Session I

105 年 6 月 24 (星期五，13:30∼15:10)

I-1 (SS1001) **Invited Session : Model Selection for Dependent Data**

I-2 (SS1003) **Invited Session : Bayesian Statistics**

I-3 (SS1004) **Invited Session : High-Dimensional Analysis**

I-4 (SS1005) **Invited Session : Statistical Computation**

I-5 (SS1006) **Invited Session : Applied Statistics**

I-7 (SS2003-2) **Invited Session : Internetional session (I)**

I-8 (SS2005) **Invited Session : Educational Statistics**

I-9 (SS2006) **Invited Session : Industrial Statistics**

I-10 (SS2012) **Invited Session : Stochastic Processes**

Graphical Regression

Hsin-Cheng Huang

Institute of Statistical Science, Academia Sinica

Abstract

Graphical models have proven useful in describing relations among interacting units. Motivated from neuroimaging analysis, we propose graphical regression to link the inverse of covariance matrix to covariates, to locate the structural change of a graph as a function of covariates. Statistically, we construct a penalized maximum likelihood for regression analysis, where sparsity constraints are imposed to estimate the graph structure through covariates. Computationally, a fast algorithm is developed based on an alternating direction method of multipliers. Some numerical and theoretical properties will also be provided.

Keywords: network, non-convex minimization, penalized maximum likelihood, precise matrix, undirected graphical model.

Model Selection for High-dimensional Time Series

Ching-Kang Ing

Institute of Statistical Science, Academia Sinica

Abstract

In the past decade, model selection for high-dimensional regression models is one of the most vibrant research topics in statistics. However, most of the attention has been devoted to situations where observations are independent, and hence time series data are precluded. In this talk, I shall address model selection problems for some high-dimensional time series models, including high-dimensional stochastic regression models and high-dimensional regression models with correlated errors. I will present rates of convergence of the orthogonal greedy algorithm (OGA) under various sparsity conditions. I will also show that when the high-dimensional information criterion (HDIC) of Ing and Lai (2011) is used in conjunction with the OGA, the resultant predictor achieves the optimal error rate. Rates of convergence of the OGA are further established under model misspecification. Applications of this latter result to model selection for high-dimensional interaction models will also be given.

Model Selection or Model Averaging for Spatial Regression

Chun-Shu Chen*

Institute of Statistics and Information Science, National Changhua University of Education

Jun Zhu

Department of Statistics, University of Wisconsin, Madison, Wisconsin, USA

Tingjin Chu

Institute of Statistics and Big Data, Renmin University of China, Beijing, China

Abstract

Model selection and model averaging are essential to regression analysis in environmental studies, but determining which of the two approaches is the more appropriate and under what circumstances remains an active research topic. In this paper, we focus on spatial regression models for spatially referenced environmental data. For a general information criterion, we develop a new perturbation-based criterion that measures the uncertainty (or, instability) of spatial model selection, as well as an empirical rule for choosing between model selection and model averaging. Statistical inference based on the proposed model selection instability measure is justified both in theory and via a simulation study. The results suggest that the performance of model selection and model averaging can be quite different for smaller models but are more comparable when the model is relatively large. For illustration, a precipitation data set in the state of Colorado is analyzed.

Keywords: data perturbation, geostatistics, information criterion, model complexity, spatial prediction

Real-time Bayesian Inference for Latent Ability Models

Ruby C. Weng

National Chengchi University

Abstract

Latent ability models relate a set of observed variables to a set of latent ability variables. It has a variety of applications, including the item response theory (IRT) models and the paired comparison models. The IRT models have been widely used in modeling educational test data, and the paired comparison models such as Bradley-Terry-Luce models are popular in modeling competition data. In this talk, I first review a Bayesian approximate method for online gaming analysis with paired comparison models, and then present an application to Internet ratings data using item response theory models.

Designing Bayesian Sampling Plans with Interval Censoring

Ming-Chung Yang*

Department of Banking and Finance, Kainan University

Lee-Shen Chen

Department of Applied Statistics and Information Science, Ming Chuan University

Abstract

This talk will present the problem of designing Bayesian sampling plans with interval censored samples. First, we derive conventional Bayesian sampling plan (BSP), and its monotonicity is studied. Based on the BSP, a new sampling plan modified by the curtailment procedure using the monotonicity is proposed. The resulting sampling plan is called the curtailed Bayesian sampling plan (CBSP), which can reduce the duration time of life test experiment. When the cost of the duration time of life test experiment is considered, we show that the Bayes risk of the CBSP is smaller than the Bayes risk of the BSP. A Monte Carlo simulation study is given to illustrate the performance of the CBSP compared with the BSP.

Keywords: curtailed Bayesian variable plan, curtailed decision function, interval censoring.

Bayesian Variable Selection for Cox Proportional Hazard Regressions with Right Censoring

Sheng-Mao Chang* and Pei-Fang Su

Department of Statistics, National Cheng Kung University

Jung-Ying Tzeng

Department of Statistics & Bioinformatics Research Center,
North Carolina State University, Raleigh, NC, U.S.A.

Abstract

In high-throughput biological studies, it is common to see datasets with much more covariates than independent subjects. The class of two-stage approaches, screening and then selection, e.g. Fan and Lv (2008), results in desired performance in the context of variable selection. In this work, a Bayesian variable selection approach for lifetime data with right censoring is considered to mimic the class of two-stage approaches. The proposed Bayesian model consists of the Cox proportional hazard likelihood, generalized g-prior on regression coefficients, and Gamma process prior on cumulative hazard. This model results in a closed-form screening statistic. Next, we apply stochastic search variable selection method, attaching selection indicators to the regression coefficients, to select important variable, and incorporate the screening screening statistic into the prior of selection indicators. Simulations and data analysis for B-cell lymphoma data are given to demonstrate the application of the methods.

Keywords: Cox proportional hazard model, generalized g-prior, stochastic search variable selection.

Integrating Multiple Random Sketches of Singular Value Decomposition

Su-Yun Huang

Institute of Statistical Science, Academia Sinica

Abstract

Low-rank singular value decomposition (SVD) of large-scale matrices is one key tool in modern data analysis and scientific computing. Rapid growing in matrix size further increases the needs and poses the challenges for developing efficient large-scale SVD algorithms. Random sketching is a promising method to reduce the problem size for computing an approximate SVD. We generalize the one-time sketching to multiple random sketches and develop algorithms to integrate these random sketches containing various subspace information in different randomizations. Such integration procedure can lead to SVD with higher accuracy and the multiple randomizations can be conducted on parallel computers simultaneously. We also reveal the insights and analyze the performance of the proposed algorithms from statistical and geometric viewpoints. Numerical results are presented and discussed to demonstrate the efficiency of the proposed algorithms. (based on joint work with D. Chang, H. Chen, T.L. Chen, C.Y. Lin and W. Wang)

Keywords: dimension reduction, high-dimensional data analysis, randomized algorithms, singular value decomposition.

An Adaptive Clustering for Functional Data

Heng-Hui Lue

Department of Statistics, Tunghai University

Abstract

We propose a new adaptive approach for clustering functional data. The data-adaptive searching method based on dimension reduction theory is proposed for estimating the basis functions and the sufficient dimension reduction space of predictors. First, these estimates are obtained through local linear approximation techniques without requiring a prespecified parametric model. A K-means clustering method is then adopted for functional clustering analysis. The two steps are iteratively proceeded until the convergence is attained. Several simulations are reported for illustration.

Keywords: clustering, dimension reduction, functional data analysis, local linear smoother, sliced inverse regression.

Sensible Functional Classification

Ci-Ren Jiang* (江其衽)

Institute of Statistical Science, Academia Sinica

Lu-Hung Chen (陳律閎)

National Chung Hsing University

Abstract

The focus of this paper is to extend Fisher's linear discriminant analysis (LDA) to both densely recorded functional data and sparsely observed longitudinal data for general c-category classification problems. We propose an efficient approach to identify the optimal LDA projections in addition to managing the nonivertibility issue of the covariance operator emerging from this extension. A conditional expectation technique is employed to tackle the challenge of projecting sparse data to the LDA directions. We study the asymptotic properties of the proposed estimators and show that the asymptotically perfect classification can be achieved in certain circumstances. The performance of this new approach is further demonstrated with numerical examples.

Keywords: classification, functional data, linear discriminant analysis, smoothing.

EL2Surv: Empirical Likelihood for Comparing Two Survival Functions

Hsin-Wen Chang*, Cheng-Chen Tsai, Pei-Yuan Tsai and Ryan Kao

Institute of Statistical Science, Academia Sinica

Abstract

Comparing two survival functions is a fundamental problem in survial analysis. Despite the classical procedures like the log-rank test or confidence bands formed by the difference between two Kaplan–Meier estimates of survival functions, there have been increasing evidence that empirical likelihood (EL) can provide more accurate confidence bands and powerful hypothesis tests. Packages `emplik` and `emplik2` have been developed for survival data analysis using EL, but only for inference regarding finite-dimensional parameters like regression coefficient or vector-valued mean. For infinite dimensional problem like simultaneous inference of two survival functions, we develop `R` functions for empirical likelihood based confidence bands and hypothesis tests. Both two-sided and one-sided tests are provided. The software will be made freely available in the R package `EL2Surv`. Simulation results demonstrating the superiority of these procedures over traditional survival methods are presented. The use of package `EL2Surv` is also illustrated through different applications.

Keywords : nonparametric likelihood ratio, `R`, right-censored data, survival analysis, two-sample problem.

開放資料之隱私去識別評估暨視覺化技術

高君豪*、謝志宏、朱宇豐、鄺郁婷

資策會資安所

摘要

近年來，由於巨量資料分析技術之興起，各領域相關應用的價值連帶提升，政府公部門及私人企業也逐漸重視異質性或跨域性資料間的關聯分析。然而各國 Open Data 政策在實際運作上之成效仍有許多改進空間。歸究其主要原因之一乃是對於「個資保護」的隱憂與不安。基於實際情境的研究已指出，即便是已針對個人機敏資訊進行去識別化處理的單一資料表，依然有機會在開放資料環境中經由異質性或跨域性資料表間的關聯分析重新識別出個人機敏資訊。有鑑於此，考量在巨量規模開放資料的情境下，使用者必須面對：1) 更多的開放資料集、2) 更大的資料規模、以及 3) 更加複雜的資料關聯後資訊外洩問題，個人機敏資料的去識別化也將成爲更困難的挑戰。基於上述困難與挑戰，本研究擬結合：1) 可規模化資料存儲系統之施作、2) 可規模化架構下各式資料去識別化模型之鑑測、及 3) 資料視覺化分析等相關技術，設計「可規模化開放資料之隱私去識別化鑑測暨視覺化技術」預計將提供國內第一套「可規模化 (Dynamic Scalable) 開放式資料隱私鑑測視覺化分析工具」，對於異質性或跨域性資料關聯之個資保護可靠度，進行 1) 高效率地評估、2)有效個人機敏資訊去識別化鑑測及簡易修正、和 3) 提供簡潔清淅、一目瞭然的使用者體驗與視覺化呈獻。如此可讓資料擁有者可以更願意、更放心的來公開手上所蒐集之資料給社會公眾，進一步加速政府開放資料政策之實施成效，並帶動開放資料相關產值效益的提升。

關鍵詞：開放資料、個人機敏資訊保護、資料去識別化、資料視覺化、巨量資料

Regularized Spatial Maximum Covariance Analysis

Wen-Ting Wang* (王文廷)

Institute of Statistics, National Chiao Tung University

Hsin-Cheng Huang (黃信誠)

Institute of Statistical Science, Academia Sinica

Abstract

In climate and atmospheric research, many phenomena involve more than one meteorological spatial processes covarying in space. To understand how one process is affected by another, maximum covariance analysis (MCA) is commonly applied. However, the patterns obtained from MCA may sometimes be difficult to interpret. In this paper, we propose a regularization approach to promote spatial features in dominant coupled patterns by introducing smoothness and sparseness penalties while accounting for their orthogonalities. We develop an efficient algorithm to solve the resulting optimization problem by using the alternating direction method of multipliers. The effectiveness of the proposed method is illustrated by several numerical examples, including an application to study how precipitations in east Africa are affected by sea surface temperatures in the Indian Ocean.

Keywords : singular value decomposition, Lasso, smoothing splines, orthogonal constraint, alternating direction method of multipliers.

Influence Analysis on Linear Regression for Symbolic Interval Data

Chun-Yi Tsai and Yufen Huang*

Department of Mathematics, National Chung Cheng University

Abstract

In the last decades, with the advent of computers, data sets become inevitably large than before. This brings the difficulty in performing standard statistical analysis. Therefore, such huge data sets must be aggregated in some fashion and the resulting summary data may be represented by lists, intervals, histograms and the like, which are called symbolic data. Linear regression is one of the most important and useful tools to analyze the data. During the fitting process, observations that are suspicious can greatly influence the results of the analysis. Therefore, detection of such influential points becomes an essential task. Many literatures have studied for the influence analysis in linear regression for classical data. However, to our knowledge, a study in the influence analysis on regression for symbolic data has not been explored in the literature. In this paper, we develop three sample versions of the influence function motivated by Hampel (1974) to identify suspicious concepts which cause seriously adverse effects on the linear regression analysis results for symbolic interval data. Also we illustrate these proposed methods with simulation studies and real data examples.

Keywords: linear regression, symbolic data, symbolic interval data.

Motion-blurred Image Restoration

Yu-Jhong Wu (吳諭忠) and Huey-Miin Hsueh* (薛慧敏)

Department of Statistics, National Chengchi University

Abstract

Camera shake for object movement always lead to occurrence of motion blur in an image. A conventional model for the blurred intensity is expressed by the convolution of the original intensity and a point spread function. A motion-blurred image can be successfully restored if an adequate estimation of the point spread function is obtained. In this study, we consider a linear point spread function, in which there involve two parameters: angle and length. We develop an estimating procedure for the two parameters based on the power spectrum of the blurred image. Through intensive experiments, the proposed method is found to produce more accurate angle detection and more stable length estimation than existing methods.

Keywords: image restoration, motion blur, point spread function, power spectrum, Radon transform.

Semiparametric Regression Analysis of Recurrent Gap Times in the Presence of Competing Risks

Chia-Hui Huang* (黃佳慧)

Department of Statistics, National Taipei University

Yi-Hau Chen (程毅豪)

Institute of Statistical Science, Academia Sinica

Ya-Wen Chuang (莊雅雯)

Division of Nephrology, Taichung Veterans General Hospital

Abstract

When a disease progression is assumed to go through several stages marked by a nonterminal, recurrent event such as relapse, or a terminal event such as death, whose occurrence terminates the progression, researchers might be concerned with the duration or gap times between successive events (stages) and would like to study the covariates effects on the gap times. In addition, how the previous events or gap times affect the current gap time may be also of interest. We propose a unifying framework for joint regression analysis of gap times between successive events. The proposed mixture modeling framework consists of a logistic regression for predicting the path of transition (to a nonterminal or terminal event) at each stage, and proportional hazards models for predicting the gap times for transition to the nonterminal and terminal events at each stage, and both the two components of models are conditional on the past event history and stage-specific covariates. As special cases, when the number of stages is fixed at one or two, the proposed framework can be applied to analysis of conventional competing risks or semicompeting risks data. We develop semiparametric maximum likelihood inference procedure for the proposed models, where the score functions are explicitly expressed as martingales, and hence the large sample theory follows directly from martingale theory. Explicit expressions for the information matrix are also derived, which facilitate direct variance estimation and convenient computation. Simulation results reveal the nice performance, and applications to two clinical studies illustrate the real utilities of the proposed model.

