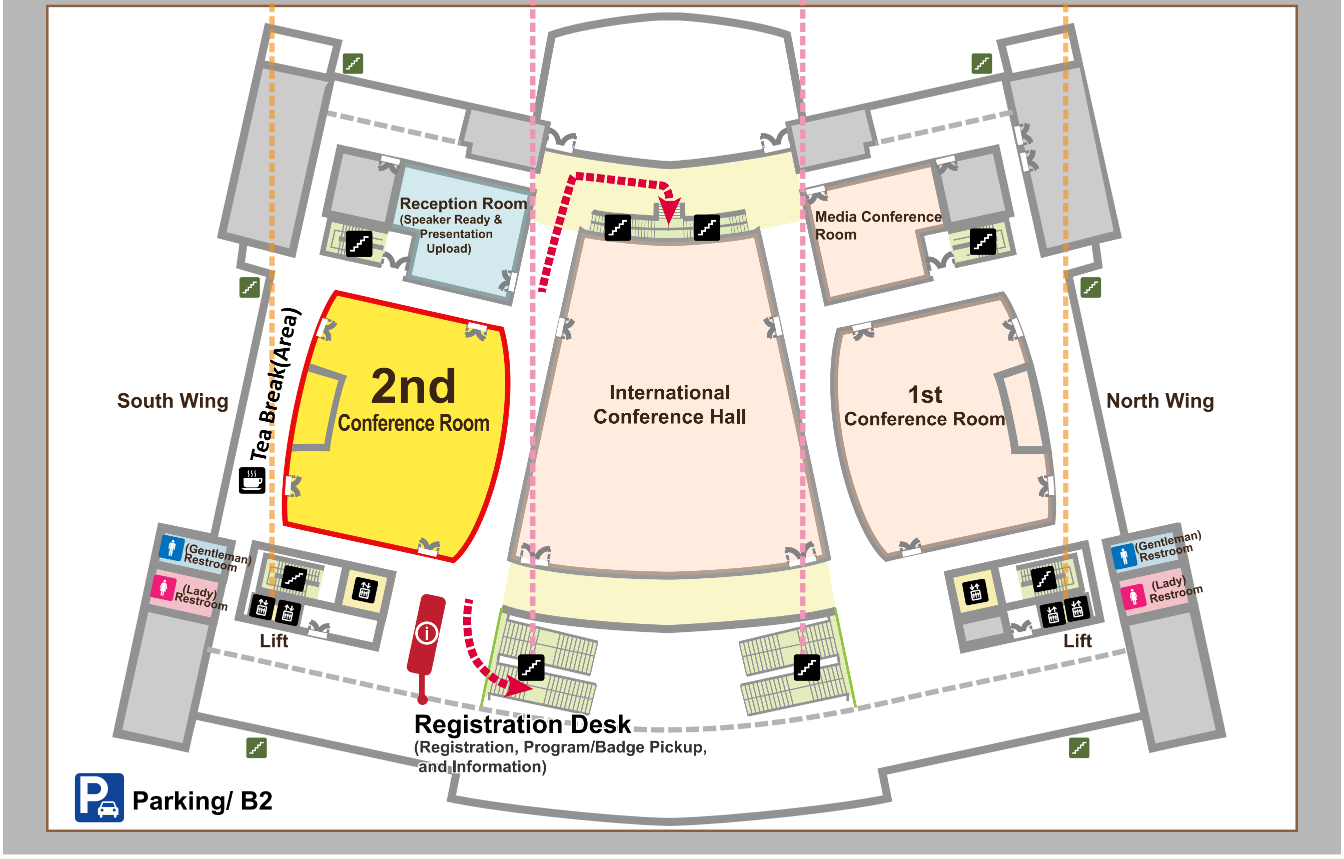
http://www3.stat.sinica.edu.tw/2019issas/

17th - 19th January,
2nd Conference Room,
Humanities and Social Sciences Building,
Academia Sinica, Taipei, Taiwan

Humanities and Social Sciences Building (Floor 4)



Humanities and Social Sciences Building (Floor 3)



2019 ISI-ISM-ISSAS Joint Conference

January 17- January 19, 2019

Institute of Statistical Science, Academia Sinica

Program

January 17 (Thursday)	
08:30-08:50	**Registration**
08:50-09:00	**Opening：Chun-Houh Chen**
09:00-10:20	**Session 1 Deep Learning** **Chair: Deepayan Sarkar**🟠
	Masaaki Imaizumi🔴 Analysis for Deep Learning by Function Estimation Theory
	Frederick Kin Hing Phoa🟢 A Two-Step Method for Quantifying Relationship between Research Subjects in the Web of Science
	Stephen Wu🔴 Title: Applications of Transfer Learning in Materials Science
10:20-10:40	**Coffee Break**
10:40-12:00	**Session 2 Multivariate/High-dimensional Data** **Chair: Hironori Fujisawa**🔴
	Anil K. Ghosh🟠 Modifications of Some Distance Based Two-Sample Tests for High-Dimensional Data
	Ming-Yueh Huang🟢 Adaptive Reduction of Curse of Dimensionality in Nonparametric Instrumental Variable Estimation
	Anil K. Ghosh🟠 Multivariate Tests of Independence Among Several Random Vectors
12:00-13:30	**Lunch**
13:30-14:20	**Session 3 Bioinformatics** **Chair: Masao Ueki**🔴
	Deepayan Sarkar🟠 Performing Gene Set Enrichment Analysis with Multi-omics Data
	Grace S. Shieh🟢 Uncovering synthetic lethal interactions for cancer therapeutics and prognostic markers

14:20-14:40	**Coffee Break**		
14:40-15:25	**Oral Presentation (1)**		
	14:40-14:45 **1. Xiaoling Dou**🔴 The Stirling and Eulerian numbers in the Edo Period	14:45-14:50 **2. Hironori Fujisawa**🔴 Recent Works on Sparse Modeling	14:50-14:55 **3. Hibiki Kaibuchi**🔴 Comparison of EVT methods for GARCH-EVT approach applied to financial time series
	14:55-15:00 **4. Shogo Kato**🔴 A measure for comparing upper and tail probabilities of bivariate distributions	15:00-15:05 **5. Takayuki Kawashima**🔴 Stochastic Gradient Descent for Doubly-Nonconvex Composite Optimization	15:05-15:10 **6. Satoshi Kuriki**🔴 Existence and Uniqueness of Maximum Likelihood Estimators of Kronecker Product Covariances
	15:10-15:15 **7. Shuhei Mano**🔴 Samplers with Computational Algebra and their Applications	15:15-15:20 **8. Masao Ueki**🔴 Quick assessment of problematic genome-wide environment interaction studies	15:20-15:25 **9. Junchao Zhang**🔴 The Effect of Transportation Benefits on Health and Consumption Among the Elderly: Quasi-Experimental Evidence from Urban China
15:25-15:35	**Coffee Break**		
15:35-16:40	**Oral Presentation (2)**		
	15:35~15:40 **10. Szu-Chi Chung**🟢 A Dimension Reduction Method	15:40~15:45 **11. Ju-Sheng Hong**🟢 Model-based causal mediation analysis	15:45~15:50 **12. Jing-Wen Huang**🟢 A systematic construction of cost-

	for Cryo-EM Image Processing	of semi-competing risk data	efficient designs for order-of-addition experiments
	15:50~15:55 **13. Cheng-Yu Hung**● Robust PCA and its Extension	15:55~16:00 **14. Chi-Wei Lai**● Spatial Modeling of Ground-Level PM2.5 in Taiwan Based on Two Types of Data	16:00~16:05 **15. En-Yu Lai**● Mediation Analyses of Ultraviolet, Air Pollution, and Structural Variations in the Human Genomes from the Taiwan Biobank
	16:05~16:10 **16. Yi-Ju Lee**● The Complexity of Schizophrenic Brain: Power Law Scaling in Resting-State fMRI Data	16:10~16:15 **17. Szu-Han Lin**● Random partition t-SNE	16:15~16:20 **18. Martin T. Lukusa**● On estimation methods for extra zeros in crash data with missing data
	16:20~16:25 19. **Siddharth Nayak**● Effective connectivity delineates putative roles for cortical regions in emotional inhibition across aging	16:25~16:30 **20. Jia-Ying Su**● HDMV: Visualization for high-dimensional mediation effects	16:30~16:35 **21. Khong Loon Tiong**● Explaining cancer type specific mutations with transcriptomic and epigenomic features in normal tissues
	16:35~16:40 **22. Shao-Hsuan Wang**● Analyzing Model Bias in Cryo-EM Single-Particle Image Processing		
16:40-16:50	**Coffee Break**		

16:50-17:50	**Poster Presentation (Presenter stands next to the poster)**
18:30-20:30	**Reception**

●All oral sessions will be held in 2nd Conference Room on the 3rd floor of the Humanities and Social Science Building.
●Poster session and lunch will be held in Lobby and Recreation Room on the 4th floor of the Humanities and Social Science Building.

