

Model Specification and Time-Varying Risk Premia: Evidence from Spot and Option Markets

Chang-Shu Chung* (鍾長恕)
National Chengchi University
Ting-Fu Chen (陳亭甫)
Feng Chia University
Shih-Kuei Lin (林士貴)
National Chengchi University

Abstract

In this paper, we attempt to answer four questions: (i) On average, what do the proportions of the stochastic volatility and return jumps account for the total return variations in S&P 500 index? In particular, which one has more influence than the other does on the total return variations? (ii) Is the fitting performance of infinite-activity jump models better than that of finite-activity jump models both in the spot and option markets? (iii) When would investors require significantly higher risk premiums? Specifically, were there significant changes in volatility and jump risk premiums during financial shocks? (iv) Whether the variance risk premiums have the predictive power on S&P 500 index returns, especially can a portfolio based on the diffusive variance risk premiums (DVRPs) gain excess returns? For the first question, we find that most of return variations are explained by the stochastic volatility, and the return jumps account for the higher percentage than the stochastic volatility at the beginning of financial crises. For the second question, we adopt the dynamic joint estimation to obtain the stochastic volatility model with double-exponential jumps and correlated jumps in volatility and the stochastic volatility model with normal-inverse Gaussian jumps fit S&P 500 index returns and options well in different criteria. For the third question, the time-varying risk premiums show that the jump risk premiums increase after the shock of the recent financial crisis, which implies that the panic of bearing jump risk in the post-crisis period causes more expected returns. For the fourth question, we find that DVRPs have the predictive power on S&P 500 index returns both in-sample and out-of-sample, with R-squared statistics of 5.40% and 3.46%, respectively. Finally, we further investigate the economic significance of the out-of-sample predictability on the basis of asset allocations with DVRPs, and the mean-variance portfolio generates substantial economic gains of over 166 basis points per annum.

Keywords:

Jump Risk Premium, Variance Risk Premium, Particle Filter, Dynamic Joint Estimation, Return Predictability.

Parameter and Interval Estimation For A Positive Autoregressive Process

HAO-YUN HUANG* (黃灝勻) 、 CHING-KANG ING (銀慶剛)

Institute of Statistics, National Tsing Hua University

WEI-YING WU (吳韋瑩)

Department of Applied Mathematics, National Dong Hwa University

Abstract

Starting from the non-negative autoregressive AR(1) model with i.i.d positive innovations, we introduce how the left or right tail behavior of the underlying error distribution affects the asymptotic property of the extreme value estimator (EVE) of the AR coefficient. Without such tail information, it is difficult to construct the confidence interval for the autoregressive (AR) coefficient via EVE. To overcome this difficulty, a procedure is proposed to detect the unknown tail behaviors first, and then construct the desired confidence interval via EVE based on the detected information. For the general non-negative autoregressive AR(p), by applying the similar procedure on linear programming estimator (LPE) in Feigin and Resnick (1997), our method can be extended to develop the confidence region for AR coefficients with the bootstrapping scheme. The performance of the proposed estimate are investigated through theoretical studies and simulation results.

Key words: Positive autoregressive processes, linear programming estimates, extreme-value estimates, regular variation indices, heavy-tailed data

Greedy variable selection for high-dimensional Cox models

Chien-Tong Lin, Yu-Jen Cheng

Institute of Statistics

National Tsing Hua University

Ching-Kang Ing

Institute of Statistics

National Tsing Hua University

Abstract

Model selection for sparse high-dimensional Cox models has broad applications to contemporary biostatistics, in particular, to extracting relevant biomarkers from high-dimensional survival data. In this talk, we propose using a greedy-type algorithm, Chebyshev Greedy Algorithm (CGA), to iteratively include covariates in the aforementioned models, and show that with probability tending to one, all relevant covariates can be included in a moderate number of iterations. We also devise a high-dimensional information criterion (HDIC) to remove the redundant covariates chosen by CGA, thereby leading to selection consistency. Finally, the proposed method is illustrated using simulated data and a diffuse large B-cell lymphoma (DLBCL) dataset.

Keywords: Chebyshev greedy algorithm, Cox model, high-dimensional information criterion, sure screening, variable selection consistency.

A SYSTEMATIC CONSTRUCTION OF COST-EFFICIENT DESIGNS FOR ORDER-OF-ADDITION EXPERIMENTS

JING-WEN HUANG(黃靖雯)、FREDERICK KIN HING PHOA(潘建興)

Institute of Statistics, National Tsing Hua University, Taiwan

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.

關鍵詞：Order-of-Addition Experiments, Cost Efficiency, Experimental Designs, Clinical Trials.

Title: Reflection on two incomplete research projects

Presenter: **Hung Chen,**

Department of Mathematics, National Taiwan University.

Abstract:

In this talk, I will present two incomplete research projects with my two graduate students. In the first project, we try to understand why LASSO works and start with the case $n=p$. It shows that Lasso won't work when the p is of the order n . However, we do not understand its reason until we realize that high-dimension geometry is quite different from low-dimension geometry. In the second project, we study estimation in hidden Markov models. The standard approach is to use EM algorithm or resorts to the marginal approach through mixture distributions. However, the approach based on mixture distribution does not utilize information of transition hidden in its surrogate. Then we realize that this limitation has been noticed in the literature. However, no general solution is being proposed and we also do not have a systematical approach to address this problem.

Is the inclusion of all data always a good idea? Inference about
marginal characteristics in paired design might shed some
insight into the issue

鄒宗山

中央大學統計研究所

摘要

Achieving optimality in statistical analysis, more often than not, entails the inclusion of all data. This practice might not always be appropriate and, very often, might complicate the analysis unnecessarily. Inference about marginal characteristics in paired design is a good example. The comparison of predictive values of two diagnostic tests in the paired scenario exemplifies the issue.