Keywords: competing risks, martingale processes, mixture model, multiple events, recurrent data.

Asymptotic Inference for Common Factor Models in the Presence of Jumps

Yohei Yamamoto

Department of Economics, Hitotsubashi University

Abstract

Financial and macroeconomic time-series data often exhibit infrequent but large jumps. Such jumps may be considered as outliers that are independent of the underlying data-generating processes and contaminate inferences on their model. In this study, we investigate the effects of such jumps on asymptotic inference for large-dimensional common factor models. We first derive the upper bound of jump magnitudes with which the standard asymptotic inference goes through. Second, we propose a jump-correction method based on a series-by-series outlier detection algorithm without accounting for the factor structure. This method gains standard asymptotic normality for the factor model unless outliers occur at common dates. Finally, we propose a test to investigate whether the jumps at a common date are independent outliers or are of factors. A Monte Carlo experiment confirms that the proposed jump-correction method retrieves good finite sample properties. The proposed test shows good size and power. Two small empirical applications illustrate usefulness of the proposed methods.

Keywords: outliers, large-dimensional common factor models, principal components, jumps.

Linear Double Autoregressive Model and Its Conditional Quantile Inference

Qianqian Zhu, Yao Zheng and Guodong Li*

Department of Statistics and Actuarial Science, University of Hong Kong

Abstract

This paper proposes the linear double autoregressive (LDAR) model by defining a linear structure on the conditional standard deviation of the double autoregressive (AR) model, which makes it convenient to apply conditional quantile estimation. The proposed model has larger parameter space than that of the commonly used AR model and only needs the existence of fractional or absolute moment for the innovation term. A necessary and sufficient condition for the strictly stationary and ergodic solution to LDAR model is derived under the case that fractional moment exist. Doubly weighted quantile estimation and a diagnostic tool are studied for the LDAR model. Simulation study and a real data analysis illustrate the performance of the doubly weighted quantile estimator and the diagnostic checking procedure.

Buffered Autoregressive Models with Conditional Heteroscedasticity: An Application to Exchange Rates

Philip L.H. Yu* (楊良河) and Wai Keung Li (李偉強)

Department of Statistics and Actuarial Science, The University of Hong Kong

Ke Zhu (朱柯)

Institute of Applied Mathematics, Chinese Academy of Sciences

Abstract

This paper introduces a new model called the buffered autoregressive model with generalized autoregressive conditional heteroscedasticity (BAR-GARCH). The proposed model, as an extension of the BAR model in Li et al. (2015), can capture the buffering phenomena of time series in both the conditional mean and variance. Thus, it provides us a new way to study the non-linearity of time series. Compared with the existing AR-GARCH and threshold AR-GARCH models, an application to several exchange rates highlights the importance of the BAR-GARCH model.

Keywords: buffered AR-GARCH model, exchange rate, nonlinear time series.

Cognitive Diagnostic Models for Classifying Skills and Misconceptions Simultaneously

Bor-Chen Kuo (郭伯臣)

Graduate Institute of Educational Information and Measurement, National Taichung University of Education

Abstract

In Psychometrics or educational measurement research, cognitive diagnostic models (CDMs) are probability models that can identify the presence and absence of students' latent attributes. Many CDMs have been proposed to model skills and concepts which are used as the evidences for constructing adaptive remedial instruction. In addition to skills and concepts, learning bugs or misconceptions are also useful information in the tutoring model of ITS. In this presentation, a novel cognitive diagnostic model for identifying skills and misconceptions simultaneously based on students' task responses will be introduced.

A Unified Penalized Likelihood Method for Structural Equation Modeling

Po-Hsien Huang

National Cheng Kung University

Abstract

Structural equation modeling (SEM) is a commonly used multivariate statistical method in psychological studies. Under SEM framework, researchers can flexibly specify their models based on available psychological theories and test the plausibility of the hypothesized models. In this dissertation, a penalized likelihood (PL) method for SEM was proposed. Compared to the usual likelihood, PL includes an additional penalty term to control the complexity of the hypothesized model. When the penalty level is chosen appropriately, PL can yield a model that balances model goodness-of-fit and model complexity. The proposed method is especially useful when limited substantive knowledge is available for model specification. An expectation-conditional maximization (ECM) algorithm was developed to maximize the PL estimation criterion with several state-of-art penalty functions. Four theorems on the asymptotic behaviors of PL were derived, including the local and global oracle property of PL estimators and the selection consistency of Akaike and Bayesian information criterion (AIC/BIC). Two simulations were conducted to evaluate the empirical performance of the proposed PL method. The practical utility of PL was demonstrated through real data examples.

Keywords: structural equation modeling, penalized likelihood, model selection, factor analysis model, MIMIC model.

Investigation of Constraint-weighted Item Selection Procedures in Computerized Adaptive Testing

Ya-Hui Su

Department of Psychology, National Chung Cheng University

Abstract

Computerized adaptive testing (CAT) has been widely used in educational and psychological assessments because it can obtain efficient and precise ability estimation with fewer items than traditional paper and pencil tests. One of the important issues in CAT is the item selection algorithm. Test specifications specify a series of constraints for including items in a test (Swanson & Stocking, 1993). The construction of assessments in CAT usually involves fulfilling a large number of statistical (e.g. target item and test information) and non-statistical (e.g. content specifications and key balancing) constraints to meet the test specifications. Although different algorithms perform item selection sequentially or simultaneously in the test assembly, the item selection in CAT is sequential by nature (van der Linden, 2005). Therefore, it is challenging, while constructing assessments, to meet the various constraints in CAT simultaneously. To improve measurement precision and test validity, the priority index (PI) and multidimensional priority index (MPI) can be used to monitor many constraints simultaneously in unidimensional CAT and multidimensional CAT (MCAT), respectively.

Since the priority index approaches can be implemented easily and computed efficiently, they are important and useful for operational CATs. This talk will first review the development of the priority index. Some studies and findings will also be included in my talk.

Keywords: CAT, priority index, multidimensional, item selection, constrained, IRT.

Profile Monitoring, Fault Detection, and Diagnosis

Jyh-Jen Horng Shiau

Institute of Statistics, National Chiao Tung University

Abstract

In many practical situations, the quality of a process or product is characterized by a relationship (or profile) between a response variable and one or more independent variables instead of by the distribution of a univariate or multivariate quality characteristic. Most research work in the literature for profile SPC assumed fixed effects and/or parametric regression models to model profiles. In this study, considering a more general and practical situation that profiles are of more flexible shapes and with the subject effect, we propose schemes for profile monitoring as well as fault detection and diagnosis (FDD). Profile monitoring schemes for Gaussian processes and non-Gaussion processes are presented, respectively. For FDD, assuming sample profiles from the in-control process and certain frequently-occurred out-of-control conditions (faults) are available, a nonprametric FDD procedure is provided for practitioners to use. When a profile signals out-of-control, the procedure will determine which of the known faults contributes to it or identify it as coming from a novel fault.

A High-dimensional Location-dispersion Model with Applications to Root Cause Detection for Wafer Fabrication Processes

Ching-Kang Ing

Institute of Statistical Science, Academia Sinica
National Sun Yat-sen University
National University of Kaohsiung

Mei-Hui Guo

National Sun Yat-sen University

Shu-Hui Yu, Shih-Feng Huang* and Hsiang-Ling Hsu

National University of Kaohsiung

Abstract

We consider the problem of selecting a high-dimensional location-dispersion model. The orthogonal greedy algorithm (OGA) in conjunction with the high-dimensional information criterion (HDIC) and TRIM used by Ing and Lai (2011) in high-dimensional homogeneity models is generalized to accommodate high-dimensional dispersion components. We prove selection consistency and derive the limiting distributions of the estimated parameters. These results are then applied to root cause detection for wafer fabrication processes, in which a problematic tool can lead to not only a location shift, but also an increase in variance. In particular, based on the variables selected by OGA+HDIC+TRIM, a novel and easy to implement procedure is proposed to identify problematic tools. Moreover, since parameter estimation involves solving nonlinear optimization problems with many variables, we provide two iterative algorithms to alleviate this difficulty. Real data analysis shows that the proposed method performs quite satisfactorily.

Keywords: high-dimensional model selection, location-dispersion model, orthogonal greedy algorithm.

MEWMA Control Chart and Process Capability Indices for Simple Linear Profiles with Within-profile Autocorrelation

Jyun-You Chiang

School of Statistics, Southwestern University of Finance and Economics, China

Y.L. Lio

Department of Mathematical Science, University of South Dakota, USA

Tzong-Ru Tsai*

Department of Statistics, Tamkang University

Abstract

A multivariate exponentially weighted moving average (MEWMA) control chart is proposed in this paper to detect process shifts during the Phase-II monitoring of simple linear profiles in the presence of within-profile autocorrelation. Furthermore, two process capability indices are proposed for evaluating the capability of in-control simple linear profile processes. Simulations are conducted to evaluate the performance of the proposed MEWMA control chart, and the use of two process capability indices is demonstrated through examples.

Keywords: autocorrelation, average run length, process capability analysis, profile monitoring

Limit Theorems for Regenerative Sequences of Random Variables

Krishna B. Athreya

Departments of Mathematics and Statistics, Iowa State University

Abstract

There are many sequences of random variables whose evolution can be broken into iid cycles. We give a number of examples of these. For these we prove ergodic theorems. We also show the uniform convergence of the empirical distribution functions in a number of cases. We also prove for some expectations variation norm convergence.

Systemic Risk and Interbank Lending

Li-Hsien Sun (孫立憲)

Institute of Statistics, National Central University

Abstract

We propose a simple model of inter-bank lending and borrowing incorporating a game feature where the evolution of monetary reserve is described by a system of coupled Feller diffusions. The optimization subject to the quadratic cost not only reflects the desire of each bank to borrow from or lend to a central bank through manipulating its lending preference but also to intend to leave deposits in a central bank in order to control the volatility for cost minimization. We observe that the adding liquidity creates the effect of flocking leading to stability or systemic risk according to the level of the growth rate. The deposit rate brings about a large number of bank defaults by diminishing the growth of the system. A central bank acts as a central deposit corporation. In addition, the corresponding Mean Field Game in the case of the number of banks N large and the stochastic game on the infinite horizon with the discount factor are also discussed. Finally, we solve for the closed-loop equilibria in the case of inter-bank lending and borrowing with clearing debt obligations using the stochastic game with delay.

Keywords: Feller diffusion, systemic risk, inter-bank borrowing and lending system, Nash equilibrium, mean field game, stochastic game with delay.

Martin Boundary for Gaussian Diffusion Processes

Wei-Da Chen (陳韋達)

國立中央大學數學研究所

Abstract

In this note, we study the Martin boundary for some Gaussian diffusion processes X_t. General theory of Martin boundary for Markov process has been well developed in the literature. See Kunita-Watanabe (1965), Dynkin (1969), Salminen (1981). One of its important applications is to give the unique representation of harmonic functions in terms of minimal Martin functions. There are very few concrete examples that properties of Martin boundary are studied, such as Brownian motion. Martin boundary for $2D$ Gaussian transient diffusion process were studied in Cranston-Orey-Rosler (1983). They studied the space time Martin boundary for such diffusion process. As a result, they obtained the representation for positive harmonic functions. The space-time Martin boundary is to consider the Martin boundary for $\widetilde{X}_t = (t, X_t) \in (0, \infty) \times \mathbb{R}^d$ as a Markov process. In this note, we show how to generalize this idea to the high dimensional space. In this case, we first observe that the space-time Martin topology becomes concrete and can be described in terms of Euclidean topology. We will discuss the convergence of some h-diffusion process, the diffusion process under h-transform, where h is a positive space time harmonic function. As a result, "Hitting" probability measure will be calculated for some examples. We also consider the relation between a positive space time harmonic function and limiting distribution of h-diffusion process.

Session II

105 年 6 月 24 (星期五，15:30∼17:10)

II-1 (SS1001) **Invited Session : Big Data Analysis with R**

II-2 (SS1003) **Invited Session : Biostatistics**

II-3 (SS1004) **Contributed Session : Survey Sampling**

II-4 (SS1005) **Invited Session : Statistical and Machine Learning (I)**

II-5 (SS1006) **Invited Session : Experimental Design**

II-7 (SS2003-2) **Invited Session : Internetional session (II)**

II-8 (SS2005) **Invited Session : Time Series**

II-9 (SS2006) **Invited/Contributed Session : Young Researcher Session (I)**

II-10 (SS2012) **Invited Session : Applied Probability**

漫談巨量資料於 R 軟體中的處理與分析

吳漢銘

淡江大學數學學系資料科學與數理統計組

摘要

巨量資料 (Big Data，或稱爲大數據) 的潮流帶動資料科學 (Data Science) 的興起。資料科學是綜合統計學、資訊科學及領域知識的一門新興學科，而 R 語言正是資料科學領域裡，最受歡迎的程式語言之一。R 具有免費，開放源碼及學習資源豐富的優點，而其功能可透過套件增強。本演講將介紹在 R 軟體中，巨量資料處理、分析與應用之實務策略，讓具統計領域的 R 愛好者有機會跨足實作巨量資料之處理與分析。

關鍵詞：巨量資料、R 軟體、資料探勘、Hadoop

R 語言在金融資料分析上之運用

吳牧恩

東吳大學數學系

摘要

R 語言爲近年來資料分析上相當熱門的工具。而在金融資料處理上，R 語言也是相當實用。本演講我們介紹幾種 R 語言在金融資料處理上常用的套件：quantmod 與 quantstrat。quantmod 套件用來抓取金融資料、繪製行情走勢，而 quantstrat 套件則用來與 quantmod 做搭配，進行策略回測與績效分析。本演講我們並展示一些金融資料分析、策略回測、資金管理等研究案例。

關鍵詞：R 語言、quantmod、quantstrat

Multi-dimensional Functional Principal Component Analysis

Lu-Hung Chen*

National Chung Hsing University

Ci-Ren Jiang

Institute of Statistical Science, Academia Sinica

Abstract

Functional principal component analysis is one of the most commonly employed approaches in functional and longitudinal data analysis and we extend it to analyze functional/longitudinal data observed on a general d-dimensional domain. The computational issues emerging in the extension are fully addressed with our proposed solutions. The local linear smoothing technique is employed to perform estimation because of its capabilities of performing large-scale smoothing and of handling data with different sampling schemes (possibly on irregular domain) in addition to its nice theoretical properties. Besides taking the fast Fourier transform strategy in smoothing, the modern GPGPU (general-purpose computing on graphics processing units) architecture is applied to perform parallel computation to save computation time. To resolve the out-of-memory issue due to large-scale data, the random projection procedure is applied in the eigendecomposition step. We show that the proposed estimators can achieve the classical nonparametric rates for longitudinal data and the optimal convergence rates for functional data if the number of observations per sample is of the order $(n/\log n)^{d/4}$. Finally, the performance of our approach is demonstrated with simulation studies and the fine particulate matter (PM 2.5) data measured in Taiwan.

Keywords : fast fourier transform, functional and longitudinal data, GPU-parallelization, local linear smoother, PM 2.5 data, random projection.