January 18 (Friday)	
09:00-10:20	**Session 4 Association & Causation** **Chair: Satoshi Kuriki🔴**
	Saurabh Ghosh🟠 A Transmission Based Association Test for Multivariate Phenotypes Using Quasi Likelihood
	Yen-Tsung Huang🟢 Causal mediation of semicompeting risks
	Daichi Mochihashi🟢 Learning Co-Substructures by Kernel Dependence Maximization
10:20-10:40	**Coffee Break**
10:40-12:00	**Session 5 Bayes Related** **Chair: Hsin-Cheng Huang🟢**
	Abhik Ghosh🟠 Robust Pseudo Bayes Estimation under Independent Non-Homogenous Set Up
	Masayo Y. Hirose🔴 An Empirical Bayes Confidence Interval in the Presence of High Leverage for Small Area Inference
	Shunichi Nomura🔴 Hierarchical Topic Models for Tensor Count Data
12:00-13:30	**Lunch**
13:30-18:00	**Local Tour**
18:30-20:30	**Banquet**

January 19 (Saturday)	
09:00-10:20	**Session 6 Probability Related** **Chair: Hideatsu Tsukahara🔴**
	Antar Bandyopadhyay🟠 “Power of Two Choices” in De-Preferential Pólya Urn Schemes
	Satoshi Ito🔴 Computation of clinch and elimination numbers in league sports based on integer programming
	Anish Sarkar🟠 Hack’s Law in a Drainage Network Model: A Brownian Web Approach
10:20-10:40	**Coffee Break**
10:40-12:00	**Session 7 Time Series Related** **Chair: Antar Bandyopadhyay🟠**
	Yoshinori Kawasaki🔴 Forecasting Financial Market Volatility Using a Dynamic Topic Model
	Yoshiyuki Ninomiya🔴 AIC for Change-Point Models and its Application to a Biological Data Analysis
	Arthur Chih-Hsin Tsai🟢 Adult age differences in inhibitory control as revealed by fMRI and drift diffusion model
12:00-13:30	**Lunch**
13:30-14:20	**Session 8 Bootstrap Resampling** **Chair: Shuhei Mano🔴**
	Hsin-Wen Chang🟢 Nonparametric Confidence Band for Activity Profiles Based on Wearable Device Data
	Wei-Chung Liu🟢 A non-parametric method for inferring food web parameters
14:20-14:40	**Closing**

Session 1

Deep Learning

Chair Person: Deepayan Sarkar

09:00~10:20

January 17, 2019

Analysis for Deep Learning by Function Estimation Theory

Masaaki Imaizumi

The Institute of Statistical Mathematics, Japan

Abstract

We investigate a theoretical reason that deep neural networks (DNNs) perform better than other models in some cases from the viewpoint of their statistical properties. While DNNs have empirically shown higher performance than other standard methods, understanding its mechanism is still a challenging problem. From an aspect of the statistical theory, it is known many standard methods attain the optimal rate of generalization errors for smooth functions in large sample asymptotics, and thus it has not been straightforward to find theoretical advantages of DNNs. This paper fills this gap by considering the following two points; non-smooth functions and low-dimensional manifolds. We derive the generalization error of estimators by DNNs with a ReLU activation, and show that convergence rates of the generalization by DNNs are almost optimal to estimate the components. In addition, our theoretical result provides guidelines for selecting an appropriate number of layers and edges of DNNs. We provide numerical experiments to support the theoretical results.

A Two-Step Method for Quantifying Relationship between Research Subjects in the Web of Science

Frederick Kin Hing Phoa[1], Hsin-Yi Lai[2], Hiroka Hamada[3] and Keisuke Honda[3]

[1]*Institute of Statistical Science, Academia Sinica, Taiwan*

[2]*Institute of Statistics, National Chiao Tung University, Taiwan*

[3] *The Institute of Statistical Mathematics, Japan*

Abstract

Pointwise Mutual Information (PMI) is a measure of association used in in- formation theory and statistics. In this work, we propose a two-step method to measure the relationship, via the PMI values, between subject types in the Web of Science. We consider a complex string of the subject type as a vector that indicates the subject components of the subject type. The first step is a classification step via deep learning to determine whether two selected subject types are independent in terms of citations. If they are determined to be dependent, then we build a model for the PMI value from a starting subject type to an eventual subject type by considering the change of subject components. The resulting model is useful for estimating the PMI values if both subject types exist in the Web of Science, or for predicting the PMI values if at least one subject type is newly introduced to the Web of Science.

Keywords: Deep Learning, Pointwise Mutual Information (PMI), Regression Analysis, Web of Science (WoS).

Applications of Transfer Learning in Materials Science

Stephen Wu

The Institute of Statistical Mathematics, Japan

Abstract

Deep neural network has shown tremendous prediction power on a large variety of problems when a significantly big data set is available for training. In many practical cases, users may either be limited by the amount of data or the amount of computation power available to perform a large scale deep neural network training. In order to expand the benefit of the powerful technology, the idea of transfer learning has become an essential method in many applications, especially in materials science. While modern material databases have grown rapidly in size. The variation of the number of data available across different material properties is still very big. When we are building a model to predict a property with only a small data set, the basic concept of transfer learning is to extract information from a well-trained deep neural network using a big data set of a related material property to assist the training of our target property. Popular methods, such as featurizer or fine-tuning, often maintains the neural network structure trained for one problem and then adjust the parameters using just a small data set of the target problem. In this study, we have shown the importance of transfer learning in many materials science problems through a variety of case studies.

Session 2

Multivariate/ High-dimensional Data

Chair Person: Hironori Fujisawa

10:40~12:00

January 17, 2019

Modifications of Some Distance Based Two-Sample Tests for High-Dimensional Data

Anil K. Ghosh

Indian Statistical Institute, Kolkata

Abstract

Testing for the equality of two high dimensional distributions is a challenging problem inmstatistics, and several distance based tests are available for it. These tests can be broadly classified into two categories: (i) tests based on averages of inter-point distances and (ii) tests based on edge-weighted graphs, where the edge-weights are defined using inter-point distances among the observations. These above mentioned tests use the Euclidean metric for distance computation, mainly due to its popularity. But, because of the concentration of Euclidean distances and violation of neighborhood structure in high dimensions, these tests often yield poor results for high dimensional data. To take care of this problem, first we introduce a class of distance functions and use them to successfully modify the tests based on averages of inter-point distances. However, this type of modification is not suitable for some of the graph based tests. For adequate modification of such tests, we use another class of dissimilarity measures. High dimensional consistency of the two types of modified tests are proved under appropriate regularity conditions. We investigate the performance of these tests using several numerical studies and discuss their relative advantages and disadvantages.

[This is a joint work with Soham Sarkar and Rahul Biswas]

Adaptive Reduction of Curse of Dimensionality in Nonparametric Instrumental Variable Estimation

Ming-Yueh Huang

Institute of Statistical Science, Academia Sinica, Taiwan

Abstract

Nonparametric estimation of instrumental variable treatment effects usually follows from various nonparametric identification results. However, such estimators often suffer from curse of dimensionality in practice since multi-dimensional covariates are common. To reduce the effect of curse of dimensionality, we study nonparametric identification of a variety of treatment effects under different sufficient dimension models. Efficiency of estimation is also studied, and we find out that unlike fully nonparametric methods, nonparametric estimators based on maximal dimension reduction from the identification results may not be efficient. Maximal dimension reduction for attaining efficiency is studied for a binary instrumental variable, and is extended to multivariate, general instrumental variables. The proposed nonparametric sufficient dimension reduction framework does not impose any constraints to the distribution of the observed data, but reduce curse of dimensionality in a data-adaptive manner.