ICH E9 (R1): Estimands and Sensitivity Analyses in Clinical Trials

王玫，醫藥品查驗中心

Abstract

ICH E9 Statistical Principles for Clinical Trials was issued in 1998. In October 2014, the Assembly of the International Council on Harmonization endorsed the formation of an expert working group to develop an addendum to ICH E9, *i.e.* ICH E9/R1. As a first major ICH statistical document over the two decades, the addendum provides a structured framework to link trial objective to proper trial design and methods for estimation and hypothesis testing. In addition, the framework introduces the concept of estimand and highlighted the importance of intercurrent events in translating the trial objective to precise definition of the treatment effect that is to be estimated.

Missing data is "*data that would be meaningful for the analysis of a given estimand but were not collected*" (E9/R1 Glossary). The missing data problem is commonly encountered in randomized controlled trials. Even defining estimands adequately, addressing intercurrent events in a proper manner, and making best efforts to collect the data required for estimation, some data may still be missing. In addition, the value reflecting the scientific question of interest using hypothetical strategy usually cannot be observed and need to be implicitly or explicitly predicted or imputed as well. Imputation of missing or unobserved data should be based on plausible assumptions. Inferences of a particular estimand should be robust to deviations from the assumptions used in the main estimator. This robustness need to be evaluated through a sensitivity analysis.

In this talk, the framework of the document will be introduced. One or two examples will be presented to demonstrate the concept.

Recent Developments of Conditionally Specified Distributions

Yuchung J. Wang (王友鐘)

Rutgers University, Camden, New Jersey 08102

ABSTRACT

One modeling strategy that has received more and more attentions recently is the conditionally specified distribution (CSD), which uses a collection of conditional models to identify a d -dimensional joint distribution. Bayesian network (BN) is a special kind of CSD because its flow of causality is forward always. We will discuss the general CSD, also known as dependence network (DN), in which the flows of reasoning form a feedback loop. DN uses pseudo-Gibbs sampler to approximate the joint distribution. Kou and Wang (2019) studied the stationary distributions of pseudo-Gibbs sampler, known as the pseudo-Gibbs distributions (PGD). Multiple imputation via chained equations (MICE) also seeks sampling values of PGD. I will concentrate on three recent researches: (a) the shared marginal distributions among $d!$ PGD; (b) among $d!$ scan sequences, which visiting schedules will lead to the correct stationary distribution (Kou and Wang, 2018); (c) efficient computations of exact PGD and verification of compatibility among the conditional models (Kou and Wang, 2019). In this talk, we show how our study of incompatible CSD has led to better understanding of the original Gibbs sampling and efficient calculation of the Gibbs distributions.

Key words: Dependence Network; Gibbs Sampling; Scan Order; Stationary Distribution.

Semiparametric Copula-based Analysis for Treatment Effects in the Presence of Treatment Switching

Chia-Hui Huang¹, Yi-Hau Chen², Jinn-Li Wang^{3,4}, and Mey Wang⁵

¹Department of Statistics, National Taipei University, New Taipei, Taiwan

²Institute of Statistical Science, Academia Sinica, Taipei, Taiwan

³Division of Hematology Oncology, Department of Pediatrics, Wan Fang Hospital, Taipei Medical University, Taipei, Taiwan

⁴Department of Pediatrics, School of Medicine, College of Medicine, Taipei Medical University, Taipei, Taiwan

⁵Center for Drug Evaluation, Taipei, Taiwan

Abstract

In controlled trials, “treatment switching” occurs when patients in one treatment group switch to the alternative treatment during the trial, and poses challenges to evaluation of the treatment effects owing to crossover of the treatments groups. In this work, we assume that treatment switches occur after some disease progression event, and view the progression and death events as two semicompeting risks. The proposed model consists of a copula model for the joint distribution of time-to-progression (TTP) and overall survival (OS) before the earlier of the two events, as well as a conditional hazard model for OS subsequent to progression. The copula model facilitates assessing the marginal distributions of TTP and OS separately from the association between the two events, and, in particular, the treatment effects on TTP and on OS in the absence of treatment switching. The proposed conditional hazard model for death subsequent to progression allows us to assess the treatment switching (crossover) effect on OS given occurrence of progression and covariates. General semiparametric transformation models are employed in the marginal models for TTP and OS. A nonparametric maximum likelihood procedure is developed for model inference, which is verified through asymptotic theory and simulation studies. The proposed analysis is applied to a lung cancer dataset to illustrate its real utility.

Keywords: Maximum likelihood; Oncology trial; Semicompeting risks; Transformation model.

Overlapping Group Screening for Detection of Gene-gene Interactions: Application to Gene Expression Profiles with Survival Trait

王价辉*、程毅豪

中央研究院 統計科學研究所

ABSTRACT

Background: The development of a disease is a complex process that may result from joint effects of multiple genes. In this article, we propose the overlapping group screening (OGS) approach to determining active genes and gene-gene interactions incorporating prior pathway information. The OGS method is developed to overcome the challenges in genome-wide data analysis that the number of the genes and gene-gene interactions is far greater than the sample size, and the pathways generally overlap with one another. The OGS method is further proposed for patients' survival prediction based on gene expression data.

Results: Simulation studies demonstrate that the performance of the OGS approach in identifying the true main and interaction effects is good and the survival prediction accuracy of OGS with the Lasso penalty is better than the ordinary Lasso method. In real data analysis, we identify several significant genes and/or epistasis interactions that are associated with clinical survival outcomes of diffuse large B-cell lymphoma (DLBCL) and non-small-cell lung cancer (NSCLC) by utilizing prior pathway information from the KEGG pathway and the GO biological process databases, respectively.

Conclusions: The OGS approach is useful for selecting important genes and epistasis interactions in the ultra-high dimensional feature space. The prediction ability of OGS with the Lasso penalty is better than existing methods. The OGS approach is generally applicable to various types of outcome data (quantitative, qualitative, censored event time data) and regression models (e.g. linear, logistic, and Cox's regression models).