Spatial Scan Statistics for Detection of Multiple Clusters with Arbitrary Shapes

Pei-Sheng Lin

Institute of Population Health Sciences, National Health Research Institutes

Abstract

In applying scan statistics for public health research, it would be valuable to develop a detection method for multiple clusters that accommodates spatial correlation and covariate effects in an integrated model. In this paper, we connect the concepts of the likelihood ratio (LR) scan statistic and the quasi-likelihood (QL) scan statistic to provide a series of detection procedures sufficiently flexible to apply to clusters of arbitrary shape. First, we use an independent scan model for detection of clusters and then a variogram tool to examine the existence of spatial correlation and regional variation based on residuals of the independent scan model. When the estimate of regional variation is significantly different from zero, a mixed QL estimating equation is developed to estimate coefficients of geographic clusters and covariates. We use the Benjamini-Hochberg procedure (1995) to find a threshold for p-values to address the multiple testing problem. A quasi-deviance criterion is used to regroup the estimated clusters to find geographic clusters with arbitrary shapes. We conduct simulations to compare the performance of the proposed method with other scan statistics. For illustration, the method is applied to enterovirus data from Taiwan.

Quantile Regression Based on a Weighted Approach under Semi-Competing Risks Data

Jin-Jian Hsieh

Department of Mathematics, National Chung Cheng University

Abstract

In this article, we investigate the quantile regression analysis for semi-competing risks data in which a non-terminal event may be dependently censored by a terminal event. Due to the dependent censoring, the estimation of quantile regression coefficients on the non-terminal event becomes difficult. In order to handle this problem, we assume Archimedean Copula (AC) to specify the dependence of the non-terminal event and the terminal event. Portnoy (2003) considered the quantile regression model under right censoring data. We extend his approach to construct a weight function, and then impose the weight function to estimate the quantile regression parameter for the non-terminal event under semi-competing risks data. We also prove the consistency and asymptotic properties for the proposed estimator. According to the simulation studies, the performance of our proposed method is good. We also apply our suggested approach to analyze a real data.

Keywords: copula model, dependent censoring, quantile regression, semi-competing risks data.

Model Selection for Marginal Regression Analysis of Longitudinal Data with Missing Observations and Covariate Measurement Error

Chung-Wei Shen* (沈仲維)

Department of Mathematics, National Chung Cheng University

Yi-Hau Chen (程毅豪)

Institute of Statistical Science, Academia Sinica

Abstract

Missing observations and covariate measurement error commonly arise in longitudinal data. However, existing methods for model selection in marginal regression analysis of longitudinal data fail to address the potential bias resulting from these issues. To tackle this problem, we propose a new model selection criterion, the Generalized Longitudinal Information Criterion (GLIC), which is based on an approximately unbiased estimator for the expected quadratic error of a considered marginal model accounting for both data missingness and covariate measurement error. The simulation results reveal that the proposed method performs quite well in the presence of missing data and covariate measurement error. On the contrary, the naive procedures without taking care of such complexity in data may perform quite poorly. The proposed method is applied to data from the Taiwan longitudinal study on aging to assess the relationship of depression with health and social status in the elderly, accommodating measurement error in the covariate as well as missing observations.

Keywords : errors-in-variables, generalized estimating equations, generalized method of moments, information criterion, missing at random

台灣地區之細懸浮微粒與急性支氣管炎盛行率及門診醫療支出之研究

楊天輔*、林麗芬

逢甲大學統計學系

摘要

界衛生組織 (WHO) 發佈報告指出，長時間暴露於空氣中的細懸浮微粒 (Particulate Matter 2.5, $PM_{2.5}$)，會引發心血管疾病、呼吸道疾病、生育不良等問題。本研究將探討台灣急性支氣管炎盛行率與門診醫療支出與細懸浮微粒之關係，主要利用台灣全民健保資料庫，以國際疾病分類病碼 ICD-9 codes：466 急性支氣管炎的患者爲研究對象，並針對 2006 年至 2011 年之就醫紀錄做研究。利用健保資料庫探討季節、區域、性別、年齡及細懸浮微粒濃度對急性支氣管炎盛行率之分佈，研究方法爲對數線性迴歸、羅吉斯迴歸、BURR 迴歸、Negative Binomial 迴歸、卡方檢定分析、交叉分析等，探討不同人口特徵與急性支氣管炎盛行率、就醫總次數以及平均門診就醫金額之差異。結果顯示，暴露在細懸浮微粒濃度越高的環境下急性支氣管炎盛行率亦越高，並且危險組群爲：女性、0 ～ 11 歲以及居住於外島者。在濃度不變的條件下本研究估計 2006 年至 2011 年台灣急性支氣管炎花費約新台幣 $1,913,260$ 萬元，若全台灣濃度降至季平均 $20\mu g/m^3$，則台灣能減少約 $1,928$ 千人就醫 (CI：901 千人，$3,281$ 千人)，並能帶來約新台幣 $98,930$ 萬元的經濟效益，節省約 5.17% 門診醫療支出。

關鍵詞：細懸浮微粒、急性支氣管炎、全民健保資料庫、健康衝擊函數

多階段群集隨機抽樣下參數估計之變異數研究

紀佩妤*、王鴻龍

台北大學統計系

于若蓉

中研院人社調查專題中心

摘要

抽樣調查教科書討論之參數變異數估計值公式有分層抽樣、群集抽樣、二階段群集抽樣等，較少提及多階段抽樣程序之參數變異數估計，本研究以中研院人社中心華人家庭計畫 RI1999 問卷調查資料，去探討多階段抽樣程序之參數變異數估計，此計畫抽樣方式使用多階段分層群集抽樣，將台灣分成九大區域，接著各區域內以鄉鎮分群，採集群隨機抽樣，抽中的鄉鎮再以村里分群，以集群隨機抽中的村里再抽出受訪者，其參數估計的變異數須符合多階段分群集抽樣方式之參數變異數公式。母體採用 1998 內政部户政司鄉鎮市區（或村里）單一年齡人口數。本研究將討論三種群推估母體總數估計式的變異數公式，比較其差異。

關鍵詞：多階段群集抽樣、參數之變異數估計、家庭動態調查

結合交叉結構反覆加權之雙重清冊整合研究

連啓雄*、王鴻龍

國立台北大學統計學系

于若蓉

中研院人社中心調查專題中心

摘要

抽樣調查的樣本結構，常使用多重反覆加權 (Raking)，將樣本結構調整成與母體一致，而一般多重反覆加權只考慮單一邊際變數，如：性別、年齡...等，加權後無法兼顧交叉結構的特性。另外，抽樣調查欲將有兩波 (或以上) 的樣本整合成一樣本，則須考慮結構是否有重疊 (overlapping) 的部分。本研究嘗試將兩樣本整合成一樣本，針對重疊的結構做調整，並結合交叉結構做多重反覆加權，探討方法的合適性及有效性。

本研究將使用中央研究院人文社會科學研究中心華人家庭動態計畫 (PFSD) 所提供 1999、2000 年的家庭動態調查資料，並嘗試使用 Skinner(1991) 整理 Bankier(1986) 的多重清冊方法後，所提出在雙重清冊 (Double Frame) 下針對重疊結構調整的估計式，整合兩年度的樣本，並結合三種交叉結構的多重反覆加權，提出兩種不同方法，比較其差異，找出最適合的方法。

關鍵詞：多重反覆加權、交叉結構、雙重清冊

樣本配置方式在子母體參數估計之精確度的比較分析—以台灣公務調查統計爲例

王晏羚*、許玉雪

國立臺北大學統計學系

摘要

一般而言市場調查結合分層隨機抽樣在進行樣本配置時通常只著重於整個母體參數估計的精確度，未將子母體參數估計的精確度列入考量。欲同時顧及每個子母體參數估計的精確度，樣本該如何進行配置是一重要議題。比較幾種不同的分析方式說明如何進行樣本配置是本文的研究重點。本文旨在探討同時考量每個子母體的參數估計要達到一定的精確度，該如何進行樣本配置。首先彙整幾種樣本配置方法，包含一般較常使用的比例配置、最佳配置以及 Wesolowski and Niemiro (2001) 提出的 eigenproblem approach 三種樣本配置方式，由理論上來比較分析這幾種樣本配置方法在參數估計的精確度。受母體資料取得之研究限制，本研究未能進行模擬分析，而以 102 年低收入户及中低收入户生活狀況調查爲例，進行實證分析，說明以上三種配置方式在此案例之應用。理論分析及實證結果都顯示 eigenproblem approach 將會較比例配置及最佳配置的結果來得好。

關鍵詞：樣本配置、分層抽樣、比例配置、最佳配置、eigenproblem approach

A Prediction Model Integrating the Expression Values of Genes and Long Non-coding RNAs for Radiosensitivity

Tzu-Pin Lu (盧子彬)

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

Abstract

Radiotherapy has become a popular and standard approach for treating cancer patients because it greatly improves patient survival. However, some of the patients receiving radiotherapy suffer from adverse effects and do not obtain survival benefits. This may be attributed to the fact that most radiation treatment plans are designed based on cancer type, without consideration of each individual's radiosensitivity. A model for predicting radiosensitivity would help to address this issue. In this study, the expression levels of both genes and long non-coding RNAs (lncRNAs) were used to build such a prediction model. Analysis of variance and Tukey's honest significant difference tests ($P < 0.001$) were utilized in immortalized B cells ($GSE26835$) to identify differentially expressed genes and lncRNAs after irradiation. A total of 41 genes and lncRNAs associated with radiation exposure were revealed by a network analysis algorithm. To develop a predictive model for radiosensitivity, the expression profiles of $NCI - 60$ cell lines along, with their radiation parameters, were analyzed. A genetic algorithm was proposed to identify 20 predictors, and the support vector machine algorithm was used to evaluate their prediction performance. The model was applied to 2 datasets of glioblastoma, The Cancer Genome Atlas and $GSE16011$, and significantly better survival was observed in patients with greater predicted radiosensitivity.

Keywords: radiosensitivity, long non-coding RNAs, microarray

Class-imbalanced Classification for High-dimensional Data

Wei-Jiun Lin* (林維鈞)

Department of Applied Mathematics, Feng Chia University

James J.Chen (陳章榮)

National Center for Toxicological Research, FDA, Jefferson, AR 72079, USA

Abstract

A class-imbalanced classifier is a decision rule to predict the class membership of new samples from an available dataset where the class sizes differ considerably. When the class sizes are very different, most standard classification algorithms may favor the larger (majority) class resulting in poor accuracy in the minority class prediction. A class-imbalanced classifier typically modifies a standard classifier by a correction strategy or by incorporating a new strategy in the training phase to account for differential class sizes. This study reviews and evaluates some most important methods for class prediction of high-dimensional imbalanced data. The evaluation addresses the fundamental issues of the class-imbalanced classification problem: imbalance ratio, small disjuncts and overlap complexity, lack of data, and feature selection. Four class-imbalanced classifiers are considered. The four classifiers include three standard classification algorithms each coupled with an ensemble correction strategy and one SVM-based correction classifier. The three algorithms are 1) diagonal linear discriminant analysis (DLDA), 2) random forests (RF), and 3) support vector machines (SVM). The SVM-based correction classifier is SVM threshold adjustment (SVM-THR). A Monte Carlo simulation and five genomic datasets were used to illustrate the analysis and address the issues. The SVM-ensemble classifier appears to perform the best when the class imbalance is not too severe. The SVM-THR performs well if the imbalance is severe and predictors are highly correlated. The DLDA with a feature selection can perform well without using the ensemble correction.

Keywords: class-imbalanced prediction, feature selection, lack of data, performance metrics, threshold adjustment, under-sampling ensemble.

Randomized SUP: A Clustering Algorithm for Large-scale Data

須上英*

台北大學統計學系

金妍秀、陳定立

中央研究院統計科學研究所

Abstract

The self-updating process (SUP) is a clustering algorithm which iteratively updates every data point according to its neighboring points. Although SUP has been shown to be particularly competitive in clustering (i) data with noise and (ii) data with a large number of clusters, the algorithm relies on the pairwise similarities between data points, which becomes computationally inefficient for large scale data.
In this talk we will present a randomized approach to overcome the computational difficulty. At each iteration, we consider only relatively small portions of data for location updates. The Law of Large Numbers guarantees that the result of the randomized updating process converges to that of the original SUP when the number of data points becomes large. Simulations as well as real data will be presented to show the clustering performance and the computational efficiency of the proposed randomized algorithm.

Keywords: SUP, clustering, randomized algorithm

Three-level A- and D-optimal Paired Choice Designs

蔡風順

Institute of Statistical Science, Academia Sinica

Abstract

In a paired choice experiment, several pairs of options are shown to respondents and the respondents are asked to give their preference among the two options for each of the choice pairs shown to them. Under the utility neutral effects coding set-up, in the literature, D-optimal designs have been obtained mostly for situations where a balanced design exists. In this paper, we consider attributes each at three levels and obtain lower bounds to the A- and D-values. We provide new A-optimal and D-optimal designs for estimating main effects under the utility neutral multinomial logit model setup. A- and D-efficient designs have also been obtained.

Keywords: choice set, main effects, Hadamard matrix.

A Generalized Class of Quaternary Code Designs with Good Resolution and Aberration

Frederick K. H. Phoa

Institute of Statistical Science, Academia Sinica

Abstract

The study of good nonregular fractional factorial designs has received significant attention over the last two decades. Recent research indicates that designs constructed from quaternary codes (QC) are very promising in this regard, although the proposed designs in the literature focused mainly on the best QC designs under maximum generalized resolution criterion. This talk introduces a generalized class of QC designs that have good resolutions and wordlength patterns when comparing to the traditional class of QC designs. In addition, a new criterion is introduced to classify fractional factorial designs when resolution and wordlength pattern are insufficient to distinguish their goodness.

Blocking in Partially Replicated Two-level Factorial Designs

Shin-Fu Tsai (蔡欣甫)

Division of Biometry, Institute of Agronomy, National Taiwan University

Abstract

Blocking is a practical technique for experimentation, which allows a researcher to precisely quantify the experimental error variance, especially when the experimental units are drastically heterogeneous. In this talk, a new series of blocking schemes, which considers some partially replicated design points in each block, is introduced for assigning the treatment combinations to several homogeneous subgroups of units. A noteworthy feature of the proposed designs is that the within-block and between-block replicates are both conducted, leading to that the model independent estimates for the variance components can be obtained easily. Based on these repeated design points, formal testing procedures are developed for identifying active factorial effects as well as significant block variance component. The proposed design and analysis methods are illustrated through a simulated experiment.

Keywords: lack-of-fit test, optimal design, orthogonal blocking, screening experiment, variance component.

Sizing Clinical Trials with Two Survival Outcomes

Tomoyuki Sugimoto*

Department of Mathematics and Computer Science, Kagoshima University Graduate School of Science and Technology, Japan

Toshimitsu Hamasaki

Office of Biostatistics and Data Management, National Cerebral and Cardiovascular Center, Japan

Scott R. Evans

Department of Biostatistics and the Center for Biostatistics in AIDS Research, Harvard School of Public Heath, USA

Takashi Sozu

Department of Management Science, Faculty of Engineering Division I, Tokyo University of Science, Japan

Abstract

Clinical trials with time-to-event outcomes as primary contrasts are common in many disease areas, such as infectious disease, oncology, and cardiovascular disease. Use of multiple endpoints creates challenges in the evaluation of power and sample size during trial design. Sample size determination for time-to-event outcomes is more complex and may require more aspects to be considered, compared with other scale endpoints such as continuous or binary outcomes.