Multivariate Tests of Independence Among Several Random Vectors

Anil K. Ghosh

Indian Statistical Institute, Kolkata

Abstract

Testing for mutual independence among several random vectors is a very challenging problem in Statistics, and it has gained significant attention in recent years. Unfortunately, most of the existing tests of independence deal with only two random vectors, and most of them cannot be meaningfully extended to test the mutual independence among several random vectors of arbitrary dimensions. However, there has been some recent development on testing mutual independence among several random variables, but these univariate tests do not have natural multivariate extensions. Here, we propose some general recipes for multivariate extensions of these tests. The resulting tests turn out to be omnibus consistent whenever we have consistency for the corresponding univariate test. We demonstrate the usefulness of the proposed methods using several numerical studies and discuss possible extensions.

[This is a joint work with Soham Sarkar, Angshuman Roy, and Alok Goswami.]

Session 3

Bioinformatics

Chair Person: Masao Ueki

13:30~14:20

January 17, 2019

Performing Gene Set Enrichment Analysis with Multi-omics Data

Deepayan Sarkar

Indian Statistical Institute, Delhi Centre

Abstract

Many high-throughput genomic technologies measure expression at the level of genes, usually to then use some statistical methods to identify differentially expressed genes. It has become increasingly common to follow up on this basic analysis by performing some form of gene set or pathway enrichment analysis, where pre-defined gene sets are evaluated for differential behaviour. With well-curated gene sets, such analyses provide greater insight into the underlying biology by increasing the power and reducing the dimensionality of the underlying statistical problem. As more and more experiments have started measuring gene expression using multiple-omics technologies (e.g., proteome, transcriptome, epigenome), it is important to investigate how these multi-omics measurements can be combined to identify interesting gene sets. In this talk, we describe a simple generalisation of a common approach for conventional gene set enrichment analysis that can be applied to multi-omics data.

Uncovering synthetic lethal interactions for cancer therapeutics and prognostic markers

Grace S. Shieh

Institute of Statistical Science, Academia Sinica, Taiwan

Abstract

Targeting synthetic lethal (SL) partners of mutated cancer genes will specifically kill cancer cells bearing the mutations but spare normal cells. Therefore, for non-druggable mutant tumor suppressor genes and oncogenes, e.g., TP53 and KRAS, synthetic lethality strategy offers an elegant alternative.

Two genes are said to have synthetic lethal (SL) interaction if their simultaneous mutations lead to cell death, but each individual mutation does not. Using synthetic lethality-based methods to develop cancer-specific therapeutics has been rapidly adapted due to its translational impact. Here, we present an integrated approach to uncover SL pairs in colorectal cancer (CRC) and lung adenocarcinoma (LADC).

This approach used 660+ collected verified SL pairs, microarray gene expression data, protein levels (immunohistochemistry staining) of ~20 selected genes and clinical features of 171/130+ CRC/LADC patients. This method resulted in 11 predicted SL pairs for CRC, including MSH2-POLB and CSNK1E-MYC previously verified in CRC. Additionally we validated CSNK1E-TP53 and CTNNB1-TP53 using RNAi and small-molecule inhibitor, and the former via mouse model. Further, CSNK1E-TP53, CTNNB1-TP53 and two other protein pairs are shown to be markers for CRC patient survival.

For LADC, of the 20+ predicted SL pairs, four pairs are consistent with literature and the synthetic lethality of TP53-PARP1 was validated in CL1-5 and H1975 cells. RAD54B↑, BRCA1↓-RAD54B↑, FEN1(N)↑-RAD54B↑ and PARP1↑-RAD54B↑ were revealed to be prognosis markers, independent from age and stage. Further, these markers were confirmed by three independent gene expression data sets. Finally, two fundamental questions on synthetic lethality will be discussed.

The results are joint work with Jan-Gowth Chang, P.-C. Yang, K-C. Chang, KL Tiong and others.

Poster Session

Chair Person: Tso-Jung Yen

14:40~17:50

January 17, 2019

The Stirling and Eulerian numbers in the Edo Period

Xiaoling Dou[1] and Hsien-Kuei Hwang[2]

[1]*Waseda University & ISM*

[2]*Institute of Statistical Science*

Abstract

Similar to the recurrence relation satisfied by the binomial coefficients, the Stirling numbers and the Eulerian numbers can also be computed by the recurrences subject to proper boundary conditions. Over the last five centuries or so, these numbers emerged naturally and were studied extensively in a large number of diverse areas, ranging from finite calculus and series summations to combinatorial structures and computer algorithms, and from statistics to spline interpolations, to name a few. While the history of the developments of these numbers in the West has been largely and factually clarified, that in the East has remained mostly obscure. In this work, we aim to provide more historical materials during the Edo Period concerning these numbers, and to specially shed further light on their evolution (including introduction and use) in the Wasan History.

We present our findings on Stirling and Eulerian numbers in the Wasan History during the Edo Period, as well as the closely connected Bell numbers. Interestingly, unlike the early developments of these numbers in the West, which are mostly computational and algebraic, those carried out by the Wasankas already are not merely computational but also were motivated by combinatorial problems, adding another rich dimension to the diversity and usefulness of these numbers. We will also present the combinatorial connections of these numbers to certain games frequently played during this Period.

Recent Works on Sparse Modeling

Hironori Fujisawa

The Institute of Statistical Mathematics, Japan

Abstract

In this poster, I talk about some resent works on sparse modeling. All the proposed methods are applicable with R packages.

Hirose, K., Fujisawa, H. and Sese, J. (2017). Robust sparse Gaussian graphical modeling. Journal of Multivariate Analysis, Vol.161, 172-190.

The sparse gaussian graphical modeling was proposed by Freedman et al. (2008). The method is well known as glasso. This paper proposes its robustified method using the gamma-divergence. The parameter estimation algorithm constructed here monotonically decreases the penalized loss function. This method outperformed existing ones in simulations, in particular, under heavy contamination, and was superior when the methods were applied to two gene expression datasets.

Takada, M., Suzuki, T. and Fujisawa, H. (2018). Independently Interpretable Lasso: A New Regularizer for Sparse Regression with Uncorrelated Variables. The 21st International Conference on Artificial Intelligence and Statistics (AISTATS).
We often think that Lasso tends not to select highly correlated explanatory variables. However, this will not be true in a high-dimensional case. To address this issue, this paper proposes an improved Lasso with a devised penalty, which suppresses to select highly correlated explanatory variables. The proposed method was superior to existing methods when the methods were applied to ten microarray datasets in the UCI database.

Kawano, S., Fujisawa, H., Takada, T. and Shiroishi, T. (2018). Sparse principal component regression for generalized linear models. Computational Statistics and Data Analysis, Vol.124, 180-196.

Consider the regression problem in a high-dimensional data. We sometimes do the dimension reduction using the PCA and then we make the regression model using few PCA scores as new explanatory variables. However, this method uses only major effects and cannot incorporate minor effects which are hidden behind the major effects. To address this issue, this paper proposes a novel penalized method. The penalized loss function is constructed from the negative likelihood and the PCA score penalty. The proposed method was applied to mouse consomic strain data and then presented more significant clusters than a usual PCA.

Comparison of EVT methods for GARCH-EVT approach applied to financial time series

Hibiki Kaibuchi* and Yoshinori Kawasaki

The Institute of Statistical Mathematics, Japan

Abstract

Managing extreme event risk in finance and insurance in vital in our modern society. It is known that the statistically justifiable modeling and prediction of rare events are inherently scarce. In order to prevent or prepare for unfavorable scenarios, the approaches based on extreme value theory (EVT) have been devised. The aim is to estimate conditional extreme quantiles (Value at Risk) using GARCH-EVT framework. For that, we: (i) pre-whiten the financial time series with a parsimonious but effective GARCH(1,1) process for forecasting volatility and an AR(1) model for forecasting conditional mean; (ii) apply the semi-parametric bias-corrected tail estimators under beta-mixing condition to the residuals from the GARCH analysis instead of the Peaks-Over-Thresholds (POT) method under I.I.D. condition. The results are illustrated on simulated data and on a financial real dataset.