Keywords: *gene-gene interaction, Lasso, overlapping group, survival prediction*

Robust likelihood analysis of panel count and interval-censored data

Chi-Chung Wen* (溫啟仲)

Tamkang University, Taiwan

Yi-Hau Chen

Academia Sinica, Taiwan

Chi-Hong Tseng

David Geffen School of Medicine at UCLA, USA

摘要

We propose a joint analysis simultaneously analyzing recurrent and nonrecurrent events subject to general types of interval censoring. The proposed analysis allows for general semiparametric models, including the classes of Box-Cox transformation and inverse Box-Cox transformation models for the recurrent and nonrecurrent events, respectively. A frailty variable is used to account for the potential dependence between the recurrent and non-recurrent event processes. We apply the pseudo likelihood for interval-censored recurrent event data, usually termed as panel count data, and the sufficient likelihood for interval-censored non-recurrent event data. Conditioning on the sufficient statistic for the frailty, and using the working assumption of independence over examination times, the sufficient likelihood does not rely on distributional assumptions on the frailty, and can deal with general interval censorship. We illustrate the proposed methodology by a joint analysis of the numbers of occurrences of basal cell carcinoma over time and time to the first

recurrence of squamous cell carcinoma based on a skin cancer dataset, as well as a joint analysis of the numbers of adverse events and time to premature withdrawal from study medication based on a scleroderma lung disease dataset.

關鍵詞：Correlated data; Joint model; Recurrent event; Semiparametric model; Survival analysis.

論文題目: Speech Signal Processing for Assistive Hearing and Speaking Devices

曹昱

中央研究院資訊科技創新研究中心

摘要

With the rapid advancement in speech processing technologies and in-depth understanding of human speech perception mechanism, significant improvement has been made in the design of assistive hearing devices [assistive listening device (ALD), hearing aids (HAs), and cochlear implants (CIs)] to benefit the speech communication for millions of hearing-impaired patients and subsequently enhance their quality of life. However, there are still many technical challenges, such as designing noise-suppression algorithms catered for ALD, HA, and CI users, deriving optimal compression strategies, improving the music appreciation, optimizing speech processing strategies for users speaking tonal languages, to name a few. In the first part of my talk, I will present our recent research achievements using machine learning and signal processing on improving speech perception abilities for ALD, HA, and CI users.

In the second part of my talk, I will present our recent progress of developing machine-learning-based assistive speaking devices. *Oral cancer ranks in the top five of all cancers in Taiwan.* To treat the oral cancer, surgical processes are often required to have parts of the patients' articulators removed. Because of the removal of parts of the articulator, a patient's speech may be distorted and difficult to understand. To

overcome this problem, we propose two voice conversion (VC) approaches: the first one is the joint dictionary training non-negative matrix factorization (JD-NMF), and the second one is the end-to-end generative adversarial network (GAN)-based unsupervised VC model. Experimental results show that both approaches can be applied to convert the distorted speech such that it is clear and more intelligible.

關鍵詞：

Deep learning, speech enhancement, voice conversion, assistive technologies

Textual Data Analytics in Finance

Chuan-Ju Wang*

Research Center for Information Technology Innovation, Academia Sinica

Ming-Feng Tsai

Department of Computer Science, National Chengchi University

Po-Chuan Chien

Research Center for Information Technology Innovation, Academia Sinica

摘要

The growing amount of public financial data makes it more and more important to learn how to discover valuable information for financial decision-making. This talk presents our recent studies on exploring and mining soft information in financial reports. This talk will cover several machine learning techniques, such as learning to rank and word embedding, on financial reports for the study of financial risk among companies and for discovering new finance keywords. A brief demonstration on our developed web-based information systems, Fin10K, will be given to show its ability to facilitate the analysis on textual information in finance.

關鍵詞：

unstructured data, volatility, financial risk, ranking, continuous language model

Style-Oriented Landmark Retrieval and Summarization

Wei-Yi Chang¹, Yi-Ren Yeh^{*2}, and Yu-Chiang Frank Wang¹

¹Research Center for IT Innovation, Academia Sinica, Taiwan

²Department of Mathematics, National Kaohsiung Normal University, Taiwan

Abstract

While the task of visual summarization aims to select representative images from an image collection, we solve a unique problem of style-oriented landmark retrieval and summarization from photographic images of a city. Instead of performing summarization or clustering on landmark images from a city, we allow the user to provide a query input which is not from the city of interest, while the goal is to retrieve and summarize the landmark images with similar style-dependent landmark images, followed by a style-consistent image summarization across landmark categories. As a result, our summarized outputs from various landmarks would exhibit similar image style as that of the query. Our experiments will confirm that the use of our proposed method is able to perform favorably against existing or baseline approaches with improved query-dependent style consistency.

Keywords: image retrieval, image summarization

Convergence of mirror descent with Armijo line search

Yen-Huan Li*

Department of Computer Science and Information Engineering
National Taiwan University

Carlos A. Riofrio

Dahlem Center for Complex Quantum Systems
Freie Universität Berlin

Volkan Cevher

Laboratory for Information and Inference Systems
École Polytechnique Fédérale de Lausanne

摘要

Standard machine learning approaches, such as maximum likelihood (ML) estimation and empirical risk minimization, require solving convex optimization problems. Most existing convex optimization algorithms require the objective function to be Lipschitz or smooth relative to a Legendre function; otherwise they lack convergence guarantees. The Lipschitz and relative smoothness conditions, however, do not hold in important applications, such as growth-optimal portfolio selection and ML quantum state tomography. In this work, we propose a local relative smoothness condition that generalizes its existing global counterpart, and prove that under the local relative smoothness condition, the associated mirror descent algorithm with Armijo line search always converges. Numerical results showed that, therefore, the mirror descent algorithm with Armijo line search was the fastest guaranteed-to-converge algorithm for quantum state tomography, empirically on real data-sets.