We discuss methods for calculating the power and sample size for randomized superiority clinical trials with two correlated time-to-event outcomes, in the three independent or dependent censoring schemes where two endpoints are of non-fatal outcomes, when only one endpoint is of fatal outcome (semi-competing risk), and two endpoints are of fatal outcomes (full-competing risk). We derive an asymptotic form of the bivariate logrank statistic in all the three censoring schemes, including the correction structure when a composite of fatal and non-fatal endpoints is considered, and investigate the behavior of power and the required sample sizes, in clinical trial of two intervention comparison with the two inferential goals, i.e., where the trial is designed to evaluate if the intervention is superior to the control on all endpoints (multiple co-primary), or the trial is designed to evaluate if the intervention is superior to the control on at least one endpoint (multiple primary).

Keywords: dependent censoring, Logrank test, multiple endpoints, semi-competing risk, time-dependent association.

Combining Evidence of Regional Treatment Effects under DREM in MRCTs

Hsiao-Hui Tsou* (鄒小蕙) and Chi-Tian Chen (陳啓天)

Institute of Population Health Sciences, National Health Research Institutes

K. K. Gordon Lan (藍光國)

Janssen Pharmaceutical Companies of Johnson & Johnson, Raritan, NJ, USA

Chin-Fu Hsiao (蕭金福) and Jung-Tzu Liu (劉容慈)

Institute of Population Health Sciences, National Health Research Institutes

Abstract

In recent years, developing pharmaceutical products through a multiregional clinical trial (MRCT) has become standard. Traditionally, treatment effect was assumed to have a fixed value over all regions in an MRCT. However, regional heterogeneity caused by differences in race, genetic fingerprints, diet, environment, culture, medical practice, and disease etiology among regions in MRCTs has been observed and may have impact upon a medicine's treatment effect. In this presentation, we will discuss the issue of combining evidence of regional treatment effects in multiregional clinical trials. We will introduce a discrete random-effects model (DREM) to account for heterogeneous treatment effects across regions for the design and evaluation of MRCTs. We will illustrate determination of overall sample size and address issues of sample size re-calculation due to the uncertainty in the treatment effect or variability assumptions.

Keywords: MRCT, DREM, regional heterogeneity, combining evidence.

Interim Evaluation of Efficacy or Futility in Group-sequential Trials with Multiple Co-primary Endpoints

Toshimitsu Hamasaki* and Koko Asakura

National Cerebral and Cardiovascular Center

Scott R. Evans

Harvard T.H. Chan School of Public Heath

Abstract

We discuss group-sequential designs in superiority clinical trials with multiple co-primary endpoints, i.e., when trials are designed to evaluate if the test intervention is superior to the control on all primary endpoints. We consider several decision-making frameworks for evaluating efficacy or futility, based on boundaries using group-sequential methodology. We incorporate the correlations among the endpoints into the calculations for futility boundaries and sample sizes as a function of other design parameters including mean differences, the number of analyses, and efficacy boundaries. We investigate the operating characteristics of the proposed decision-making frameworks in terms of efficacy/futility boundaries, power, the Type I error rate and sample sizes, while varying the number of analyses, the correlations among the endpoints, and the mean differences. We provide an example to illustrate the methods and discuss practical considerations when designing efficient group-sequential designs in clinical trials with co-primary endpoints.

Keywords: error-spending method, futility, multiple endpoints, non-binding boundary, type I and type II error adjustments.

Combined Multistep Forecasts in Time Series Using Composite Least Squares

Nan-Jung Hsu* (徐南蓉) and Chun-Hsien Li (李俊賢)

Institute of Statistics, National Tsing-Hua University

Abstract

Multi-step forecasts have long been studied in time series literature. Among linear forecasts, the direct and plug-in (also called recursive or iterated) methods are both popular in use and theoretically justified while their empirical performance relative to the other is depending on the tradeoff between the bias and estimation variance, which is typically sensitive to the working model, the forecast horizon, and the underlying data scenario. This work concerns about a combined method via a composite least squares (CLS) approach hoping to take the advantages of both the direct forecast and the plug-in forecast. To optimize the combined strategy in CLS, we further suggest a cross validation criterion to minimize the multistep forecast mean square errors. Simulation studies show that the proposed CLS-forecast performs effectively and adaptive towards to the better one among the traditional direct and plug-in forecasts under a variety of linear and nonlinear data generating scenarios. Some issues regarding model selection and computations are also addressed.

Keywords: composite least squares, cross-validation, multi-step forecasts

Asymptotic Inefficiency of BIC and Asymptotic Efficiency of TSIC: The Case of an $I(d)$ Process

Shu-Hui Yu* (俞淑惠)

National University of Kaohsiung

Chor-Yiu Sin

National Tsing Hua University

Abstract

We consider in this paper an $I(d)$ autoregressive (AR) process, is an unknown integer. While Sin and Yu (2015) show that Akaike's information criterion (AIC) is asymptotically inefficient when the lag order is finite; this paper shows that when the lag order is infinite with (a) exponentially decaying AR coefficients, or (b) algebraically decaying AR coefficients, Bayesian information criterion (BIC) is asymptotically inefficient. These results motivate us to combine the strengths of AIC and BIC, yield a so-called two-stage information criterion (TSIC) for a general $I(d)$ AR process. We show that TSIC is asymptotically efficient in the aforementioned three scenarios. The paper concludes with a simulation study.

Keywords: asymptotic efficiency, $I(d)$, Bayesian information criterion (BIC), mean squared prediction error (MSPE), penalty term, same-realization prediction

Conditional Tail Expectation for Non-stationary Processes

Hwai-Chung Ho (何淮中) and Hung-Yin Chen (陳虹吟)

Institute of Statistical Science, Academia Sinica

Henghsiu Tsai* (蔡恆修)

Institute of Statistical Science, Academia Sinica

Abstract

The present paper studies the estimation of the conditional tail expectation (CTE) for non-stationary integrated processes of returns which follows a general class of multivariate stochastic volatility model. The non-parametric estimate of the CTE we propose is easy to implement, and shown to be asymptotically normal with non-standard normalization. The coverage ratios for confidence intervals obtained in the Monte Carlo experiments are consistent with the theoretical findings, and demonstrate the superiority of our approach as well. Results on the estimation of the CTE for the long-horizon returns of the S&P 500 index and other world's major indexes are also presented.

Keywords: asymptotic normality, conditional tail expectation, integrated process, stochastic volatility model.

Markov Limit of Line of Decent Types in a Multitype Branching Process

Jyy-I (Joy) Hong (洪芷漪)

Department of Applied Mathematics, National Sun Yat-sen University

摘要

In a multitype (d types) supercritical positively regular Galton-Watson branching process, we want to investigate some properties about the population in the past, especially the type of the ancestors. Let $\{X_n, X_{n-1}, \ldots, X_0\}$ denote the types of a randomly chosen (i.e., uniform distribution) individual from the nth generation and this individual's n ancestors. It is shown here that this sequence converges in distribution to a Markov chain $\{Y_0, Y_1, \ldots\}$ with transition probability matrix $(p_{ij})_{1\le i,j\le d}$ and Y_0 having the stationary distribution.

Keywords: Branching process, Markov, supercritical, critical, ancestor, line of descent, types.

Value at Risk for Integrated Returns and Its Applications to Equity-linked Insurance

Hung-Yin Chen* (陳虹吟) and Henghsiu Tsai (蔡恆修)

Institute of Statistical Science, Academia Sinica

Hwai-Chung Ho (何淮中)

Institute of Statistical Science, Academia Sinica
Department of Finance, National Taiwan University

Abstract

The present paper investigates the distribution quantile for integrated portfolio returns that follow a general class of multivariate stochastic volatility model. We propose a non-parametric quantile estimate that incorporates the rate with which the true quantile diverges as the integration horizon expands. The asymptotic normality established for the estimate enables us to construct the confidence interval for the true quantile. Monte Carlo experiments are conducted to demonstrate both the consistency and the advantages of our approach. Results on quantile estimates for the return distribution of the S&P 500 index are also presented.

Keywords: quantile, Value at Risk, stochastic volatility model, integrated returns.

Reference Curve Selection in a Class of Misaligned Curves

Yu-Hsiang Cheng* (鄭宇翔)

Institute of Statistical Science, Academia Sinica

Tzee-Ming Huang (黃子銘)

Department of Statistics, National Chengchi University

Abstract

Functional data usually have a common pattern with some variation in time. In order to estimate the common shape, curves need to be aligned to a reference curve. Some previous studies show the choice of reference curve affect the estimation accuracy. In this talk, several approaches for choosing reference curve will be introduced. Simulation results will be presented to demonstrate the performance of these approaches.

Keywords: curve alignment, reference curve

Evaluation of Benefit and Consistency of Treatment Effect under a Discrete Random Effects Model in Multiregional Clinical Trials

Chi-Tian Chen* (陳啓天) and Jung-Tzu Liu (劉容慈)

Institute of Population Health Sciences, National Health Research Institutes

Gordon Lan (藍光國)

Janssen Pharmaceutical Companies of Johnson & Johnson, Raritan, NJ, USA

Chyng-Shyan Tzeng (曾晴賢), Chin-Fu Hsiao (蕭金福) and Hsiao-Hui Tsou (鄒小蕙)

Institute of Bioinformatics and Structural Biology, National Tsing Hua University

Abstract

Consider the impact on the drug effect due to regional differences; the traditionally uniform treatment effect assumption may be inappropriate in an multiregional clinical trial (MRCT). Lan and and Pinheiro (2012) proposed a discrete random effects model (DREM) to address the heterogeneous treatment effects among regions. However, the benefit of the overall drug effect and the consistency of the treatment effect in each region are two major issues in MRCTs. In this research, the power function is derived for a beneficial treatment effect under DREM and the overall sample size determination in an MRCT. We assess the treatment benefit and consistency simultaneously under DREM based on the Method 2 in "Basic Principles on Global Clinical Trials" guidance published by the Japanese Ministry of Health, Labour, and Welfare. We also optimize the sample size allocation to reach maximum power for the benefit and consistency. A practical problem in MRCT design is that the different treatment effects across regions are unknown. Thus, we provide some guidelines on the design of MRCTs with consistency when the regional treatment effects are assumed to fall into a specified interval.

Keywords: multiregional clinical trial, discrete random effects model, consistency, power for benefit and consistency, optimal sample size allocation.

Importance Sampling Methods for Sensitivity Estimations

Hui-Ming Pai (白惠明)

Department of Statistics, National Taipei University

Abstract

The estimation of derivatives, especially of rare events, is an interesting but challenging problem. It has applications in a wide range of fields including industrial engineering, epidemiology and finance, etc.. For example, the derivative of ruin probability of stochastic processes can finds its application in communication network, risk management and vice versa. One of the common approaches to estimate the derivative is the Monte Carlo method. In this talk, we are developing an importance sampling Monte Carlo algorithm for estimating derivative of probabilities or functionals of probabilities. The technique is based on the connection of the problem with a differential game and its associated Isaacs equation. In terms of the subsolution of the Issacs equation, one can derive dynamic change of measures which make the importance sampling scheme asymptotically optimal.

The Localization of One-dimensional Self-attractive Polymers

Chien-Hao Huang (黃建豪)

Mathematics Division, National Center for Theoretical Sciences

Abstract

A polymer chain with attractive forces is considered. The energy of this model is related to the L2 norm of the space variable of Brownian local time processes. The behavior of the polymer chain converges to a mixture of Markov processes under the transformed measure.

Keywords: large deviations, weak convergence, attractive interactions.

Phase Retrieval with Random Illumination

Gi-Ren Liu (劉聚仁)

Department of Mathematics, National Tsing Hua University

Abstract

In many areas of imaging science, it is difficult to determine the phase of linear measurements. For example, in the X-ray imaging, the detector can only measure the Fourier magnitude of the received optical wave. In this setting, the goal of the phase retrieval is to reconstruct the unknown image from its Fourier magnitude data. Due to the absence of the phase information, the phase retrieval does not have a unique solution. In 2012, Albert Fannjiang proved that randomly modifying the phases of the unknown image by a mask can lead to a unique solution up to a global phase factor. Apart from considering the uniqueness of the phase retrieval problem, the simulation results show that random illuminations also significantly improve the numerical performance of the Error-Reduction (ER) algorithm, which is the most popular phase-retrieval algorithm. This talk describes the mathematical formulation of the phase retrieval problem, why the random illumination can improve the performance of the ER algorithm, and what breakthroughs about the convergence of the ER algorithm have been made compared with related works.

Session III

105 年 6 月 25 (星期六，10:20∼12:00)

III-1 (SS1001) **Invited Session : Data Mining/Big Data (I)**

III-2 (SS1003) **Contributed Session : Experimental Design/Biostatistics**

III-3 (SS1004) **Invited Session : Sampling Survey**

III-4 (SS1005) **Invited Session : Statistical and Machine Learning (II)**

III-5 (SS1006) **Contributed Session : Applied Statistics (I)**

III-6 (SS2002-1) **Contributed Session : Mathematical Statistics/ High-Dimensional Analysis**

III-7 (SS2003-2) **Invited Session : Internetional session (III)**

III-8 (SS2005) **Invited Session : Econometrics**

III-9 (SS2006) **Contributed Session : Young Researcher Session (II)**

III-10 (SS2012) **Invited Session : Social Statistics**

大數據商業化策略

鄭宇庭

國立政治大學統計學系

摘要

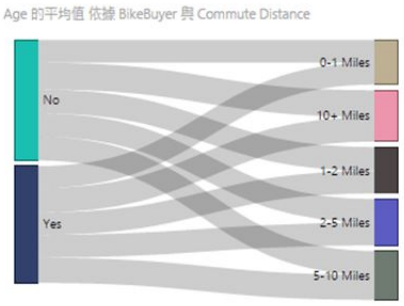
從 2012 年到 2016 年，大數據持續發威，IDC 市場調查機構預測 2017 年大數據市場將價值 324 億，很多企業、研究機構正在利用大數據激盪出最有創意的構想，用數據的力量來創造出大量的價值。從讓資料搜尋分析平台，大數據新創公司這幾年內如雨後春筍般出現，2016 年這些「Big Data Startups」也會持續挖掘大數據的價值、微調其企業策略，期盼在市場中展露頭角。將介紹他們如何應用大數據創造出商業價值，也從中瞭解大數據的應用方向及潛能。