Keywords: extreme values, financial time series, GARCH, quantitative risk management

A measure for comparing upper and tail probabilities of bivariate distributions

Shogo Kato

The Institute of Statistical Mathematics, Japan

Abstract

It is well known that the lack of fit in tails of probability distributions leads to erroneous results in statistical analysis. In this study we propose a measure to compare upper and lower tail probabilities of bivariate distributions. It is seen that the expression for the proposed measure can be simplified if bivariate distribution functions are represented using copulas. With this representation, some properties of the proposed measure are investigated. It is shown that the limit of the proposed measure as a tuning parameter goes to zero can be expressed in a simple form under certain conditions on copulas. A sample analogue of the proposed measure is given and its asymptotic normality is shown. A nonparametric test of symmetry in upper and lower tails based on the sample analogue of the measure is presented. As an illustrative example, the presented measure is applied to stock daily returns of three stock indices.

[This is joint work with Toshinao Yoshiba (Bank of Japan) and Shinto Eguchi (Institute of Statistical Mathematics).]

Stochastic Gradient Descent for Doubly-Nonconvex Composite Optimization

Takayuki Kawashima

The Institute of Statistical Mathematics, Japan

Abstract

The stochastic gradient descent has been widely used for solving composite optimization problems in big data analyses. Many algorithms and convergence properties have been developed. The composite functions were convex primarily and gradually nonconvex composite functions have been adopted to obtain more desirable properties. The convergence properties have been investigated, but only when either of composite functions is nonconvex. There is no convergence property when both composite functions are nonconvex, which is named the doubly-nonconvex case. We investigate convergence properties for the stochastic doubly-nonconvex composite optimization problem.

[This is a joint work with Prof. Hironori Fujisawa (ISM)]

Existence and Uniqueness of Maximum Likelihood Estimators of Kronecker Product Covariances

Satoshi Kuriki

The Institute of Statistical Mathematics, Japan

Abstract

Suppose that we have iid samples from a vector-valued Gaussian distribution. When the covariance matrix of the Gaussian distribution has no structure, the MLE uniquely exists with probability one if and only the sample size n is equal to or greater than the size of the covariance matrix. However, this is not the case where the covariance matrix has a structure and is described with a fewer number of parameters. In this presentation, we consider the case where the covariance matrix is the Kronecker product of two matrices (m1×m1 and m2×m2 matrices). We show that the existence and the uniqueness of the MLE are characterized by a rank of an m1×m2×n tensor. In particular, when the sample size n is 2, the problem can be reduced by Kronecker's canonical form of two matrices. The tensor rank is given explicitly as the solution of an integer programming. The Groebner basis computation is also useful when m1 and m2 are small.

[Joint work with Mathias Drton, University of Copenhagen]

Samplers with Computational Algebra and their Applications

Shuhei Mano

The Institute of Statistical Mathematics, Japan

Abstract

Diaconis and Sturmfels (Ann. Statist. 1998) introduced the notion of Markov bases for the Markov chain Monte Carlo, which are bases of transitions of the Markov chain. They showed that a Markov basis is given by a Gröbner basis of a toric ideal of the polynomial ring determined by sufficient statistics. By considering the ring of differential operators, it is shown that a direct sampling is possible by constructing a Markov chain whose sample path follows the target distribution. For the direct sampler, the cost we must pay is computation of normalized constants which involve hypergeometric polynomials. Some applications including sampling from non-decomposable contingency tables and posterior sampling from Bayesian mixture models with non-exchangeable priors will be presented.

Keywords: Bayesian mixture model, computational algebra, contingency table, Gröbner basis, hypergeometric system, sampler

Quick assessment of problematic genome-wide environment interaction studies

Masao Ueki

Statistical Genetics Team, RIKEN Center for Advanced Intelligence Project, Japan

The Institute of Statistical Mathematics, Japan

Abstract

Gene-environment (GxE) interaction is one potential explanation for the missing heritability problem. A popular approach to genome-wide environment interaction studies (GWEIS) is based on regression models involving interactions between genetic variants and environment variables. Unfortunately, GWEIS encounters systematically inflated test statistics more frequently than a marginal association study. Problematic behavior may occur due to poor specification of the null model in GWEIS. Improved null model specification may resolve the problem, but the investigation requires many time-consuming analyses of genome-wide scans, e.g. by trying out several transformations of the phenotype. It is therefore helpful if we can predict problematic behavior beforehand. We present a simple closed-form formula to assess problematic behavior of GWEIS under the null hypothesis of no genetic effects. It requires only phenotype, environment variables, and covariates, enabling quick identification of problematic studies. Applied to real data from the Alzheimer's Disease Neuroimaging Initiative (ADNI), our formula identified problematic studies from among one hundred GWEIS considering each metabolite as the environment variable in GxE interaction. Our formula is useful to quickly identify problematic GWEIS without requiring a genome-wide scan.

[This is a joint work with Masahiro Fujii of Alfresa Pharma Corporation and Gen Tamiya of Tohoku University, Japan.]

The Effect of Transportation Benefits on Health and Consumption Among the Elderly: Quasi-Experimental Evidence from Urban China

Ting Yin[1], Zhigang Yin[2], Junchao Zhang[3]

[1]Research Institute of Economy, Trade and Industry (RIETI), Japan

[2]Shanghai Research Center on Aging, China

[3]Institute of Statistical Mathematics, Japan

Abstract

This study estimates the causal effect of transportation subsidies or similar benefits on the health of the elderly. We exploit a discontinuity in the probability of receiving transportation benefits induced by an age-based policy to take account of the endogeneity of treatment status. Our baseline IV results indicate that receiving public transportation benefits significantly improves elderly people's health condition by approximately 10 percentage points. The results are robust under different specifications and placebo tests. Further tests on possible channels show that the health effect is driven by increasing food consumption and health care utilization, but not by the amount of exercise done.

A Dimension Reduction Method for Cryo-EM Image Processing

Szu-Chi Chung

Institute of Statistical Science, Academia Sinica, Taiwan

Abstract

With the recent breakthrough in the camera together with microscopy automation and advancement of algorithms, cryo-EM has become a mainstream technique to solve structures of macromolecules at near atomic resolution. As a result, single particle cryo-EM has awarded Nobel Prize in chemistry in 2017. However, further extending to atomic resolution has been hindered by both the noisy nature of the images and the heterogeneity of samples, caused by very low doses of electrons during imaging and the flexibility of particles before plunge-freezing, respectively. Enhancement of the signal-to-noise Ratio (SNR) and differentiate the conformation states of these images is thus the key for solving higher resolution 3D structure. In this work, we propose a dimension reduction method called Two Stage Dimensional Reduction (2SDR) and a clustering method called Distributed Robust Multi-Reference Alignment (DRMRA). We first demonstrate that 2SDR can realize effective de-noising and can be applied to different stages of the workflow of cryo-EM processing including assessing micrographs, screening particles, and enhancing 2D alignment performance. Second, we show that the DRMRA can speed up the time-consuming 2D clustering process with more robust results compare with state-of-the-art methods. Finally, several experiment cryo-EM data sets are used to test our method. By applying 2SDR strategy, we can obtain better 2D class average results and enable single-frame workflow. In addition, we show that DRMRA has the potential to find more conformation states in 2D analysis. These results are encouraging and may help us to overcome the obstacles so to push the resolution limit of final 3D map. These techniques and other state-of-the-art algorithms are integrated in our currently developing software platform called ASCEP in order to build a better workflow for Cryo-EM.

[This is a joint work with Po-Yao Niu, Ting-li Chen Su-Yun Huang, Wei-Hau Chang and I-Ping Tu.]

Model-based causal mediation analysis of semi-competing risk data

Ju-Sheng Hong*, Shu-Hisen Cho* and Yen-Tsung Huang†

Institute of Statistical Science, Academia Sinica, Taiwan

Abstract

The semi-competing risks data are commonly found in biomedical research in which a primary outcome (e.g. death) may censor an intermediate event (e.g. cancer incidence) but not vice versa. We propose a model-based approach formulating the semi-competing risks as a causal mediation problem. Here we construct a mediation model with the intermediate and primary events, respectively as the mediator and the outcome. Indirect effect is defined as an effect of the exposure on the primary outcome mediated through the intermediate event and direct effect is an effect not mediated through the intermediate event. We construct proportional hazards estimators for direct and indirect effects, with time-varying weights, which are estimated by a series of logistic regression. Based on the martingale theory and functional delta method, the asymptotic properties are established for the proposed estimators. Using simulations, we evaluate the finite-sample performance of the proposed estimators. The utility of our proposed method is illustrated in a hepatitis study of liver cancer survival.