關鍵詞：Mirror descent, Armijo line search, relative smoothness

以因果圓派模式進行疾病歸因

李文宗

台灣大學流行病學與預防醫學研究所

摘要

因果圓派模式是流行病學因果推論重要的概念架構，近年來在方法學上更有重要突破，得以用於流行病學實際資料之分析。本演講介紹以因果圓派模式進行疾病歸因的方法。本演講介紹如何計算因果圓派成圓率、因果圓派權重、以及可歸因危險性。此方法得以將疾病分別歸因於各危險因子之獨立作用，以及其間之各種可能交互作用。此處之歸因具有機率之概念，並且滿足歸因機率總合為1的特性。除了危險因子及其交互作用外，此方法也可對中介因子進行適當的歸因。在一個二元性危險因子以及一個二元性中介因子的情境下，疾病的致病過程共有六種可能途徑。本方法可以進行路徑歸因，分別計算各路徑的權重機率。本演講所介紹疾病歸因方法，對於疾病致病機轉的瞭解、公共衛生介入方案可能效果的估計、法庭上侵權行為責任歸屬的判定等等，應有所助益。

關鍵詞： 因果圓派模式，因果推論，流行病學方法論，疾病歸因，危險因子，中介因子，交互作用。

Coups and Democratization: A Mediation Analysis

Yen-Tsung Huang (黃彥棕) and Wen-Chin Wu* (吳文欽)

Academia Sinica

Abstract

Do coups facilitate democratization under dictatorships? Recent studies are inconclusive on this question. On the one hand, coups are good for democratization because the post-coup governance will be improved as the political leaders who have survived coups or replaced the previous ones would adopt policies that enhance their political legitimacy. On the other hand, after coups political leaders may coercively suppress the opposition forces to reconsolidate their authoritarian rules. In this paper, we adopt a mediation analysis to reconcile this debate. In particular, we argue that some dictatorships need to improve their governance or even choose to democratize after coups but other do not. As a result, coups have different mediation effects among different authoritarian regimes on democratization. With the data of dictatorships from 1950 to 2010, we find that coups are more likely to facilitate democratization among military dictatorships but not others. Our finding shed light on literature of coups and comparative democratization.

Keywords: coup, democratization, dictatorships, mediation analysis

Causal mediation of semi-competing risks

黃彥棕

中央研究院統計學研究所

摘要

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

關鍵詞：causal inference; causal mediation model; martingale; Nelson-Aalen estimator; semi-competing risks

Estimating the Dynamic Effects of Regime Transitions by Local Projections

Hsin-Yi Lin
Yu-Hsiang Hsiao

Abstract

We employ Local Projections to estimate the responses of outcomes to regime transitions. We emphasize that regime transitions have two special features: First, the effects of new regimes and mature regimes can be different. Second, transitions to the new regime and reversals to the original one can have different effect sizes. We show that splitting the panel data into two non-overlapped subsamples and respectively estimating the first-difference Local Projections estimation can properly capture the above two features of regime tr

服務創新參與的前因與創新績效之探討：雙元機制的觀點

白佩玉*

國立政治大學企業管理學系

蔡顯童

國立台北大學企業管理學系

鍾君宇

台北醫學大學通識教育中心

摘要

本研究根據心理學及組織行為相關研究，提出「服務創新參與」之概念性研究架構，探究服務創新參與及創新績效的前因及調節因素。本研究以台灣資訊服務業為實證情境，結合資訊服務業員工的兩階段問卷資料及員工實際創新表現的客觀資料，並以結構方程模式進行假設驗證。本研究發現：員工服務創新績效受到員工服務創新投入的影響，而員工對於服務創新的自我效能感及組織認同為服務創新參與的雙元機制（心理機制及社會機制）。本研究亦發現：顧客參與及同事支持有助於服務創新自我效能的提升；同事支持及成員相似性也有助於員工認同感深化。此外，員工的顧客導向程度，以及員工感到受信任的程度，皆能強化其服務創新與服務績效的正向關係。在學理上，本研究延伸了服務創新等相關研究，探討資訊服務業員工參與服務創新的決策過程。透過雙元影響途徑，讓企業主能同時考量員工投入服務創新的心理及社會機制，以及影響服務創新參與及服務創新績效的形成過程。根據實證結果，本研究也提出具體建議供企業參考。

關鍵詞：服務創新、自我效能感、社會認同、顧客導向

房價增值會提高居住滿意度嗎？

李美杏

助理教授，台北大學統計系，新北市三峽區大學路 151 號，E-mail:

mshing@mail.ntpu.edu.tw

彭建文

教授，台北大學不動產與城鄉環境學系，新北市三峽區大學路 151 號，E-mail:

cwpeng@mail.ntpu.edu.tw

陳冠儒

博士班研究生，台北大學不動產與城鄉環境學系，新北市三峽區大學路 151 號

E-mail: s810573101@gm.ntpu.edu.tw

摘要

本文目的在探討房屋的持有時間與近年來房價的大幅增值對居住滿意度的影響，採用行政院主計處「104 年住宅狀況抽樣調查」資料，分別針對室內滿意度、室外滿意度與整體滿意度進行分析。實證結果顯示：一、在住宅持有情況部分，持有時間愈長，室內與整體滿意度愈低，由於持有時間同時與屋齡和增值幅度呈高度正相關，說明持有時間增加可能產生兩個效果，一是屋齡增加導致的實質環境品質下降，另一是持有時間增加導致增值幅度增加，理論上前者對滿意度的影響是負向，後者的影響則不一定，最後淨效果為負向。屋齡與持有年數交叉項係數對室內、室外與整體滿意度的影響顯著為正，顯示當屋齡高且持有時間也長的住戶有較高的滿意度，說明繼續居住的正面效果仍高於屋齡增加所產生的負面影響，可以視為家戶因為對環境滿意而長期居住，而非無力換屋勉強居住；屋齡低且持有年數短的住戶也會有較高的滿意度，但若是屋齡偏高且持有期間短，也就是購買屋齡愈久的中古屋，則會有較低的滿意度。二、持有期間的房價增值對室內與整體滿意度的影響在各地區不同，但房價增值對室外滿意度有負向影響，在各地區的顯著性不一。房價增值可能來自於鄰里環境的改善，也可能來自於金融市場的影響，前者對滿意度應有正向效果，後者則未必；再者，房價增值也可能帶來財富效果，而提高對居住的滿意度，三者淨效果在各地區的影響不同。在全國房價漲幅最高的台北市，是唯一房價增值對室內滿意度有正向影響的縣市，說明在此地區財富效果的影響大於因房價增值帶來的負向影響。

關鍵詞：居住滿意度、室內居住環境滿意度、住宅戶外環境滿意度。

Using Mendelian Randomization Approach to Infer the Causality between Serum Uric Acid and the Risk Factors of Cardiovascular Disease

Kuang-Mao Chiang (姜廣茂) 、Mahantesh I. Biradar
Institute of Biomedical Sciences, Academia Sinica, Taipei, Taiwan

Hsin-Chou Yang(楊欣洲) 、Huang, Yen-Tsung(黃彥棕)
Institute of Statistical Science, Academia Sinica, Taipei, Taiwan

Wen-Harn Pan (潘文涵)
Institute of Biomedical Sciences, Academia Sinica, Taipei, Taiwan

摘要

Serum uric acid (SUA) has gradually been recognized as a potential risk factor for cardiovascular disease and has been found to cluster with multiple components of metabolic syndrome (MetS). However, whether the relationship is causal remains controversial.

Taking advantage of the data from 10,000 participants of Taiwan Biobank, we carried out Mendelian randomization analyses with SUA weighted genetic risk score (WGRS) as instrumental variables and metabolic syndrome components as the outcomes. Since central obesity is generally considered as the core of MetS, we also generate the waist circumference WGRS to investigate the causal relationship between central obesity and the other MetS components as well as SUA.

Using generalized linear model, we found that SUA-WGRS was significantly associated with SBP, DBP, triglyceride (positively for the above), and HDL-C

(negatively). But it was not associated with fasting serum glucose, waist circumference or BMI. On the other hand, waist circumference WGRS was associated with all the components of MetS including uric acid.

Our MR investigation confirms that central obesity is a causal agent of all MetS components including hyperuricemia and suggests that hyperuricemia does augment risk of MetS, but not through accumulating fat or hyperglycemia.

關鍵詞：

Mendelian randomization, Serum uric acid, cardiovascular disease, metabolic syndrome, causality.

drVM: a new tool for efficient genome assembly of known eukaryotic viruses from metagenomes

Hsin-Hung Lin and Yu-Chieh Liao*

Institute of Population Health Sciences, National Health Research Institutes, Miaoli,
Taiwan

Abstract

Virus discovery using high-throughput next-generation sequencing (NGS) has become more commonplace. However, although analysis of deep NGS data allows us to identify potential pathogens, the entire analytical procedure requires competency in the bioinformatics domain, which includes implementing proper software packages and preparing prerequisite databases. Simple and user-friendly bioinformatics pipelines are urgently required to obtain complete viral genome sequences from metagenomic data.

This manuscript presents a pipeline, drVM (detect and reconstruct known viral genomes from metagenomes), for rapid viral read identification, genus-level read partition, read normalization, de novo assembly, sequence annotation and coverage profiling. The first two procedures and sequence annotation rely on known viral genomes as a reference database. drVM was validated via the analysis of over 300 sequencing runs generated by Illumina and Ion Torrent platforms to provide complete viral genome assemblies for a variety of virus types including DNA viruses, RNA viruses and retroviruses. drVM is available for free download at: <https://sourceforge.net/projects/sb2nhri/files/drVM/> and is also assembled as a Docker container, an Amazon machine image and a virtual machine to facilitate seamless deployment.

drVM was compared with other viral detection tools to demonstrate its merits in terms of viral genome completeness and reduced computation time. This substantiates the platform's potential to produce prompt and accurate viral genome sequences from clinical samples.

Keywords

Next-generation sequencing (NGS), bioinformatics, metagenomics, detection, reconstruction

Pharmacometabolomics study of four anti-hypertensive drugs
in the Han Chinese population of Taiwan

Yu-Jen Liang (梁佑任)

Institute of Statistical Science, Academia Sinica

Kuang-Mao Chiang (姜廣茂)

Institute of Biomedical Sciences, Academia Sinica

Chia-Min Cheung (鍾加明)

Graduate Institute of Clinical Medical Science, China Medical University

Ming-Shi Shiao (蕭明熙), Mei-Ling Cheng (鄭美玲)

Department of Biomedical Sciences, College of Medicine, Chang Gung University

Hsin-Chou Yang (楊欣洲)*

Institute of Statistical Science, Academia Sinica

Wen-Harn Pan (潘文涵)*

Institute of Biomedical Sciences, Academia Sinica

Abstract

In this study, we conducted liquid chromatography-mass spectrometry (LC-MS) based untargeted and targeted pharmacometabolomics study to evaluate pharmacometabolomics responses to hypertensive drugs in young-onset hypertensive patients in the Han Chinese population of Taiwan. We recruited 250 young-onset hypertensive patients from five study groups containing 50 non-medication patients and 50 patients for each of four medication groups. Our software Statistical Metabolomics Analysis—An R Tool (SMART) was applied to analyze the data from the initial data preparation, preprocessing, quality control, batch effect detection to the downstream association analysis. In summary, in total, 4,159 molecular features were found in 1,492 replicate samples. The association analysis identified two drug-associated molecular features significantly associated with the four studied drugs commonly. The two features are the phospholipid-derived molecules that can modify the function of physicochemical properties of membrane and alter ion channel transport systems for Na^+ , K^+ , and Ca^{2+} and increase vascular tone in essential hypertension patients.