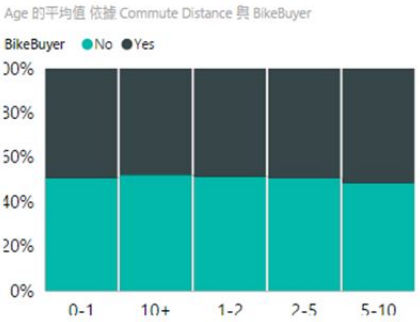
消費行爲之大數據視覺化分析

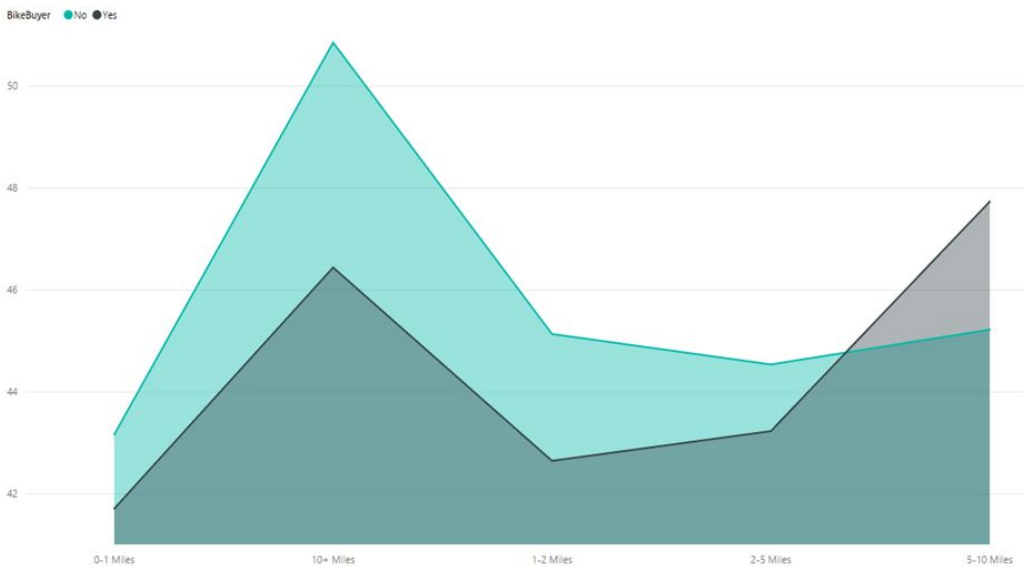
顧明*、吳格非

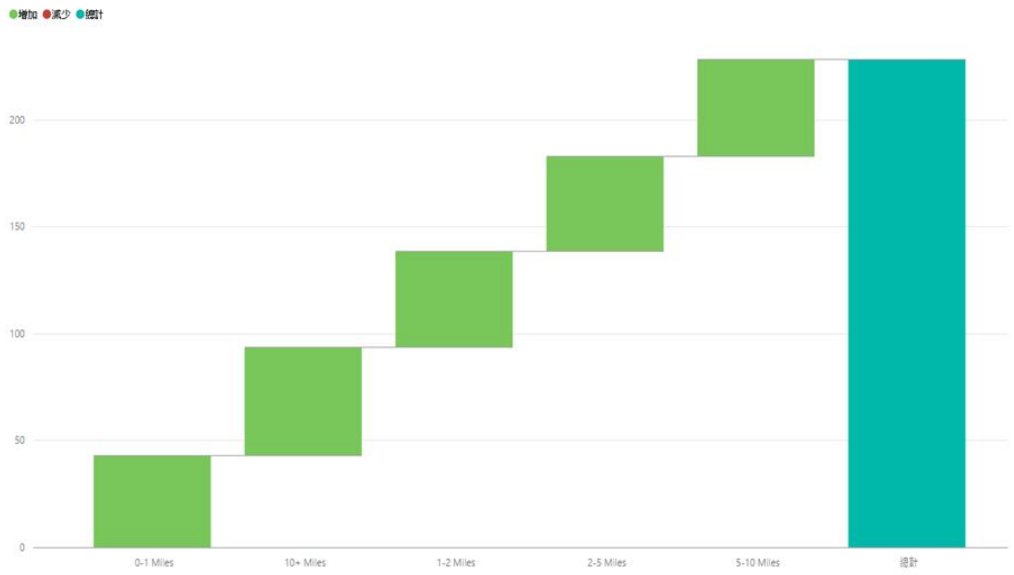
輔仁大學商學研究所博士班

摘要







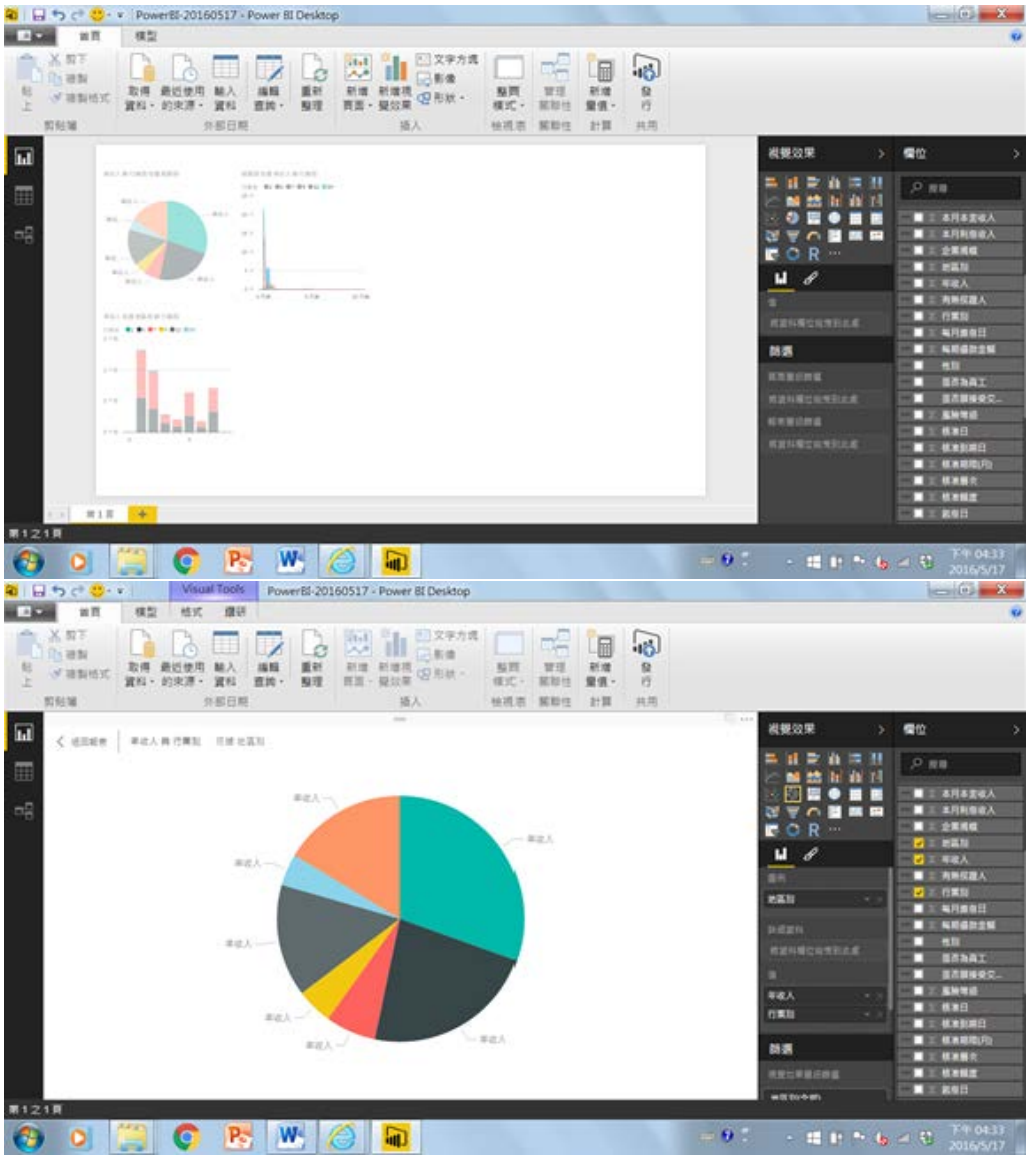

增加
減少
總計
200
150
100
50
0
0-1 Miles
10+ Miles
1-2 Miles
2-5 Miles
5-10 Miles
總計

金融財務之大數據視覺化分析

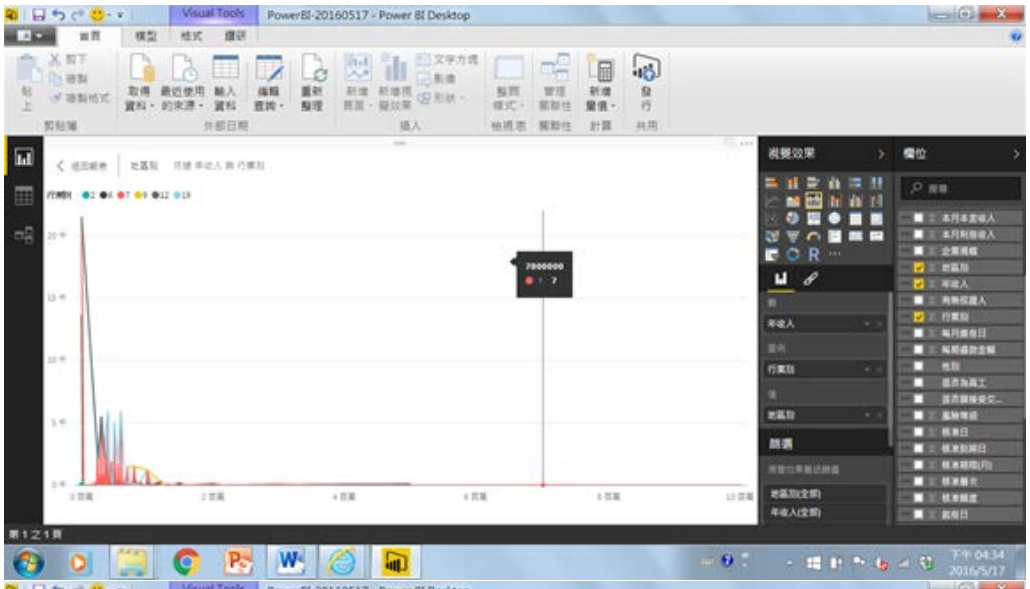
杜明晃*、陳世訓

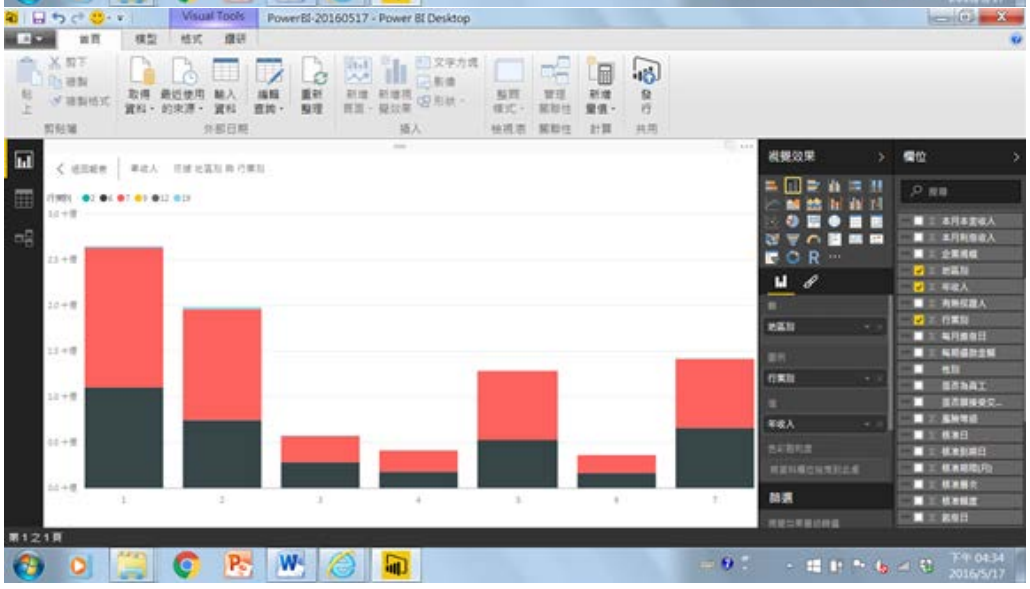
輔仁大學商學研究所博士班

摘要



[III-1-3]