Keywords: causal mediation model; Cox proportional hazards model; functional delta method; semi-competing risk

----

*Equal Contribution

†Institute of Statistical Science, Academia Sinica, Taipei, Taiwan.; email: ythuang@stat.sinica.edu.tw

Applications of Transfer Learning in Materials Science

Jing-Wen Huang[1], Frederick Kin Hing Phoa[2] and Yuan-Lung Lin[2]

[1]*Institute of Statistics, National Tsing Hua University, Taiwan*

[2]*Institute of Statistical Science, Academia Sinica, Taiwan*

Abstract

An order-of-addition (OofA) experiment aims at investigating how the order of factor inputs affects the experimental response, which is recently of great interest among practitioners in clinical trials and industrial processes. Although the initial framework was established for more than 70 years, recent studies in the design construction of OofA experiments focused on their properties of algebraic optimality rather than cost-efficiency. The latter is more practical in the sense that some experiments, like cancer treatments, may not easily have adequate number of observations. In this work, we propose a systematic construction method for designs in OofA experiments from cost-efficient perspective. In specific, our designs take the effect of two successive treatments into consideration. To be cost-efficient, each pair of level settings from two different factors in our design matrix appears exactly once. Compared to recent studies in OofA experiments, our designs not only handle experiments of one-level factors (i.e. all factors are mandatorily considered), but also factors of two or more levels, so practitioners may insert placebo or choose different dose when our designs are used in an OofA experiment in clinical trials for example.

Robust PCA and its Extension

Cheng-Yu Hung and I-Ping Tu

Institute of Statistical Science, Academia Sinica, Taiwan

Abstract

Principal Component Analysis (PCA) has been used in an overwhelming manner for data analysis. However, PCA did not perform well when data did not follow the model well like sensor failure or corrupted sample. Candes et al. (2011) proposed Robust Principal Component Analysis (RPCA) to recover the data and proved that it can perform very well when data has the sparsity property for the signal with a low rank background. Unfortunately, the FRET data set does not satisfy the working condition. Here, we employ a sampling scheme to enable the application for the FRET data. For extremely large number of pixel image application, RPCA may suffer from computation loading. Thus, we also extend RPCA to a high order SVD version.

Spatial Modeling of Ground-Level PM2.5 in Taiwan Based on Two Types of Data

Chi-Wei Lai and Hsin-Cheng Huang

Institute of Statistical Science, Academia Sinica, Taiwan

Abstract

There are two systems to monitor fine particulate matter (PM2.5) in Taiwan. One consists of 77 monitoring stations of the Environmental Protection Administration, which provides high-quality measurements. The other one involves a large number of low-cost internet-of-things devices called AirBoxes, which produce less precise measurements but with much broader coverage. In this research, we propose a spatial model to obtain spatial prediction at any location in Taiwan by combining these two types of data. In addition, we develop a Shiny application that automatically identifies unusual measurements and shows the current PM2.5 concentration map with uncertainty quantification based on the proposed method.

Keywords: fine particulate matter, kriging, regression calibration.

Mediation Analyses of Ultraviolet, Air Pollution, and Structural Variations in the Human Genomes from the Taiwan Biobank

En-Yu Lai, Wan-Ping Lee, Wen-Chi Pan, Ming-Wei Su, Chen-Yang Shen, and Yen-Tsung Huang

Institute of Statistical Science, Academia Sinica, Taiwan

Abstract

Structural variation is a DNA region that shows changes in copy number, sequence orientation or chromosomal location. Previous studies have suggested a link between air pollution and genetic variation in animal experiments and longitudinal studies, but the sample size is rather limited. It is imperative that a population-based study is conducted to document the potential hazard of the environmental exposure such as air pollution and ultraviolet to human genome and health. The Taiwan Biobank has been collecting biological specimens and conducts the whole-genome sequencing in order to build the reference genome of the Taiwanese population. In this study, we aim to characterize the causal relationship among ultraviolet, air pollution and structural variations. We applied a mediation model to describe the influence of ultraviolet toward structural variants through air pollution. The preliminary results showed a strong effect from ultraviolet to structural variants mediated by air pollution. Validation studies are needed to confirm this interesting finding.

The Complexity of Schizophrenic Brain: Power Law Scaling in Resting-State fMRI Data

Yi-Ju Lee [1*], Albert Yang [2, 3], Su-Yun Huang [4] and Shih-Jen Tsai [5, 6]

[1]*Taiwan International Graduate Program in Interdisciplinary Neuroscience, National Yang-Ming University and Academia Sinica, Taiwan*

[2]*Beth Israel Deaconess Medical Center, Harvard Medical School, U. S. A.*

[3] *Institute of Brain Science, National Yang-Ming University, Taiwan*

[4]*Institute of Statistical Science, Academia Sinica, Taiwan*

[5]*Department of Psychiatry, Taipei Veterans General Hospital, Taiwan*

[6]*Division of Psychiatry, School of Medicine, National Yang-Ming University, Taiwan*

Abstract

Schizophrenia as one of the major psychiatry disorders involves a complex set of neurocognitive deficits. Interdisciplinary approach such as complexity science is ideal to investigate the underlying mechanism. Power law scaling as a well-validated principle in physics is often used to evaluate dynamical systems. Such method optimizes the quantification of heterogeneous information, extracting fundamental features from spatial-temporal neuroimaging data across levels. In this research, we adopt the nonlinear property of a complexity, investigating the change of power-law characteristics in a large-scale schizophrenic and healthy resting-state fMRI data. The whole brain resting-state functional and anatomical magnetic resonance imaging data of 200 schizophrenia patients and 200 age and sex-matched healthy Han Chinese (age mean=43.56; male = 49.5% for both groups) were retrieved from Taiwan Aging and Mental Illness cohort. Pwelch function was modified to estimate the power spectra of preprocessed image signal. To compare the power-law behavior of image signal between two groups, general linear model is adopted. The results of voxel-wise exploration in grey matter (55749 voxels) show that schizophrenia patient has significantly more positive power law spectrum slope than healthy subject at 4 clusters with extent threshold of k= 35 voxels (p =0.02; corrected p (FRD-cor) <0.001 at voxel level): left precuneus, left medial dorsal nucleus, right inferior frontal gyrus, and right middle temporal gyrus. On the other hand, healthy subject shows significant higher power law slope at right putamen and left putamen. These located regions with complexity abnormality found in schizophrenia indicates over or insufficient brain activities in the schizophrenic brain, corresponds with clinical observations such as auditory hallucinations, attentional control and stop-signal inhibition. These findings support “the loss of brain complexity hypothesis”, suggesting that neuronal dynamics in healthy states exhibit multiscale variability, a characteristic of power law behavior, and pathological state of brain is associated with the breakdown of neuronal dynamics. As a useful method to distinguish pathological from healthy states in biological system, power law scaling has great potential to offer a powerful diagnostic biomarker of psychiatric disorders in the near future.

Random partition t-SNE

Szu-Han Lin, Ting-Li Chen and I-Ping Tu

Institute of Statistical Science, Academia Sinica, Taiwan

Abstract

Data visualization has been recognized as an important tool in exploring the heterogeneity of high dimensional data, for which many statistical methods have been used including Principal Component Analysis (PCA), Multidimensional Scaling (MDS) and Laplacian Eigenmaps (2003, Belkin and Niyogi). Along the development of data visualization methods, t-Distributed Stochastic Neighbor Embedding (tSNE) proposed in 2008 by Laurens van der Maaten and Geoffrey Hinton has been the first to successfully separate the 10 digit groups of MNIST data set into 10 clusters, yielding a total of 5997 citations to date. There are two key features in tSNE for its success: 1) it transforms the similarity matrix into a distribution (e.g. Gaussian distribution and t-distribution) for both the input data of high dimension and the visualization in two dimensions; 2) it minimizes the KL divergence between these two distributions by the gradient descent algorithm. However, as the data volume becomes huge, the computation for similarity matrix becomes a burden and the application faces an overwhelming barrier. We propose a random partition algorithm for t-SNE, which we name RP-tSNE, to accommodate large volume data sets for data visualization. In addition to providing a proof for the consistency of RP-tSNE, we will demonstrate the usage of RP-tSNE on a cryo-electron microscopy image data set with 103,363 images.