Keywords: Pharmacometabolomics, Hypertension, SMART, anti-hypertensive drugs.

General retrospective mega-analysis framework for rare variant association tests

Li-Chu Chien (簡莉珠)^{1*}, Yen-Feng Chiu (邱燕楓)²

¹Center for Fundamental Science, Kaohsiung Medical University, Kaohsiung,
Taiwan

²Institute of Population Health Sciences, National Health Research Institutes, Miaoli,
Taiwan

ABSTRACT

Here, we describe a retrospective mega-analysis framework for gene- or region-based multimarker rare variant association tests. Our proposed mega-analysis association tests allow investigators to combine longitudinal and cross-sectional family- and/or population-based studies. This framework can be applied to a continuous, categorical or survival trait. In addition to autosomal variants, the tests can be applied to conduct mega-analyses on X-chromosome variants. Tests were built on study-specific region- or gene-level quasi-score statistics and, therefore, do not require estimates of effects of individual rare variants. We used the generalized estimating equation approach to account for complex multiple correlation structures between family members, repeated measurements, and genetic markers. While accounting for multilevel correlations and heterogeneity across studies, the test statistics were computationally efficient and feasible for large-scale sequencing studies. The retrospective aspect of association tests helps alleviate bias due to phenotype-related sampling and type I errors due to misspecification of phenotypic

distribution. We evaluated our developed mega-analysis methods through comprehensive simulations with varying sample sizes, covariates, population stratification structures, and study designs across multiple studies. To illustrate application of the proposed framework, we conducted a mega-association analysis combining a longitudinal family study and a cross-sectional case-control study from Genetic Analysis Workshop 19.

KEYWORDS:

pedigree data; mixed study design; retrospective association tests; family study

The adaptive combination of Bayes factors method and its applications to genetic studies

林苑俞

Institute of Epidemiology and Preventive Medicine, College of Public Health,
National Taiwan University, Taipei, Taiwan; Department of Public Health, College of
Public Health, National Taiwan University, Taipei, Taiwan

摘要

The adaptive combination of Bayes factors (ADABF) method can serve as a powerful test for the joint analysis of rare and common variants. Most gene-based association tests are underpowered given a large proportion of neutral variants within a gene. Our ADABF removes the variants with smaller Bayes factors and so it is robust to the inclusion of neutral variants. The ADABF method is more powerful than other association tests when there are only few variants (in a gene or a region) associated with the phenotypes. It can be applied to genome-wide association studies or next-generation sequencing data, continuous traits or dichotomous traits, unrelated subjects or case-parent trios, and it allows for covariates adjustment. Besides, more than other gene-based association tests, the ADABF method further indicates which variants enrich the significant association signal. Furthermore, ADABF is recommended for genome-wide ‘gene–environment interactions’ ($G \times E$) analyses. Recently, it has also been developed as a polygenic approach to detect $G \times E$ when external information is unavailable.

關鍵詞：diastolic blood pressure, systolic blood pressure, gene–alcohol interaction, Taiwan Biobank, multiple-testing correction

Compound Cox: univariate feature selection and compound covariate for predicting survival

Takeshi Emura (Speaker)

Graduate Institute of Statistics, National Central University, Taiwan

Shigeyuki Matsui

Department of Biostatistics, Nagoya University Graduate School of Medicine, Japan

Hsuan-Yu Chen

Institute of Statistical Science, Academia Sinica, Taiwan

Abstract: Univariate feature selection is one of the simplest and most commonly used techniques to develop a multigene predictor for survival for cancer patients. Presently, there is no software tailored to perform univariate feature selection and predictor construction. We develop the `compound.Cox` R package [1] that implements univariate significance tests (via the Wald tests or score tests) for feature selection. We provide a cross-validation algorithm to measure predictive capability of selected genes and a permutation algorithm to assess the false discovery rate. We also provide three algorithms for constructing a multigene predictor (compound covariate, compound shrinkage, and copula-based methods), which are tailored to the subset of genes obtained from univariate feature selection. We demonstrate our package using survival data on the lung cancer patients. We examine the predictive capability of the developed algorithms by the lung cancer data and simulated data. The developed R package, `compound.Cox`, is available on the CRAN repository. The statistical tools in `compound.Cox` allow researchers to determine an optimal significance level of the tests, thus providing researchers an optimal subset of genes for prediction. The package also allows researchers to compute the false discovery rate and various prediction algorithms.

[1] Emura T, Matsui S, Chen HY (2019) `compound.Cox`: univariate feature selection and compound covariate for predicting survival, *Comput Methods Programs Biomed* 168: 21-37

關鍵詞： cancer prognosis, copula, Cox regression, cross-validation, dependent censoring, false discovery rate, gene expression, multiple testing

A measure for comparing upper and lower tail probabilities of bivariate distributions

Shogo Kato (Institute of Statistical Mathematics, Japan) ****SPEAKER****

Toshinao Yoshiba (Bank of Japan, Japan)

Shinto Eguchi (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 talk 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. An estimator of the proposed measure is given and its asymptotic normality is shown. A nonparametric test based on the estimator is presented to test the symmetry in upper and lower tails. As an illustrative example, the presented measure is applied to stock daily returns of Nikkei225 and S&P500.

Robust estimation of causal effects under data combination using instrumental variables

Baoluo Sun

National University of Singapore, Singapore

Abstract

Although instrumental variable (IV) methods are widely used to estimate causal effects in the presence of unmeasured confounding, the IVs, exposure and outcome are often not measured in the same sample due to complex data harvesting procedures. Under these settings, two-sample IV methods have been widely used in the health and social sciences. We propose new multiply robust locally efficient estimators of the causal effect of exposure on the outcome, and illustrate the methods through simulation and an econometric application on public housing projects.