Visual Tools
PowerBI-20160517 - Power BI Desktop
7000000
第 1 之 1 頁

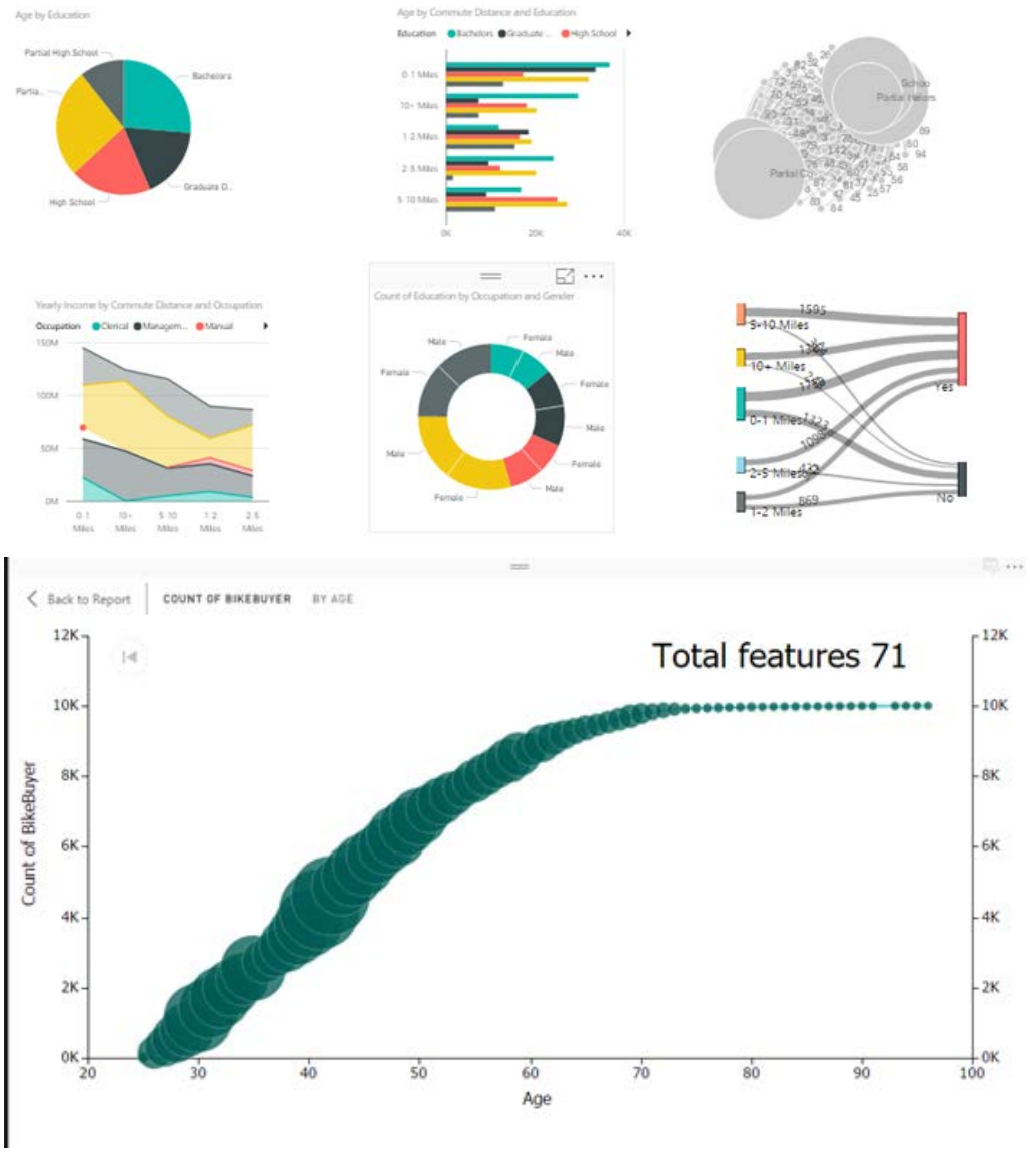

Visual Tools
PowerBI-20160517 - Power BI Desktop
第 1 之 1 頁

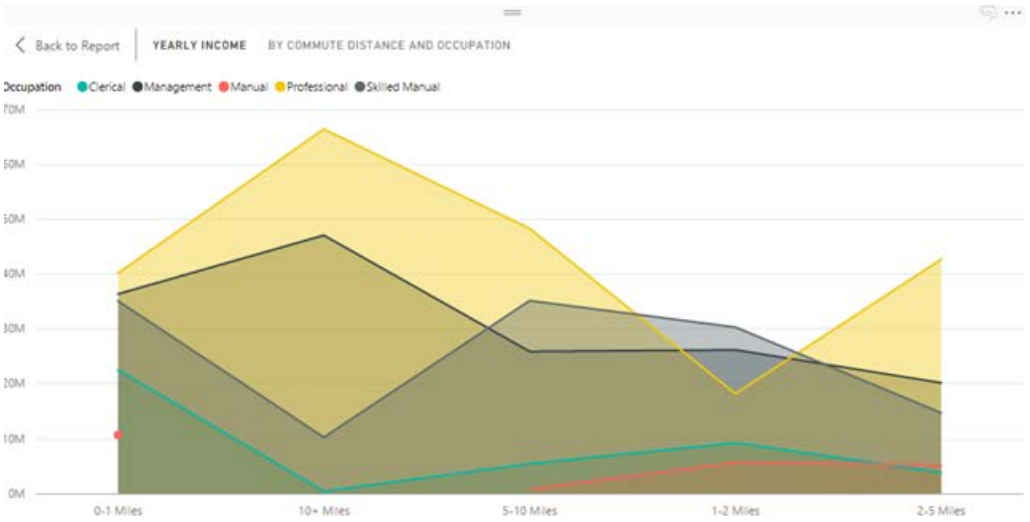
從國內六大報系探討臺北市國際觀光旅館與新聞媒體之關係研究

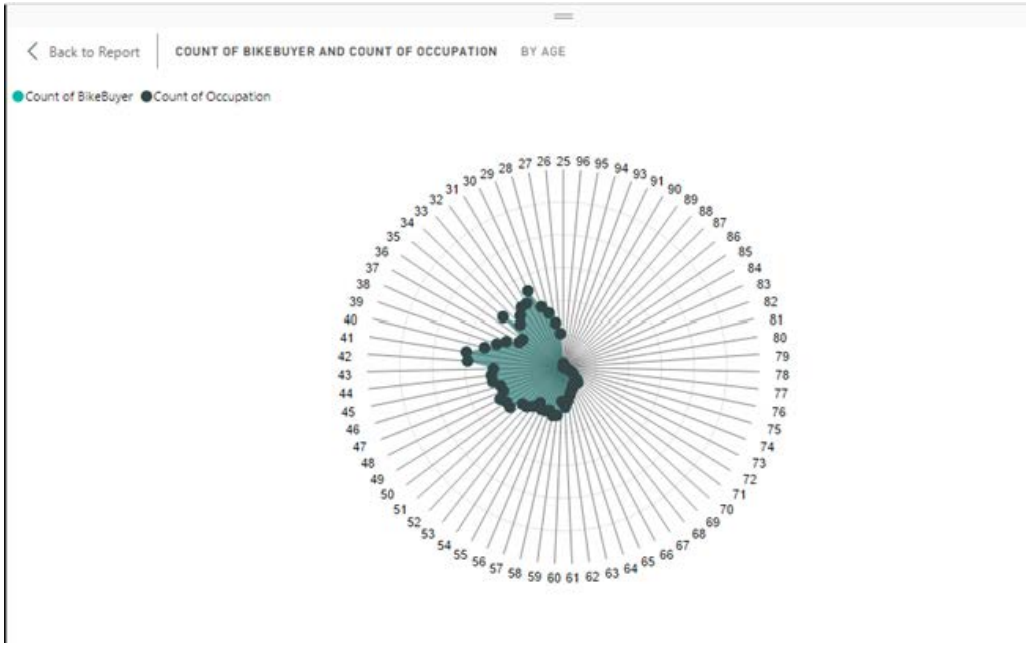
廖美蘭、車成緯*

輔仁大學商學研究所博士班

摘要




Back to Report
YEARLY INCOME
BY COMMUTE DISTANCE AND OCCUPATION
Occupation
Clerical
Management
Manual
Professional
Skilled Manual
70M
50M
40M
30M
20M
10M
0M
0-1 Miles
10+ Miles
5-10 Miles
1-2 Miles
2-5 Miles


Back to Report
COUNT OF BIKEBUYER AND COUNT OF OCCUPATION
BY AGE
Count of BikeBuyer
Count of Occupation

Statistical Methods for Detecting Modifier Genes of Survival in a Mouse Model of Dravet Syndrome

Chen-Hung Kao

Institute of Statistical Science, Academia Sinica

Abstract

The evidence for the influence of genetic background (modifier genes) on the major genes is old and has been reported by many researchers in plant, fruit fly, mice and human, etc. These modifier genes modulate the penetrance, dominance and expressivity of a major gene, so that the expression of a specific genotype at the major gene will show a wide phenotypic range (in a quantitative fashion) under different genetic background. These modifiers act the same way as quantitative trait loci (QTL) in controlling the quantitative traits. Therefore, the modifiers have been regarded as QTL, and statistical methods of QTL mapping have been applied to identify the modifiers in numerous studies. Motivated by a study of a strain-dependent difference in survival (age at time of early death) in the mouse model of Dravet syndrome, we outline and propose statistical analyses to detect and analyze modifier loci responsible for the severity in the study. The issues related to QTL mapping are also presented and discussed.

Keywords: QTL mapping, subpopulations, modifiers, survival analysis.

Sample Size Re-estimate for Cluster Randomized Trial

Shu-Mei Wan

Department of Finance, Lunghwa University of Science and Technology

Chien-Hua Wu*

Department of Applied Mathematics, Chung-Yuan Christian University

Abstract

Cluster randomization trials have been grasped attention since the publication of the article by Cornfield. The clusters of randomization in clinical trials are clinics and/or hospitals, for example. The correlation among observations within the cluster, typically measured by intracluster correlation coefficient (ICC), tends to be positive. It is well known that such trials may have substantially underestimated the size of sample if ICC is not taking into account in the sample size calculation. van Breukelen and Candel propose simple guidelines for calculating sample sizes of cluster randomized trials.

Because of uncertainty in the initial estimate of ICC, it is desirable to reevaluate the sample size in the middle of the trial. The European Union's "Note for Guidance" (1995) from the Commission for Proprietary Medical Products (CPMP) addresses the importance issue that the sample size recalculation is performed such that all persons involved in the study remain blinded to the treatment group allocation. Gould and Shih (1992) uses an EM algorithm to estimate individual group means and variances for continuous cases, and Shih and Zhao (1997) stratifies the randomization procedure to re-estimate the sample size for binary outcome. Wu(2015) and Wu(2016) provide an extension of the work by Shih and Zhao (1997) to continuous endpoints and crossover design, respectively. However, none of them evaluate the sample size for the cluster randomized trials in the middle of study without breaking the blind. In this article, we study the sample sizes of cluster clinical trials.

Keywords: ICC, sample size, crossover design, cluster randomized trials

A New Construction Method for Space-filling Designs via Factor Collapse

Cheng-Yu Sun* and Frederick Kin Hing Phoa

Institute of Statistical Science, Academia Sinica

Shaowei Cheng

National Tsing-Hua University

Abstract

Space-filling designs have been commonly used in computer experiment, random sampling, and other scientific studies. We propose a new method to construct space-filling designs. Using techniques in Galois field, we collapse the factors' levels in a regular fractional factorial design (FFD) and the strength of the collapsed design is enhanced. By reversing the factor collapse to relabel the factor levels, the resulting regular FFD enjoys good space-filling properties. Compare with the existing methods in level permutations, the factor collapse method is efficient, especially when the number of factor level is high.

Moreover, we propose a criterion, namely maximal strength efficiency, to distinguish two collapsers when they result in collapsed designs with the same strengths. This criterion not only maximizes the strength of the collapsed design, but also maximizes the proportion of the projected sub-designs that are full factorials. This method can be sequentially applied when the collapsed design is a regular FFD, which is possible under certain situations. Some demonstrations are presented, including one constructed design that is better than the existing one in terms of discrepancies.

On Estimating the Mean of a Quantitative Sensitive Character by Using Randomized Response Model

Chih-Li Wang (王智立)

Department of Applied Statistics and Information Sciences, Ming Chuan University

Abstract

Social surveys sometimes include sensitive or threatening issues of enquiry for which people are not inclined to state honest response. Social stigma and fear of reprisals often lead respondents to give biased, misleading or even erroneous responses when approached with a direct response (DR) survey method. To procure reliable data for estimating the proportion of persons possessing a qualitative sensitive attribute, Warner (1965) proposed an ingenious procedure called randomized response (RR) technique. Consider a dichotomous population in which every person belongs either to a sensitive group A or to the non-sensitive complement A^C. The problem of interest is to estimate the population proportion π of individuals who are members of A.

Greenberg *etal*. (1971) observed that the randomized response model need not be restricted to qualitative trials, but may be applied to the cases of quantitative characters as well for the sensitive problem. Moreover, Singh (1999) also observed that the RR procedure to the cases of quantitative characters as well. Let X be a sensitive quantitative character and Y be a non-sensitive quantitative character. The purpose is to estimate the unknown mean μ_x of X.

The paper mainly focuses on how to employ the randomized response model to assist the survey of sensitive character so that the rejection rate of interview can be reduced and the privacy of interviewees will be protected well. We expect to obtain better estimator and interview quality. Toward these purposes, the following works have to be done: 1. I propose a weighted mean randomized response model for sensitive problem when the data is quantitative. Further, I also considered a distribution of the random number S, which leads the estimation strategies to be available in practical application. 2. It is founded that the proposed model performs better than traditional ones.

Keywords: agriculture survey, primary farm household, sampling strategy, stratified sampling, stratum boundary, optimal allocation

Sampling Design for the Primary Farm Household Survey of the Taiwanese Agriculture Study

Chang-Tai Chao* (趙昌泰) and Chien-Min Huang (黃倩旻)

National Cheng Kung University

Chiu-Yen Lee (李秋嬿), Shiow-Ing Lee (林秀霙) and Yu-Wen Liu (劉玉文)

Council of Agriculture, Executive Yuan

Abstract

The Census of Agriculture, Forestry, Fishery and Animal Husbandry is an important official survey in Taiwan. However, enormous survey cost and effort for the data processing are required for such a nationwide census, hence this census can only be conducted every five years. Therefore, certain annual agriculture survey is necessary for the realization of the current related industry information, so that proper policy can be formulated timely. A sampling strategy for the Taiwanese primary farm household survey is constructed in this research. A stratified random sampling design, in which the optimal stratum boundary and allocation are carried out based on the 2010 census data, is proposed for the purpose to enhance the estimation precision and investigate certain subpopulations of interest. The result indicates that the performance of the proposed stratified sampling is more advantageous than simple random sampling without replacement and other stratified design, such as one with stratification boundaries to equalize the stratum size and optimal within-stratum sample size under a comparable total sample size. In fact, the proposed design has been practiced successfully since 2014, and further modification of the proposed design will be addressed.

Keywords: agriculture survey, primary farm household, sampling strategy, stratified sampling, stratum boundary, optimal allocation

A Bayesian Approach in Analyzing Randomized Response Techniques for Multi-level Attribute Data

Shu-Hui Hsieh* (謝淑惠)

Research Center for Humanities and Social Sciences, Academia Sinica

Shen-Ming Lee (李燊銘)

Department of Statistics, Feng Chia University

Abstract

In survey research, Warner (1965) pioneered randomized response techniques for overcoming the challenge of collecting reliable answers to sensitive survey questions such as drug abuse history, homosexual activities and AIDS, abortion experience, family income, etc. In general, the randomized response method guarantees that individual attributes are not revealed by the responses, and as long as respondents understand and trust this protection, this can reduce non-response and underreporting biases. In this study, we develop a Bayesian approach for estimating proportions of individuals who possess one of multiple possible non-overlapping attributes using a single sensitive question by Hsieh et al. (2016), since a Bayes estimator may assume values solely within the parameter space. To investigate the performance of Bayesian estimators, we conduct a simulation study of the relative efficiency of the proposed method and maximum likelihood method. The technique is illustrated using data from the 2012 Family and Gender Module of the Taiwan Social Change Survey to estimate the proportions of individuals of different sexual orientations.

Keywords: randomized response techniques, Bayesian approach, Taiwan social change survey, sexual orientations.

Sparse Representation for Time Series Classification

Yuh-Jye Lee

Department of Applied Mathematics, National Chiao Tung University

Abstract

The problem of time series classification has been studied for over a decade. In the era of Internet of Things, time series data has become a major data type data, much effort has been devoted to this issue. The approaches to time series classification can be categorized into three types, including distance-based, model-based, and feature-based approaches. In this research, we focus on feature-based methods, which represent time series into a set of characterized values. However, features generated by most of existing representation techniques are not completely interpretable. Due to this fact, a novel time series representation, envelope, is proposed. The envelope is a profiling for a set of time series. This is a supervised feature extraction method that encodes time series into three numbers, -1, 0 and 1. If time series value falls into the envelope then encodes it as 0. We use -1 and 1 to represent the value falls outside below and above respectively. It is always important to find the most discriminating features for data mining tasks. Hence, we need a good heuristic to decide the size of the envelope in order to have a better performance either in data classification or anomaly detection tasks. Moreover, this new representation enjoys the characteristic of sparsity which is an essential property for applying compressed sensing. With this advantage, we can benefit from high transmission efficiency, the reduction of required storage and model complexity. Furthermore, the transformed features are interpretable via visualization. Envelope shows the shape of time series and defines the similarity between which. Disclosed below is the demonstration of the effectiveness of proposed method on numerous benchmark datasets.

A Low-rank Based Estimating and Testing Procedure for Matrix-covariate Regression

Hung Hung* and Zhi-Yu Jou

Institute of Epidemiology and Preventive Medicine, National Taiwan University

Abstract

Matrix-covariate is now frequently encountered in many biomedical researches. It is common to fit conventional statistical model by vectorizing the matrix-covariate. This strategy, however, results in a large number of parameters, while the available sample size is relatively small to have reliable analysis results. To test the significance of matrix-covariate, score test is widely used to overcome the curse of high-dimensionality by treating the coefficient of interest as a random vector. Although score test performs well in many situations, it cannot provide an estimate for the effect sizes of matrix-covariate. When the research interest focuses on the effect sizes, one still needs to fit a conventional regression model, and therefore faces the problem of high-dimensionality again. In this work, we overcome the problem of high-dimensionality by utilizing the inherent structure of matrix-covariates. The advantage is that estimation and hypothesis testing can be conducted simultaneously as in the conventional case, while the estimation efficiency and detection power can be largely improved due to a parsimonious parameterization. Our method is applied to test the significance of gene-gene interactions in the PSQI data, and is applied to test if electroencephalography is associated with the alcoholic status in the EEG data.

Learning Cross-domain Landmarks for Heterogeneous Domain Adaptation

Yi-Ren Yeh* (葉倚任)

Department of Mathematics, National Kaohsiung Normal University

Yao-Hung Tsai (蔡曜宏) and Yu-Chiang Wang (王鈺強)

Research Center for Information Technology Innovation, Academia Sinica

Abstract

While domain adaptation (DA) aims to associate the learning tasks across data domains, heterogeneous domain adaptation (HDA) particularly deals with learning from cross-domain data which are of different types of features. In other words, for HDA, data from source and target domains are observed in separate feature spaces and thus exhibit distinct distributions. In this work, we will present a novel learning algorithm of Cross-Domain Landmark Selection (CDLS) for solving the above task. With the goal of deriving a domain-invariant feature subspace for HDA, our CDLS is able to identify representative cross-domain data, including the unlabeled ones in the target domain, for performing adaptation. In addition, the adaptation capabilities of such cross-domain landmarks can be determined accordingly. This is the reason why our CDLS is able to achieve promising HDA performance when comparing to state-of-the-art HDA methods. We conduct classification experiments using data across different features, domains, and modalities. The effectiveness of our proposed method can be successfully verified.