On estimation methods for extra zeros in crash data with missing data

Martin T. Lukusa and Frederick Kin-Hing Phoa

Institute of Statistical Science, Academia Sinica, Taiwan

Abstract

It is common that the frequencies of casualties of roads crashes exhibit a right skewed distribution characterized by a phenomenon of extra zeros. Among all count regression models, the zero-inflated models are the most appropriate tools used to account for the overdispersion and the zero-inflated features. In addition, crash data are often collected after a crash has happened and consequently, they are potentially subject to missing data. Missing data may be present in count response variable or factors that highly influence crashes occurrence. We assume that the missingness is at random and propose three types of robust estimations that implement nonparametric nuisance functions in first stage. The proposed estimators of a zero-inflated regression model in the presence of missingness are all asymptotically consistent and robust. The merit of the proposed methods is confirmed by simulation study and real examples.

Keywords：Crash data, Extra-zeros, Missing data, Nuisance functions, Robust estimators.

Effective connectivity delineates putative roles for cortical regions in emotional inhibition across aging

Siddharth Nayak[1,2], Chih-Chan Hsu[1] and Arthur C. Tsai[1]

[1]Institute of Statistical Science, Academia Sinica, Taiwan

[2]Graduate Student of Taiwan International Graduate Program Interdisciplinary Neuroscience (TIGP - INS) Academia Sinica and National Cheng Kung University (NCKU)

Abstract

In order to look at putative roles for cortical regions in emotional inhibition across aging, we designed an emotional stop signal task with 59 subjects (30 old; mean - 66 years, 29 young; mean - 24 years). Dynamic causal modelling (DCM) results highlight premotor area (PMd) – superior temporal gyrus (STG) and medial prefrontal cortex (MPFC) – basal ganglia (BG) connections as being necessary for emotional inhibition. MPFC, BG and PMd play a crucial role in cognitive control and the DCM results obtained are in line with previous studies in literature. We used four model spaces of 4 BOLD regional time series in each model space to look deeper into the neuronal connections associated with emotional inhibition. Bayesian model reduction (BMR) and Bayesian model averaging (BMA) helped us chalk down the significant models at the within-subject level. Finally, we defined a second level Parametric Empirical Bayes (PEB) model to look at between-subject connectivity. Bayesian Model comparison (BMC) was instrumental in model comparisons across groups in each of the model space we had previously defined. We have correlated between-subjects effect and questionnaire scores (trait anxiety, BIS/BAS) to look into brain - behavior correlation. Our findings give partial support to compensation-related utilization of neural circuits' hypothesis (CRUNCH). Brain - behavior correlations give insights into clinical perspective for aging while DCM gives insights on neuronal interactions involved in aging.

Keywords: Aging, Basal ganglia, Connectivity, Dynamic causal modelling, Medial prefrontal cortex.

HDMV: Visualization for high-dimensional mediation effects

Jia-Ying Su, Yen-Tsung Huang and Chun-Houh Chen

Institute of Statistical Science, Academia Sinica, Taiwan

Abstract

Hypothesis-driven mediation analyses have become popular in social science and medical research. However, there is a lack of exploratory tools for researchers to visualize high-dimensional mediation. Here we consider mediation effects of p mediators and define the total mediation effect as an inner product of the exposure-mediator association $\boldsymbol{\alpha} = \left(\alpha_1, \dots, \alpha_p\right)^T$ and the mediator-outcome association $\boldsymbol{\beta} = \left(\beta_1, \dots, \beta_p\right)^T$. Estimators for the associations $\widehat{\boldsymbol{\alpha}} = \left(\hat{\alpha}_1, \dots, \hat{\alpha}_p\right)^T$ and $\widehat{\boldsymbol{\beta}} = \left(\hat{\beta}_1, \dots, \hat{\beta}_p\right)^T$ can be obtained from standard statistical methods such as regression modeling. We propose a new proximity matrix for visualization as well as mediator clustering where the proximity matrix is constructed as a covariance of $\left(\hat{\alpha}_1\hat{\beta}_1, \dots, \hat{\alpha}_p\hat{\beta}_p\right)^T$. The proposed proximity matrix has the advantage of integrating the information about the p element-wise mediation effects and the correlation of the p mediators. The mediators can be clustered by hierarchical clustering tree guided by rank-two ellipse (HCT-R2E) algorithm based on the proposed matrix. Our simulation studies show that compared with the traditional correlation matrix, HCT-R2E using the proposed proximity matrix better classifies the effect directionality and dependence structure of mediators. We further demonstrate the utility of HDMV by a lung cancer study from The Cancer Genome Atlas (TCGA), investigating the mediation effect of smoking on lung cancer mortality mediated by a large number of DNA methylation loci.

Explaining cancer type specific mutations with transcriptomic and epigenomic features in normal tissues

Khong-Loon Tiong and Chen-Hsiang Yeang

Institute of Statistical Science, Academia Sinica, Taiwan

Abstract

Most cancer driver genes are involved in generic cellular processes such as DNA repair, cell proliferation and cell adhesion, yet their mutations are often confined to specific cancer types. To resolve this paradox, we explained mutation frequencies of selected genes across tumor types with four features in the corresponding normal tissues from cancer-free subjects: mRNA expression and chromatin accessibility of mutated genes, mRNA expressions of their neighbors in curated pathways and the protein-protein interaction network. Encouragingly, these transcriptomic/epigenomic features in normal tissues were closely associated with mutational/functional characteristics in tumors. First, chromatin accessibility was a necessary but not sufficient condition for frequent mutations. Second, variations of mutation frequencies in selected genes across tissue types were significantly associated with all four features. Third, the genes possessing significant associations between mutation frequency variations and pathway gene expression were enriched with documented cancer genes. We further proposed a novel bivariate gene set enrichment analysis and confirmed that the pathway gene expression was the dominant factor in cancer gene enrichment. These findings shed lights on the functional roles of genes in normal tissues in shaping the mutational landscape during tumor genome evolution.

Analyzing Model Bias in Cryo-EM Single-Particle Image Processing

Shao-Hsuan Wang

Institute of Statistical Science, Academia Sinica, Taiwan

Abstract

Single particle cryogenic electron microscope (cryo-EM) has become a popular method to determine the structure of biological molecules to near-atomic resolution. When cryo-EM is applied to imaging biological molecules, the data is recorded in a micrograph containing many particle projections in unknown orientations. A big challenge of cryo-EM image analysis is that the signal to noise ratio is extremely low (SNR <0.1), because the molecules are photographed with low-exposure to minimize structural degradation caused by radiation. 2D clustering, a crucial step in cryo-EM analysis, is to group projections of similar particle orientations such that the resulting averages can greatly enhance the signal to noise ratio of those abundant views and allow them to be labeled. Meaningful clustering depends on good image alignment, for which all possible rotations and translations are exhaustively searched to find the most fitted solution. However, image alignment for highly noisy data can be strongly biased toward the reference model, which is referred as model bias phenomenon. In this talk, I will show how we investigate model bias from a mathematical and statistical perspective. We propose an index to quantify model bias and provide the consistency and asymptotic theorems.

[This is a joint work with Yi-Ching Yao, Wei-Hau Chang and I-Ping Tu.]

Session 4

Association & Causation

Chair Person: Satoshi Kuriki

09:00~10:20

January 18, 2019

A Transmission Based Association Test for Multivariate Phenotypes Using Quasi Likelihood

Saurabh Ghosh

Indian Statistical Institute, Kolkata

Abstract

The classical transmission disequilibrium test (TDT) (Spielman et al. 1993) based on the trio design is an alternative to the population based case-control design to detect genetic association as it protects against population stratification. Since the manifestation of most complex diseases are governed by multiple precursor traits, it has been argued that it may be a more prudent strategy to study a multivariate phenotype comprising these precursors. One of the statistical challenges in analyzing multivariate phenotypes is to incorporate both quantitative and qualitative variables in the vector of phenotypes. We modify the classical TDT for quantitative traits based on logistic regression (Waldman et al. 1999, Haldar and Ghosh 2015) to include multivariate phenotypes. We adopt a quasi likelihood approach based on Generalized Linear Regression to develop a test of association for multivariate phenotypes. Since the Generalized Estimating Equation (GEE) approach (Gourieroux, Monfort, and Trognon 1984; Liang and Zeger 1986) used for solving the quasi likelihood equation is highly influenced by outliers, we use the modified Resistance Generalized Estimating Equation approach (RGEE) (Hall, Zeger, and Bandeen-Roche 1996, Preisser and Qaqish 1999) to down weight the outliers. We also explore a modified model that includes information on allelic transmission from both parents. We perform extensive simulations under a wide spectrum of genetic models and different correlation structures between the components of a multivariate phenotype. We compare our method with the FBAT test procedure (Lake et al. 2002) as well as separate univariate analyses of the component phenotypes and find that the proposed method that incorporates information on both parents is more powerful than the other approaches. We apply our method to analyze a multivariate phenotype related to alcoholism using data from the Collaborative Study on the Genetics of Alcoholism (COGA) project.