Transferred Regression: some theoretical considerations

Jinfang Wang

Yokohama City University, Japan

Abstract:

Transferred regression is a novel regression method, proposed by the author and his collaborators recently, which utilizes categorized information in table forms as prior information for a second stage Bayesian analyses. This approach uses parametric models to predict the cell probabilities, and estimate the regression parameters by minimizing the Kullback-Leibler divergence from the data to the predictive distribution. In this talk we will concentrate on theoretical properties such as efficiencies of the proposed methods. We will also illustrate the usefulness of transferred regressions using comprehensive medical check up data for diabetes prediction.

The performance of penalized logistic regression with rare events data

Yunchan Chi

National Cheng-Kung University, Tainan, Taiwan

ycchi@mail.ncku.edu.tw

Grace S. Shieh

Institute of Statistical Science

Academia Sinica, Taipei, Taiwan

gshieh@stat.sinica.edu.tw

Ding-Huang Lin

National Cheng-Kung University, Tainan, Taiwan

ken476397321@gmail.com

Abstract

It is well known that the maximum of likelihood estimator (MLE) of regression parameters of logistic regression is seriously biased estimators when event rate is very small, especially under small sample size. Consequently, King and Zeng (2001) proposed a modified estimating method to reduce the bias of the MLE of regression parameters of logistic regression. Recently, most researchers use Firth's (1993) penalized regression to reduce the bias of the MLE of regression parameters of logistic regression.

Based on a simulation study, this paper examines the performance of variable selection methods in selecting important variables in logistic regression, including smoothly clipped absolute deviation (SCAD), least absolute shrinkage and selection operator (LASSO), and adapted LASSO. The simulation results will be presented and the application to prediction of cancer drug response will be illustrated.

Does Gene-Gene interaction play a role in genetic predisposition to diseases?

Nelson Tang, Chinese University of Hong Kong

During the past decade, GWAS dominated the field of human genetics and have been very successful in identifying disease causing /predisposition loci of many common diseases and human traits. All these analyses were performed at a single SNP level by performing chi square for a single factor SNP in a 2 x 3 table. On the other hand, in model organisms and animals, interaction between genes or between mutations are commonly reported. So the first question in mind 10 years ago was if SNP-SNP interaction also occurs in common diseases in human. Given that 500,000 SNPs were genotyped on average, the testing space became $500,000^2$ when interaction was analysed. At that time, algorithms to analyse SNP-SNP interaction were few or it took months to complete one chromosome.

We developed a log-linear model with approximate solution to provide a rapid screening of potential interacting SNPs. It was called BOOST. The results could be generated but sometime, the genetic mechanisms would not be apparent. Subsequently, another test constrained on either dominant-dominant or recessive-recessive interactions were developed. Simulation analysis showed that statistical power of common GWAS samples sizes ($N=2000$) should be sufficient to pick up meaning interactions. This algorithm was applied to 2 Psoriasis datasets and revealed complex interactions among genes in the vicinity of the HLA locus. This may be a realistic impression of real life interactions that is one acting in cis. Wet laboratory experiment also revealed interaction working in cis between microsatellites and SNPs in the IGF1 gene. This long term quest for gene interaction provided us some insight into the interaction pattern that was overlooked before.

Integration analysis of multi-omics data using hierarchical structural component model and its application to pancreatic ductal adenocarcinoma

Yongkang Kim and Taesung Park

Department of Statistics
Seoul National University, Seoul, Korea

Abstract

In personalized medicine era, different omics data are commonly generated from each patient. Identification of multi-markers from these multi-omics data is one of the most challenging issues. Although many methods have been developed to identify candidate markers for each type of omics data, few methods can facilitate multi-marker identification from multi-omics data. In this talk, we will introduce a hierarchical structural component model (HisCoM) for integrative analysis of multi-omics data. We apply the HisCoM model to integrate microRNA and mRNA data. It is well known that microRNAs affect phenotypes only indirectly, through regulating mRNA expression and/or protein translation. To put this knowledge into practice, we suggest a microRNA-mRNA integration model for survival time analysis (“mimi-survival”) that accounts for this biological relationship, to efficiently study and identify such integrated markers. In real data analysis, mimi-survival successfully identified more informative microRNA-mRNA integration sets for survival time of pancreatic ductal adenocarcinoma (PDAC) patients, compared to the other existing methods. Through this application to pancreatic cancer data, our proposed model was shown to effectively identify integrated microRNA/target mRNA sets as markers for prognosis, providing a much broader biological interpretation.

醫療理賠、成本分類及估計

吳孟瑾*、俞淑惠

國立高雄大學統計學研究所

摘要

長期以來，人壽保險市場的主力產品為醫療險和意外險。而住院理賠是醫療險理賠金的最大項目。本研究分析歷史住院索賠數據為達到兩個目的—找出住院風險高的客戶群，作為核保時保費調整的參考依據，降低保險公司的承保風險；預測下一年度的住院天數，對下一年度的住院理賠準備金更精準的估計，使公司的資金更靈活運用。針對資料特性及目的我們發展兩種分析方法—根據零膨脹廣義卜瓦松迴歸模型（zero inflated generalized Poisson regression model，簡稱 ZIGP）推廣的多點膨脹截斷廣義卜瓦松迴歸模型（multiple inflated truncated generalized Poisson regression model，簡稱 MITGP）以及混和機器學習及統計方法的三階段模型。實證研究採用 Heritage Provider Network 提供的兩年歷史索賠資料與住院天數資料，並使用四種指標—均方根對數誤差（Root Mean Square Logarithmic Error，簡稱 RMSLE）、平均住院天數、預測準確率、住院總花費與其他常見的分析方法進行比較。實證研究顯示，MITGP 能準確預測下一年度的平均住院天數；三階段模型 RMSLE 較小。因此，使用 MITGP 對於保險公司估計下一年度的住院理賠準備金較為精準；三階段模型對於找出住院風險高的客戶群較有幫助。