Keywords: heterogeneous domain adaptation, landmark selection, semi-supervised learning

消費者體驗行銷與購後行爲之研究—以麗寶樂園爲例

羅琪*

中華大學餐旅管理學系

溫芳玉、許雅雯、洪嫩璇、曾佩楹

中華大學應用統計學系

摘要

從行銷的觀點來看，體驗行銷是消費者透過對事件的觀察或參與，感受到某些刺激所誘發的思維認同或購買行爲。所以體驗行銷不是著墨於產品本身，而是提供一個知覺的、情感的、認知的、行爲的情境，讓消費者與商品產生互動；不同於傳統的行銷方式，體驗行銷傳達的是消費者的觀感或使用心得。因此以體驗行銷來吸引消費者將是未來的趨勢潮流 (Schmitt, 1999)。在樂園的經營管理上，若能有效運用體驗行銷，則可以提供遊客各種不同的體驗，並發展出與競爭者截然不同的特色，進而提昇遊樂園的經營績效。因此本專題是以麗寶樂園爲研究範圍，探討遊客生活型態、遊園所感受到的體驗形式，及購後行爲之關係。了解不同遊客的生活型態與消費特性，有助於各種行銷策略的擬定，提高入園遊客的體驗感受，促進其重遊意願，進而提昇遊樂園的經營績效。使用的統計方法有敘述統計、因素分析、變異數分析、信度分析及迴歸分析，使用的軟體爲SPSS。

研究結果發現：(1) 園內各項體驗媒介大致皆能帶給遊客特殊之體驗感受，由其是在園內設計上。在機械遊具、主題活動、解說標示設施、餐點飲料等媒介上，也提供遊客較深的感受。而吉祥物則較無法提供給遊客獨特的體驗；(2) 生活型態可分爲「流行時髦」、「精打細算」、「生活休閒」、「重視家庭」、「愛好社交」、「易受影響」六個因素構面。而購後行爲可分爲「正向推薦」、「負向抱怨」、「背叛」三個因素構面。(3) 遊客在職業、教育程度、個人月收入、與誰同行、預計停留時間、其他遊樂園的經驗項目上，對體驗結果呈現顯著的差異；另外遊客在性別、個人月收入、同行人數、與誰同行、預計停留時間項目上，對購後行爲呈現顯著的差異。(4) 行動體驗對正向推薦有顯著影響，而感官體驗、關聯體驗則對負向抱怨有顯著影響，思考體驗對背叛有顯著影響。

最後本專題建議園方可從加強體驗媒介之設計、塑造全方位的主題特色、依據遊客生活型態加以設計不同體驗活動、重點推畫體驗活動，以加強遊客購後行爲意圖四方面來著手，以提高遊客在不同體驗形式的體驗結果，進而提高遊客遊園後的正向購後行爲意圖。

關鍵詞：主題遊樂園、體驗行銷、策略體驗模組、生活型態、購後行爲

An Efficient Analysis of Change Points via Swarm Intelligence

Hsin-Hao Chen*, Livia Chang and Frederick Kin Hing Phoa

Institute of Statistical Science, Academia Sinica

Abstract

Evolutionary algorithm has received considerable attention in current statistics researches. Recently, Phoa (2016) and Phoa et al. (2016) proposed a nature-inspired metaheurisic method, namely the Swarm Intelligence Based (SIB) method, for efficient optimization in discrete and continuous domains. In this paper, we adopt the SIB method for analyzing composite functions that consists of multiple change points. Instead of simply applying the standard SIB framework, our algorithm introduces a new operation called VARY that allows the adjustment of number of change points to be included during the optimization process. Numerical results show that our algorithm successfully points out the location of change points accurately by using a small number of iterations, and the deviation between the fitted model and the true function is small.

Network Exploration by Complements of Graphs with Graph Coloring

Tai-Chi Wang* (王泰期)

National Center for High-performance Computing
National Applied Research Laboratories

Frederick Kin Hing Phoa (潘建興) and Yuan-Lung Lin (林遠隆)

Institute of Statistical Science, Academia Sinica

Abstract

Network data have become very popular with the growth of technologies and social applications such as Twitter and Facebook. However, few visualization tools are created for exploring large-scale networks. We propose a simple and quick procedure to explore a network in this study. The algorithm changes the edge representation based on the complement of a simple graph and the partition method of vertex coloring. Furthermore, the colors provide additional information on top of the partitions. Our proposed method is demonstrated in some famous networks.

Keywords: visualization, greedy Algorithm, network Partition, statistics of Labels, N-clique

飛行員雙眼深度知覺電腦化測驗之研發及資料分析

李紀蓮*、朱信、張維剛、文羽西、江國超

國軍高雄總醫院岡山分院航空生理訓練中心

黃碧群

國立成功大學心理系

摘要

前言：深度知覺 (立體感) 是國軍飛行人員在三度空間執行飛行任務的重要能力，爲甄選飛行人員視力檢查項目之一，爲精進現行儀器式深度知覺測驗的效度及探討不同視力組別其深度知覺之差異性，乃規劃本研究。

方法：(1) 進行電腦化雙眼深度知覺測驗的設計與研發，建構電腦化測驗之信效度，爲發展實用性之測驗奠基；(2) 探討電腦化測驗與儀器式深度知覺測驗之相關性，以爲電腦化測驗信效度建構之參考。

結果：(1) 完成「電腦化雙眼深度知覺測驗軟、硬體」的開發及設置，測驗類型區分「4 個圓圈」、「2 條直線」兩項，難度各自區分爲 7 級及 11 級。(2) 樣本計 110 人，含飛行員 (生) (66.3%)、一般軍事人員 (33.7%)，平均年齡 26.2 ± 4.8 歲 (20 ∼ 40 歲)，樣本區分爲視力良好組 (70.9%) 及戴鏡組 (29.1%)。(3) 儀器式深度知覺測驗分析：良好組與戴鏡組白天及夜間的立體感均無顯著差異。(4) 電腦化深度知覺測驗分析：單因子多變量變異數分析，良好組與戴鏡組在電腦化測驗「4 個圈閾值」及「2 條直線斜率 log」變項均無顯著差異。(5) 電腦化測驗與儀器式深度知覺測驗答對比率分析：將電腦化測驗 4 個圈閾值轉換爲角秒，並比照儀器式深度知覺測驗分級，二者 9 題全答對、答對7題以上、答對 5 題以上者分別爲 11.4%、29.5%、50.0% 及 46.5%、67.3%、89.1%，顯現電腦化量測立體感之靈敏度較儀器式深度知覺測驗爲佳。

結論：(1) 順利完成電腦化雙眼深度知覺測驗軟體開發及硬體設置，有效達成研究目標，提升國軍航空眼科學專業研究的能力及研究能量；(2) 電腦化雙眼深度知覺測驗量測精確度較儀器式深度知覺測驗爲佳，可有效區分深度知覺能力的差異；(3) 未來可運用於飛行員人員甄選或深度知覺能力訓練用。

關鍵詞：深度知覺測驗 Depth Test (stereoscopic vision test)、飛行員 Pilot、電腦化測驗 Computerlized Test

Random Partition Lasso for Large Scale Variables

Yen-Shiu Chin* and Ting-Li Chen

Institute of Statistical Science, Academia Sinica

Abstract

Sparsity of high-dimensional statistical models is essential for improved understanding and interpretation. The lasso (Tibshirani, 1996), a L1-norm penalized approach, is one of the most popular for high dimensional data since its sparse solutions and feasible computation compared with classical methods. However, for extremely high-dimensional datasets, e.g. more than one hundred thousand covariates, the lasso will take too much of time and space in computing. We introduce the random partition lasso: do the lasso for each smaller subset of predictor variables based on randomly partitioning. First of all, the procedure is efficient to screen out a great number of irrelevant covariates. By executing another lasso for the selected candidates from each subset, our proposed algorithm can select the same variables as the lasso does, but in much less time. Furthermore, it has a better chance to identify truly relevant variables in multiple small random subsets than in one large complete set. We prove that under some reasonable necessary conditions similar to the Irrepresentable Condition for the lasso, the random partition lasso is asymptotically sign consistent. Moreover, in certain of high-dimensional settings, our proposed algorithm identify true models while the classical lasso fails. We will present simulations which support our claim.

Keywords: high dimensional data, variable selection, Lasso, regularization, sparsity.

Non-Gaussian Process for Trait Evolution

D. C. Jhwueng (鍾冬川)

Department of Statistics, Feng-Chia University

Abstract

Phylogenetic comparative methods (PCMs) mainly use the Gaussian processes such as Brownian motion or Ornstein-Uhlenbeck process to describe continuous trait evolution on the phylogenetic tree. Under this assumption, trait values observed at the tips of the tree (comparative data) have multivariate normal distribution. However, for a group of related species, some traits show the diversity where the datasets have long tail and skewed distribution which violates the normality assumption common to the PCMs. In this project, we consider to use non Gaussian processes to describe continuous trait evolution. We develop a maximum likelihood framework for parameters estimation and apply our method to analyze empirical data from literature as well as simulation study to assess the model adequacy.

Keywords: dependence, phylogenetic comparative method, non-Gaussian process, diffusion bridges, trait evolution

A Unified Robust Score Statistic for Population Means Comparison

Hsiao-Yun Liu (劉小筠)

The General Education Center, Ming Chi University of Technology

Abstract

This dissertation deals with comparison of population means, similar to that of analysis of variance, in a way that the knowledge of the underlying distributions is absent. We develop a novel robust score test statistic that is akin to the familiar (observed-expected)/expected formula, with extra terms incorporating impact of the unspecified population moments.
We derive the test by correcting the score statistics from models including gamma, normal, Poisson, negative-binomial and inverse-Gaussian. These models, in spite of their diversity, give rise to a single unified corrected robust score statistic. Conditions under which our new robust test is more powerful than current competitors are provided. Finite sample performance is demonstrated via simulations and real data analysis.

Keywords: robust score statistic, analysis of variance.

On Unbiasedness and Sufficiency via Rao-Blackwell and on Correcting Traditional Kernel Smoothing

高正雄

國立中正大學統計科學所與國立成功大學統計學系

Abstract

Part I: Creating a minimum-variance unbiased estimator
The talk shall present the idea of using Rao-Blackwell theorem to create minimum-variance unbiased estimator by using the principle of moments to obtain an unbiased estimator first and having it followed by using a sufficient statistic to reduce the variance of the unbiased estimator, where the sufficient statistic is to obtainable from use of the principle of maximum likelihood. It shall also be shown that the effect to estimation accuracy by sufficiency is significantly more important than that by unbiasedness. Some illustrative examples shall be given.
Part II: The mistaken traditional Kernel smoothing
In the talk an integral representation of a function by introduction of symmetric kernel shall be presented. A new kernel smoothing result shall then follow, and the new kernel smoothing is to be shown that the traditional kernel smoothing often can produce seriously wrong approximation, which is a fact widely known by many for decades. Applications of the new kernel smoothing to interpolation and extrapolation, density estimations and nonlinear regressions shall also be discussed.

Sample Size Determination for the Assessment of Average Bioequivalence

Chieh Chiang* (姜杰) and Chin-Fu Hsiao (蕭金福)

Institute of Population Health Sciences, National Health Research Institutes

Jen-pei Liu (劉仁沛)

Institute of Population Health Sciences, National Health Research Institutes
Division of Biometry, Department of Agronomy, National Taiwan University
Institute of Epidemiology and Preventive Medicine, National Taiwan University

Abstract

The 1992 US Food and Drug Administration (FDA) indicates that two drugs are said to be average bioequivalent (ABE) if the log-transformed mean difference of pharmacokinetic (PK) responses lies in $(-0.223, 0.223)$. To assess ABE, the most widely used approach is the two one-sided test (TOST) procedure. More specifically, ABE is concluded when a $100(1-2\alpha)\%$ confidence interval for mean difference falls within (-0.223, 0.223). As known, bioequivalent studies are usually conducted by crossover design. However, in the case that the hale-life of a drug is longer, a parallel design for the bioequivalent study may be preferred. In this study, the two-sided interval estimation such as Satterthwaite's, Cocharn-Cox's or Howe's approximations is used for assessing parallel ABE. We show that the asymptotic joint distribution of the lower and upper confidence limits is bivariate normal, and thus the sample size can be calculated based on the asymptotic power so that the confidence interval falls within (-0.223, 0.223). Simulation studies also show that the proposed method achieve sufficient empirical power. A real example is provided to illustrate the proposed method.

Keywords: average bioequivalence, confidence interval, sample size determination

A Simulation Study to Compare Multiple Comparison Procedures for Multi-arm Clinical Trials

Kentaro Sakamaki

Department of Biostatistics, Yokohama City University Graduate School of Medicine, Japan

Abstract

In some multi-arm clinical trials, Hochberg procedure or Dunnett procedure is applied for multiplicity adjustment. Both procedures need similar assumption to control the familywise error rate. As mentioned in some studies, (step-down) Dunnett procedure may be more powerful than Hochberg procedure. Although, the reason regarding to selection of multiple comparison procedures is not clear because these procedures are hardly compared. Besides these procedures, the fixed sequence procedure is commonly used in multi-arm clinical trials because of the dose-response assumption. However, there is a few studies regarding to the operating characteristics of the fixed sequence procedure in various dose-response scenarios.
In this talk, we compare multiple comparison procedures for multi-arm clinical trials in a variety of simulation scenarios. To evaluate the operating characteristics of multiple comparison procedures, we use various measurements, such as marginal power, disjunctive power, and a probability of optimal choice.

Keywords: multi-arm clinical trials, multiple comparison procedures, operating characteristics, simulation study.

Influence Analysis for the Area under Receiver Operating Characteristic Curve with Application to Classification Assessment

Bo-Shiang Ke*

Institute of Statistics, National Chiao Tung University

Yuan-chin Ivan Chang

Institute of Statistical Science, Academia Sinica

Abstract

Performance measures are essential in classification research. Besides being used in assessment selection scenarios, these measures are sometimes included in the construction process of classifiers. Hence, how to prevent the selection and evaluation of classifiers from being affected by individual cases is an important research problem. Some model-based influential analyses are discussed in the literature. However, there is a lack of studies that focus on performance measures directly. In order to identify the influential observations that affect the estimate of the area under receiver operating characteristic curve, which is one of the most important classification performance measures, we propose several indexes based on the concepts of the influence function and local influence. Cumulative lift charts are used to equipoise the disagreements among the proposed indexes. Two real data sets are used for illustration purposes, and we compare the results of the new model, which are constructed after removing the influential cases identified by the proposed method, to those of the original ones reported in the literature.

Keywords: AUC, partial AUC, influence function, local influence, cumulative lift chart.

Using Heteroskedasticity-consistent Variances for OLS: Reviews and Some New Results

Cheng-Few Lee

Rutgers Business School, Rutgers University, U.S.A

Chor-Yiu Sin*

Department of Economics, National Tsing Hua University

Abstract

This paper reviews some of the properties of the heteroskedasticity consistent variances and heteroskedasticity non-consistent variances, also known as robust variances and non-robust variances, for OLS (ordinary least squares). Unlike the related papers in the literature, we discuss separately (i) the cases where the explanatory variables are strictly exogenous (see, for instance, Chapter 7 of Wooldridge, 2010); and (ii) the cases where the explanatory variables may or may not be strictly exogenous. The latter cases allow weakly dependent explanatory variables such as those generating from an autoregressive process. New results on the original robust variance (denoted by HC_0) and its variants (denoted by HC_1, HC_2, HC_3, HC_4 and HC_j, see Hausman and Palmer, 2012) are also derived. The conditional finite-sample distribution is derived only for the strictly exogenous data, while the asymptotic distribution is derived regardless the data is strictly exogenous or not. Our new results aim to throw light on the heteroskedasticity-robust part of the variance-covariance matrix, both in cross-sectional regressions and panel regressions (with time and/or firm effects).