[This is a joint work with Hemant Kulkarni]

Causal mediation of semicompeting risks

Yen-Tsung Huang

Institute of Statistical Science, Academia Sinica, Taiwan

Abstract

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

Keywords: causal inference; causal mediation model; martingale; Nelson-Aalen estimator; semi-competing risk

Learning Co-Substructures by Kernel Dependence Maximization

Daichi Mochihashi

The Institute of Statistical Mathematics, Japan

Abstract

Associative information like "cold front passes"->"rains", "dine with a friend"->"have a happy time" are crucial for causal inference, natural language processing, market basket analysis or artificial intelligence in general.

However, it is often difficult to identify the ranges of information to be included for association, because they do not necessarily obey to the syntax.

We regard this problem as an extraction of substructures from paired sentences that maximize mutual information. Specifically, we employ HSIC (Gretton+ 2005) as a measure for dependence defined over the Euclidean space where words are embedded by word2vec, and estimate binary latent variables of each word that denotes inclusion or exclusion from the knowledge. MCMC sampling from a Gibbs distribution whose energy is HSIC optimizes these latent variables.

Experiments on synthetic and actual large corpora revealed that our proposed method has superior predictive accuracy of association over the heuristic rules employed so far.

[This is a joint work with Sho Yokoi (Tohoku University).]

References

[1] "Learning Co-Substructures by Kernel Dependence Maximization".
Sho Yokoi, Daichi Mochihashi, Ryo Takahashi, Naoaki Okazaki, Kentaro Inui. IJCAI 2017, pp.3329-3335, 2017.

Session 5

Bayes Related

Chair Person: Hsin-Cheng Huang

10:40~12:00

January 18, 2019

Robust Pseudo Bayes Estimation under Independent Non-Homogenous Set Up

Abhik Ghosh

Indian Statistical Institute, Kolkata

Abstract

Although Bayesian inference is an immensely popular paradigm among a large segment of scientists including statisticians, most of the applications consider objective priors and need critical investigations (Efron, 2013). And while it has several optimal properties, one major drawback of Bayesian inference is the lack of robustness against data contamination and model misspecification, which becomes pernicious in the use of objective priors. Using the popular Density Power Divergence, a Robustified (pseudo) Posterior Density has been developed for independent identically distributed data by Ghosh and Basu (2016), which successfully rectify the non-robustness of the Bayes inference against data contaminations and put more weights to the prior following Bayesian philosophy. This hybrid approach of robust pseudo-Bayes inference has recently been extended to the general parametric set-up by Ghosh and Basu (2017) who have also proved the exponential consistency of the resulting pseudo-posterior under general set-up.

In this talk, we will discuss the detailed applications of the robust pseudo-Bayes inference, based on the density power divergence, under the independent non-homogeneous (INH) set-ups. We have first simplified the necessary conditions for their exponential convergence results from Ghosh and Basu (2017). Then, a new Bernstein von-Mises type convergence result for the (suitably standardized) robust posterior to the normal distribution is developed under general INH set-ups along with its asymptotic expansion. These asymptotic results are then used to prove the high breakdown property of the resulting pseudo-Bayes estimators in a location model. The influence function analysis has also been developed to prove the claimed robustness properties of the proposed density power divergence based robust pseudo-posterior and related estimators. Finally our proposal has been applied to linear and logistic regression models with fixed-designs. Also, all theoretical properties and the required assumptions are simplified for the Bayes estimation of the regression coefficient, robustly against the outliers and data contaminations.

An Empirical Bayes Confidence Interval in the Presence of High Leverage for Small Area Inference

Masayo Y. Hirose

The Institute of Statistical Mathematics, Japan

Abstract

Empirical Bayes confidence interval is widely used especially when the sample size within each area is not large enough to make reliable direct estimates based on the design-based approach. Especially, there exists a second-order corrected confidence interval which achieves a smaller length than that of the confidence interval based on the direct estimates. However, this interval may have an issue in the presence of high leverage and small number of areas. In this talk, we will introduce an empirical Bayes confidence interval which has milder condition of the leverage than such empirical Bayes confidence interval. Moreover, we will also show our confidence interval being more tractable. Furthermore, we will also report the results of our simulation study for showing overall superiority of our confidence interval method over the other methods.

Hierarchical Topic Models for Tensor Count Data

Shunichi Nomura

The Institute of Statistical Mathematics, Japan

Abstract

We propose a classification method for hierarchically structured count data using extended topic models. We consider the count datasets arranged in multilevel categories and allocate topic layers to category layers at the corresponding levels. The topics in each layer are interpreted as category patterns in the lower layers. Despite of numerous combinations in multilevel categories, our model provides simple and interpretable results by sharing the topics in each layer. Hyperparameters in Dirichlet prior distributions and latent topics for respective counts are estimated by MCMC and variational Bayes methods.

We apply the proposed model to step count data recorded by activity monitors. The hourly step count is arranged in two-level categories, days of the week and hours of the day, for each one-week units. Our model identified several daily patterns as sub-topics and weekly patterns as super-topics in ambulatory activities.

[This is joint work with Michiko Watanabe and Yuko Oguma at Keio University.]

Session 6

Probability Related

Chair Person: Hideatsu Tsukahara

09:00~10:20

January 19, 2019

“*Power of Two Choices*” in De-Preferential Pólya Urn Schemes

Antar Bandyopadhyay

Indian Statistical Institute, New Delhi & Kolkata

Abstract

We will consider an implementation strategy for *weighted de-preferential* in Pólya Urn scheme. De-preferential urn schemes were first introduced by Bandyopadhyay and Kaur (2018) and Kaur (2019). Let U_n be the configuration of the urn at time n with a total of K colors. Starting with U_o, a non-empty urn, at every time step we will consider selecting d colors with replacements and reinforcing a selected color i with probability proportional to a non-increasing weight function of the proportion of selected colors. This indicating the *negative reinforcement*, in the sense that the least proportion color is most likely to be reinforced. For $d=1$ the model is trivial and is the random reinforcement model. For $d = K$ the model was first studied by Bandyopadhyay and Kaur (2019) for linear but deceasing w, and later by Kaur (2018) for general w. In this talk, we will show that for any $2 \leq d \leq K$ the almost sure convergence to uniform vector holds and we will also show that the asymptotic of the fluctuations around the limit under mild regularity conditions on w, is Gaussian. We will further show the so called “*power of two choices*” phenomenon holds here, in the sense that $d=2$ achieves the optimal asymptotic efficiency.

[This is a joint work with Gursharn Kaur]

Computation of clinch and elimination numbers in league sports based on integer programming

Satoshi Ito[1]*, Yuji Shinano[2] and Chi-Hao Wu[3]

[1] *The Institute of Statistical Mathematics, Japan*

[2] *Zuse Institute Berlin, Germany*

[3] *Institute of Statistical Science, Academia Sinica, Taiwan*

Abstract

At an early stage of a season of some sports league, if a team wins all of its remaining games, then the team will secure the pennant race; or conversely, if a team loses all of its remaining games, then the team will be in the cellar. At any moment during the season, unless a team has a chance to be eliminated from some specified situation (such as league championship or playoff berth) even when the team wins all remaining games, there exists a minimal number of future wins sufficient for the team to achieve the situation; and conversely, unless a team has a chance to achieve the specified situation even when the team loses all remaining games, there exists a minimal number of future losses sufficient for the team to be eliminated from the situation. These minimal numbers of future wins and losses are respectively called the *clinch* and *elimination* numbers for the team at the moment. We will formulate several mathematical optimization models for finding these numbers and show a generic computational framework based on integer programming for solving these optimization models in this talk.