關鍵詞：ZIGP、MITGP、RMSLE。

Financial Indices Prediction by Integrating Sentiment Analysis with Factors of International Stock Markets

Yun-Hsuan Chan (詹韻玄)* and Shuen-Lin Jeng (鄭順林)

National Cheng Kung University

Abstract

The trend prediction of the stock markets is one of the major challenges in financial time series prediction. Recently, the influence of social networks on stock market predictions has been brought to attention. Constructing critical news features is one way of sentiment analysis. Different from most of the research using forums as their sources of sentiment analysis, this study chooses the multiple news sources with lower noise as news features. Apart from the impact of the news features, the stock prices are affected by many other factors such as technical indicators and economic factors. Therefore, we apply both technical and news features to predict the trend of financial indices and individual stocks. Simultaneously, we combine the factors of the international stock markets relationships between the United States, China, Hong Kong and Taiwan. We aim at looking through various methods to construct the critical news features which are from the sources of multiple regions. Three statistical models are used to fit the technical and news features to predict the market trends. They are vector autoregression (VAR), vector error correction model (VECM) and multivariate adaptive regression splines (MARS). The results show that models with news features increase the accuracy of trend prediction. Finally, the relationships of financial indices and individual stocks in multiple regions are summarized.

Keywords: International Stock Markets, Technical Analysis, Sentiment Analysis, Prediction, Classification

Optimal Momentum Strategies under Trend Transitions

Yu-Han Shiau (蕭宇瀚)

National University of Kaohsiung

Abstract

Momentum investment strategies have been widely discussed. In the literature, a specified market may follow momentum or contrarian momentum during a specified time period. Therefore, considering momentum with possible transitions, investors can profit from both markets. In this study, we will develop the optimal momentum investment strategies under possible trend transitions. To this purpose, we introduce a modified Implied Price Risk index (IPR index) of Chung and Ho (2014) to determine the status of asset's momentum and estimate the transition matrix. From these results, we could predict the future status of assets to decide the underlying assets. Finally, we execute some case studies. In our case studies, we apply the proposed methods to US stock market from 1999 to 2016. The portfolio performances will be compared to those of Jegadeesh and Titman (1993) and Chuang and Ho (2014). The results show that short-term investments considering transition probability in contrarian momentum market would do better performance. With optimal asset allocations also help us reduce investment risks.

關鍵詞：

Momentum strategy, contrarian strategy, implied price risk index, transition matrix, optimal asset allocation.

Portfolio Selection via Network Approach

Mu-En Hsieh and Meihui Guo
Department of Applied Mathematics
National Sun Yat-sen University

Abstract:

How to choose stocks from the market to form a portfolio is an important issue for investment. According to modern portfolio theory one can either assemble a portfolio of assets such that the expected return is maximized for a given level of risk or the risk is minimized for a given expected return. This article focuses on constructing financial networks for portfolio selection. We adopt the planar maximumly filtered graph method to construct the financial network of stock return time series. We use an eigenvector centrality measure, defined in Zhao et al. (2017), to construct mean-variance portfolios. Empirical studies of several stock markets and different time periods show that those portfolios composed of peripheral stocks with low temporal centrality scores, perform consistently better than portfolios composed of central stocks. The results suggest that the temporal centrality measure provides a useful portfolio optimization and risk management tool.

Keywords and phrases : centrality, financial network, planar maximumly filtered graph, portfolio, time series.

Prediction of Potential Customers with Data Streams

謝宜臻
Yi-Jen Shie

國立中山大學應用數學系
Department of Applied Mathematics
National Sun Yat-sen University

Abstract

Prediction of potential customers is an important issue for e-commerce companies. In this study, we analyze this problem for two data sets. The first dataset, based on the Kaggle competition called "Acquire Value Shopper Challenge", contains all the given transactions with size over 22GB. The objective is to make prediction for customers who remain loyal after a promotional period. The second dataset is from a local e-commerce company, which contains customer purchase history and product information. The objective is to make prediction for potential customers who will buy new product given their past purchase information. First we transfer the data stream to a new data format where columns represent aggregated features and rows represent customers. The aspect of feature engineering was based on the aggregation of customer transactions in the past. We consider the prediction models including XGBoost classifier and support vector machine. To evaluate the models, the area under the operating characteristic curve (AUC) and prediction accuracies are adopted. The results provide useful information for e-companies to keep their advertising budget in line with promotional and marketing goals.

Key words : e-commerce, data streams, aggregated features, XGBoost classification, support vector machine

On Sovereign Credit Ratings Classification and Evaluation

Shu-Hao Chin(金書豪)*, Shu-Hui Yu(俞淑惠)

Institute of Statistics, National University of Kaohsiung

Abstract

In financial study, credit ratings are a critical issue. Credit ratings represent the debt-paying ability of a debtor. Since a single credit ratings result from the credit ratings agencies may not suitable for all financial purposes, it is essential to make a specific ratings/classification result for a specific financial purpose. In this study, we take sovereign credit ratings for an example. In order to develop the new ratings/classification, we consider risk management of exchange rates as our financial purpose, and then use the VaR estimates as the evaluation criterion for the ratings/classification performance. However, since the forward value of exchange rates are unobservable, we proposed some modified CreditMetricsTM methods, including empirical method and time-series-based method, to solve the problem. For case studies, we consider 46 countries from 1997 to 2017. Therefore, we also develop a method to apply the time series analysis on the panel data type. The results indicate that the proposed method has satisfactory performance.

Keywords: Credit ratings; Exchanges rates; Value at Risk (VaR); CreditMetricsTM; Panel data