Computing cost varies from league to league, and there are two structural factors that significantly affect the cost. One such factor is the treatment of ties (draws). If ties are not allowed (as in Major League Baseball), if a tie is converted to a fixed score (e.g., a loss, or a pair of a half win and a half loss), or similarly if some winning point system is used (as in most football leagues where three/one/zero points are awarded for a win/tie/loss), then everything will be done linearly (as far as the second factor explained below is not concerned). On the other hand, if ties are allowed and especially if the winning percentage (WP) defined as the number of wins divided by the total number of wins and losses is used for team standings (as in Taiwan's Chinese Professional Baseball League), then nonlinearity resulting from the winning percentage directly influences the computation. Note that a tie is then worth a pair of WP wins and (1 – WP) losses and the value of a tie becomes higher as the WP goes up.

The second influential factor to the computing cost is the presence of tiebreaking criteria for season standings, which plays a more crucial role from a computational point of view. Some sports leagues permit joint champions, and in some leagues one-game, a-series-of-games or round-robin tiebreaker is additionally played among the tied competitors, but some leagues provide tiebreaking criteria, which include head-to-head (considering only results of games among the tied), intra-district (in case of multiple districts in the league), scoring differential (the difference between points scored and those conceded) and so on. When a variety of tiebreaking criteria must be taken into consideration in calculating clinch and elimination numbers, the resulting model can be logically complex and computationally expensive to solve.

Keywords: clinch number, combinatorial optimization, elimination number, integer programming, round-robin tournament, sports league, tiebreaking criteria, winning percentage.

Hack's Law in a Drainage Network Model: A Brownian Web Approach

Anish Sarkar

Indian Statistical Institute, Delhi Centre

Abstract

Hack (1957) while studying the drainage system in the Shenandoah valley and the adjacent mountains of Virginia, observed a power law relation $l{\sim}a^{0.6}$ between the length l of a stream from its source to a divide and the area a of the basin that collects the precipitation contributing to the stream as tributaries. We study the tributary structure of Howard's drainage network model of headward growth and branching studied earlier by Gangopadhyay, Roy and Sarkar (2004). We show that the exponent of Hack's law is 2/3 for Howard's model. Our study is based on a scaling of the process whereby the limit of the watershed area of a stream is area of a Brownian excursion process. To obtain this, we define a dual of the model and show that under diffusive scaling, both the original network and its dual converge jointly to the standard Brownian web and its dual.

[This is a joint work with Rahul Roy and Kumarjit Saha]

Session 7

Time Series Related

Chair Person:

Antar Bandyopadhyay

10:40~12:00

January 19, 2019

Forecasting Financial Market Volatility Using a Dynamic Topic Model

Yoshinori Kawasaki

The Institute of Statistical Mathematics, Japan

Abstract

This study employs big data and text data mining techniques to forecast financial market volatility. We incorporate financial information from online news sources into time series volatility models. We categorize a topic for each news article using time stamps and analyze the chronological evolution of the topic in the set of articles using a dynamic topic model. After calculating a topic score, we develop time series models that incorporate the score to estimate and forecast realized volatility. The results of our empirical analysis suggest that the proposed models can contribute to improving forecasting accuracy.

[This is a joint work with Prof. Takayuki Morimoto of Kwansei Gakuin University.]

Keywords: dynamic topic model, forecasting, heterogeneous autoregressive model, text data, volatility.

AIC for Change-Point Models and its Application to a Biological Data Analysis

Yoshiyuki Ninomiya

The Institute of Statistical Mathematics, Japan

Abstract

Change-point problems have been studied for a long time not only because they are needed in various fields but also because change-point models contain an irregularity that requires an alternative to conventional asymptotic theory. The purpose of this study is to derive the AIC for such change-point models. The penalty term of the AIC is twice the asymptotic bias of the maximum log-likelihood, whereas it is twice the number of parameters, $2p_o$, in regular models. In change-point models, it is not twice the number of parameters, $2m+2p_m$, because of their irregularity, where m and p_m are the numbers of the change-points and the other parameters, respectively. In this study, the asymptotic bias is shown to become $6m+2p_m$, which is simple enough to conduct an easy change-point model selection. In simulation studies, it is shown that the derived AIC provides more reasonable model selection than the naive AIC and that the difference between the two AICs is not negligible. Moreover, the derived AIC is used in a biological data analysis, which is to detect a change in brain waves of a patient who has epilepsy.

Adult age differences in inhibitory control as revealed by fMRI and drift diffusion model

Arthur Chih-Hsin Tsai

Institute of Statistical Science, Academia Sinica, Taiwan

Abstract

Many situations require one to inhibit the impulse to react to others' emotional facial displays. Considerable evidence, however, points to age-related changes in response inhibition in older adults. In addition, evidence also shows more automatic regulation of affect in older compared to younger adults. How these two opposing processes operate in young and older adults remains unclear. In this study, a drift diffusion model was used to each participant's behavioral data to extract components of psychological processing, including measures of caution, motor execution time, and stimulus processing speed. The individual differences in the resulting components were further examined in relation to fMRI activation to evaluate the underlying neural correlates and the emotional modulation in young and older adults. Thirty-two young and 32 older normal adults underwent an emotional stop signal paradigm (ESSP) in an fMRI experiment with disgusted and neutral emotional faces. Participants were instructed to make a target response (Go) as quickly as possible to face stimuli unless a red border appeared (between 40 to 400 ms of face onset), in which case they were to withhold their response (Stop). Young adults made more Stop than Go errors to neutral faces. Disgusted faces increased Stop false alarms, decreased Go hits, and also reduced reaction times compared to neutral faces. Older adults made more Go than Stop errors to neutral faces. Disgusted faces increased Stop false alarms, increased Go Hits, and also reduced reaction times compared to neutral faces. Two-sample t-tests are used to delineate the role of aging in emotional inhibition tasks. The results show that older adults tend to engage more neural circuits than young adults including striatum and hippocampus regions. Younger adults show more focal activations in right inferior frontal cortex, inferior parietal lobule and postcentral gyrus for inhibition contrasts. Using drift-diffusion modeling principles, our findings suggest whereas negative affect increases the rate of evidence accumulation for target and stop feature decisions in younger adults, negative affect lowers the decision criterion in older adults.

With contribution by: Siddharth Nayak, Chih-Chan Hsu, Chii-Shyang Kuo, Shuo-Heng Li, Joshua Oon Soo Goh, Yi-Ping Chao, Li Jingling, and Su-Ling Yeh

Keywords: drift-diffusion model, emotional inhibition, emotional stop signal paradigm, fMRI, ICA

Session 8

Bootstrap Resampling

Chair Person: Shuhei Mano

13:30~14:20

January 19, 2019

Nonparametric Confidence Band for Activity Profiles Based on Wearable Device Data

Hsin-wen Chang[1]* and Ian McKeague[2]

[1]Institute of Statistical Science, Academia Sinica, Taiwan

[2] Columbia University, United States

Abstract

The motivation for this talk comes from applications to health care monitoring in which there is a need to analyze activity profiles constructed from wearable device data. We introduce a nonparametric likelihood ratio approach that makes efficient use of the activity profiles to provide a confidence band for their means. The procedure is calibrated using bootstrap resampling. A simulation study shows that the proposed procedure outperforms competing Wald-type functional data approaches. We illustrate the proposed methods using wearable device data from an NHANES study.

A non-parametric method for inferring food web parameters

Wei-Chung Liu

Institute of Statistical Science, Academia Sinica, Taiwan

Abstract

A food web is a representation of who eats whom in an ecosystem. Food webs are often generated by a single dataset aggregated from one or several surveys. Point estimates of food web parameters can be calculated from the data, but how to quantify their corresponding interval estimates still remains elusive. Here, a simple bootstrap-based resampling procedure is proposed for inferring food web parameters. First, for a particular food web parameter, we obtain its point estimate by calculating the corresponding statistics from the original food web. Second, we generate a resampled food web by sampling with replacement the same number of species from the original food web, and for each resampled species we record how many prey items it consumes in the original food web. Third, a resampled species is allowed to consume its original prey species if such a species is also present; if not then it instead consumes the resampled species that is most topologically similar to its original prey species. Several resample food webs can be constructed from which the sampling distribution and the interval estimate for this particular statistics can then be determined. We demonstrate our methodology on two different food web datasets and discuss its application in comparing food webs of various sizes and complexity.

Institute of Statistical Science, Academia Sinica