Keywords: asymptotic distribution, conditional finite-sample distribution, non-robust variance, robust variance, strictly exogenous, weakly dependent.

Tests for the Presence of Clusters in Heterogenous Panel Data Models

Chang-Ching Lin* (林常青)

Department of Economics, National Cheng Kung University

Shou-Yung Yin (殷壽鏞)

Institute of Economics, Academia Sinica

Abstract

In this paper, we propose a set of tests for the existence of clusters when the slope parameters are heterogeneous in a panel of different groups but group membership is unknown to the econometrician. It is well-known that improperly imposing the assumption of slope homogeneity may lead to severe consequences of inference. Arranging observations into groups and then detecting the presence of clusters is one way to resolve the problems caused by model heterogeneity. However, the conventional tests for slope homogeneity cannot be applied to the groups because splitting might lead to dependent data. We propose an approach, which does not suffer from the above caveats. Particularly, we first estimate the slope coefficients for all individuals and then use them to perform the tests. This proposed approach, based on the multivariate L-statistics for whether the slope coefficients come from the same population or not, enables us to confirm the clusters more accurately. Finally, our proposed methods are applied to the existence of "convergence clubs" in the growth literature.

Keywords: L-statistics, bootstrapping, heterogeneous panel data.

Hierarchical Factor Models with Possible Non-stationary Components

Shih-Hsun Hsu (徐士勛)

Department of Economics, National Chengchi University

Abstract

While facing the large dimensional data, the factor model, which assumes the main fluctuations of all variables of interest are driven by only a few common factors, has thus become popular, and lots of its variants are introduced in the literature. In particular, to gain a better understanding of factors, the so-called top-down hierarchical factor model is established by imposing more economic structures on factors. Nevertheless, there are a couple of limitations in the exiting hierarchical factor models: (1) they work for the stationary data only, and (2) the number of factors of each layer must be presumed by researchers in advance of employing the maximum likelihood estimation or Bayesian methods. This paper thus aims to get round these limitations, while keeping the advantages of top-down hierarchical factor model. The non-stationary data as well as non-stationary factors and idiosyncratic errors are allowed, the number of factors of each layer is determined by the data instead of presumption by researchers, and the proposed estimation procedure is implementable by applying principal component analysis from top layer to bottom layer recursively. The corresponding asymptotic properties of the proposed approach are discussed in detailed, and good finite-sample performance is also shown by some Monte Carlo simulations. In essence, the proposed framework is new in the literature and can be a comparable alternative to the existing top-down hierarchical factor models, while facing the possible non-stationary data.

Keywords: common factor, non-stationarity, hierarchical factor model, principal component analysis, variance decomposition.

Experimental Design with Circulant Property and Its Application to fMRI Experiments

Yuan-Lung Lin* (林遠隆) and Frederick Kin Hing Phoa

Institute of Statistical Science, Academia Sinica

Ming-Hung Kao

Arizona State University, Tempe, Arizona, USA

Abstract

Experimental designs have been widely used for cost-efficiency. Orthogonal arrays are commonly used to study the effects of many factors simultaneously, but they do not exist in any sizes. Recently, orthogonal arrays with circulant property receive great attention and are applied to experiments in many fields, such as functional magnetic resonance imaging (fMRI). fMRI is a pioneering technology for studying brain activity in response to mental stimuli. Efficient fMRI experimental designs are important for rendering precise statistical inference on brain functions, but a systematic construction method for this important class of designs does not exist. In this work, we propose an innovative and unified construction method for efficient, if not optimal, fMRI designs via circulant almost orthogonal arrays (CAOAs). Since circulant Hadamard matrices, that can also be viewed as circulant orthogonal arrays of symbols two and strength two, have been conjectured nonexistence, CAOAs are considered. We characterize this new class of efficient designs and propose a systematic construction via a newly invented algebraic tool called complete difference system (CDS). We not only prove the equivalence relation of CDS and CAOA, but also construct many classes of CAOAs with very high efficiency. Finally, we apply these efficient CAOAs to fMRI experiments, demonstrating that our constructed designs have better properties than the traditional designs in terms of cost-efficiency. Abstract begins here.

Keywords: optimal designs, circulant orthogonal arrays, fMRI.

Uncertainty Quantification on Linear-time System

Peter Chang-Yi Weng* (翁章譯) and Frederick Kin Hing Phoa

Institute of Statistical Science, Academia Sinica

Abstract

Uncertainty quantification (UQ) is a new and hot branch of computer experiment that provides a calibration on parameter estimates in a computer model when small perturbations exist. In this paper, we first study the continuous-time linear system described by discretizing the partial differential equations. This type of system has many important applications such as circuits, signal processing, spectroscopy, control theory and many others. In order to understand the random errors, we add some perturbations to the parameters in the system that involve some uncertainties. In order to choose the optimal control to minimize a cost function, we solve the continuous-time Riccati differential equation where the solutions represent the approximations of the corresponding random process. Moreover, we discuss the continuous-time algebraic Riccati equation when we consider the infinite time interval. We provide the sufficient conditions for the existences of the stabilized solutions of the stochastic continuous-time linear system. This study is extended to the discrete-time linear system. Some numerical simulations of the stochastic linear-time systems are presented.

Keywords: uncertainty quantification, Riccati differential equation, Riccati difference equation, stochastic model, optimal control

Two-level Minimum Aberration Designs under a Conditional Model with a Pair of Conditional and Conditioning Factors

Rahul Mukerjee

Indian Institute of Management Calcutta

C. F. Jeff Wu

H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology

Ming-Chung Chang*

Institute of Statistical Science, Academia Sinica

Abstract

Two-level factorial designs are considered under a conditional model with a pair of conditional and conditioning factors. Such a pair can arise in many practical situations. With properly defined main effects and interactions, an appropriate effect hierarchy is introduced under the conditional model. A complementary set theory as well as an efficient computational procedure, supported by a powerful recursion relation, are developed to implement the resulting design strategy, leading to minimum aberration designs. This calls for careful handling of many new and subtle features of the conditional model as compared to the traditional one.

Keywords: bias, Complementary set, effect hierarchy, model robustness, orthogonal array, regular design, universal optimality, wordlength pattern.

D-optimal Design for Quadratic Regression Model without Intercept on q-cube

Wun-Hau Cheng* (鄭文豪) and Fu-Chuen Chang (張福春)

Department of Applied Mathematics, National Sun Yat-sen University

Abstract

We construct a solution of the D-optimal designs for the multivariate quadratic regression without intercept on q-cube. The optimal support consists of all vertices and midpoints of edges on q-cube if $q \geq 2$. Illustrative examples are given and the performance of these designs are compared with D-optimal designs for the multivariate quadratic regression with intercept on q-cube.

Keywords: cube, D-optimal designs, multivariate quadratic, quadratic regression model, without intercept

Identifying Emotional Contagion in Social Networks from Contact Diaries

Ta-Chien Chan* (詹大千)

Research Center for Humanities and Social Sciences, Academia Sinica

Tso-Jung Yen (顏佐榕) and Jing-Shiang Hwang(黃景祥)

Institute of Statistical Science, Academia Sinica

Yang-chih Fu (傅仰止)

Institute of Sociology, Academia Sinica

Abstract

Several studies of emotional contagion through social networks have been reported in the literature, finding that human emotions of both happiness and depression might affect friends of the indexed person. However, the actual transmission process within a social network and how the network's structure affected the contagion have seldom been discussed due to lack of proper data. With contact diary data collected from the online platform ClickDiary (http://cdiary.tw), we tried to examine how an individual's mood might be affected by the degree of social connectedness. The contact network data consist of 133 participants (egos) and 16,139 contacted subjects (alters). The tie between two alters in one personal contact network was determined by whether the pair were familiar with each other, which was confirmed by the ego. Mood scores of the alters during the contacts were also given by the ego. Mixed-effects models were applied to analyze the relationship between an individual's mood score and the scores of surrounding people, while adjusting for contact intensity, different types of contacts, individual characteristics and others. We found that an alter's mood was strongly correlated with the average mood score of degree 1 alters. An alter's mood was also strongly correlated with the average mood score of degree 2 alters, but the effect became half of that in degree 1 separation. The effects were not significant for those alters separated beyond two degrees.

Keywords: contact diary, emotional contagion, social networks, happiness

Analysis of ClickDiary Data: Some Initial Results

Tso-Jung Yen

Institute of Statistical Science, Academia Sinica

Abstract

ClickDiary is an online platform aiming to collect people's contact details and health information on a daily basis. Like other social media, ClickDiary collects data via participants' personal networks. Unlike other social media, ClickDiary traces face-to-face contact, asking participants to identify the type, location, and duration of the contact. It also requires participants to identify personal information of each contacted individual and the relationship between each contacted individual. We analyzed data collected by ClickDiary during a four-and-half-month period from May 1, 2014 to September 15, 2014. The data contain 61,258 contact records generated from 86 independent personal networks with sizes ranging from 3 to 652. We investigated daily contact in egos' personal networks by formulating regression models to identify factors that influence quality of the contact. Results from our analysis show that quality of the contact between an ego and an alter is associated with the alter's position in the ego's personal network. In particular, the more people with whom the alter knows and is familiar in the ego's personal network, the more likely the ego will feel beneficial after contacting the alter (This is a joint work with Ta-Chien Chan, Yang-chih Fu, and Jing-Shiang Hwang).

Keywords: contact diary, egocentric networks, strong ties, weak ties, generalized linear mixed models.

Upward Daily Contact in Social Hierarchy: Benefits of Reaching into Structural Holes in Ego-centered Networks

Thijs Velema* (韋岱思)

Institute of European and American Studies, Academia Sinica

Jing-Shiang Hwang (黃景祥)

Institute of Statistical Science, Academia Sinica

Yang-Chih Fu (傅仰止)

Institute of Sociology, Academia Sinica

Abstract

One central thesis in social network studies argues that the outcome of social networking would be more beneficial (in terms of attaining instrumental goals such as getting jobs, achieving promotions, and increasing salary levels) when an individual (ego) gets connected with someone (alter) who is ranked higher in the social hierarchy. However, recent studies have questioned this thesis by showing that people in higher ranked positions do not necessarily share valuable information and resources in their daily contact with lower ranked others. This opens up the question under what circumstances people are more likely to report benefits from their daily contact with higher ranked alters. We study this question by applying multilevel models to analyze the contact diaries, collected through Click Diary, of 137 diary keepers kept over a period of seven months in 2014. Results indicate that, compared to the contact with lower ranked alters, diary keepers are in general more likely to report benefitting from the contact with alters in similar or higher positions in the social hierarchy. Such benefits of reaching upward, furthermore, are particularly evident when higher ranked alters are familiar with fewer other alters in ego's personal network. The findings indicate that people are especially likely to benefit from their daily contact with higher ranked alters when these alters are surrounded by structural holes and brokerage opportunities.

Keywords: social network analysis, contact diaries, click diary, social structure, brokerage opportunities

Session IV

105 年 6 月 25 (星期六，13:30∼14:40)

IV-1 (SS1001) **Invited Session : Big Data (II)**

IV-2 (SS1003) **Contributed Session : Biostatistics (I)**

IV-3 (SS1004) **Contributed Session : Biostatistics/Biostatistics and Bio-informatic**

IV-4 (SS1005) **Contributed Session : Statistical and Machine Learning**

IV-5 (SS1006) **Contributed Session : Applied Statistics (II)**

IV-6 (SS2002-1) **Contributed Session : Financial/Econometrics (I)**

IV-7 (SS2003-2) **Invited Session : Internetional session (IV)**

IV-8 (SS2005) **Contributed Session : Time Series**

IV-9 (SS2006) **Contributed Session : Young Researcher Session (III)**

IV-10 (SS2012) **Contributed Session : Spatial Statistics/other**

運用網絡輿情分析於校務管理應用

王翠蘭*、葉玲瓏、丘錦發、姜葉飛

輔仁大學商學研究所博士班

摘要

大數據科技發展全面改變了訊息的傳播方式與速度，高等教育領域中的校園事件，常因網路傳播效應引發各式似是而非的議論，造成第一線學務工作者莫大的壓力，值此輿論壓力排山倒海而來之際，學生事務工作者如何拿捏職權分際，維護校園秩序、甚至趁機教育社會大眾，實在需要明確的立場、完備的法規與客觀的案例給予正面的支持；本文遂藉由近期發生之「名導演蔡OO，未申請逕赴輔大賣票宣傳事件」爲例，運用文字探勘技術，分析網路訊息，就媒體報導與網友留言對觀方式，呈現不同觀點的報導或議論脈絡，輔以學務管理立場與法源，支持第一線學生事務工作者之所當爲。

關鍵詞：大數據、高等教育、網路傳播效應、文字探勘、學生事務工作者

應用文字探勘技術預測台灣股價短期漲跌趨勢配合深層學習方法

陳建輝*、張光昭、楊志清

輔仁大學統計資訊學系應用統計碩士班
政治大學統計系

摘要

台灣股票市場一直是國內理財投資的重要標的。如果我們可以事先得知股票的漲跌趨勢，則投資人們就可以從其中獲得利益。一般來說，我們將投資人分爲兩大類，第一類爲大户投資人，其資金雄厚，投資計畫周延。而另一類則爲散户投資人。散户投資人的投資決策取決於大眾間傳播的訊息，且較無計畫性。但是散户們的投資狀況又將會影響個股走向。此研究應用文字探勘技術於股市新聞，計算出其正負面情感，再搭配文本語料庫找出關鍵字詞，後應用 Google 趨勢技術，找出關鍵字詞的關注度，藉以得到投資人關注情況。搭配股市技術指標 (Technology Index)，使用深層學習技術建立預測模型，預測股價漲跌狀況。

研究結果顯示，加入情感指標及關注度指標可以提高約十分的預測正確率。其中對於股市下跌的正確預測率有相當大的幫助。

關鍵詞：投資行爲、股價預測、台股、股票市場、個股趨勢、文字探勘、深層學習

運用 R Shiny 建立健康資訊服務平台

陳俊傑*

輔仁大學統計資訊學系應用統計碩士班

廖珮珊、陳銘芷

輔仁大學商學研究所博士班

摘要

在現今科技日新月異、網路資訊普及化之下，人們對於生活品質也逐漸提升，更講究生活態度。在一個豐衣足食的時代下，現代文明病也悄悄的進入我們的生活。例如：高血壓、高血糖、高血脂等，一些會影響我們身體健康的疾病。目前三高 (高血壓、高血糖及高血脂) 仍然是相當重要的議題，在三高之下很容易衍生出心血管疾病、代謝症候群、中風等相關的衍生疾病，而這些重大的疾病將會威脅到人們寶貴的性命。在現代人生活步調較爲忙碌之下，很容易就疏忽自己的健康，而導致後過不堪設想，讓健康也跟時間一起流失掉。本研究以中高年齡層健康檢查中的資料作爲建模來源，並且利用 R Shiny 建構平台，提供給中高年齡層的人們能夠在忙碌之餘，利用此平台做自我的生理檢測，更能夠防範未然。同時在這資訊時代快速進步中，伴隨而來的就是資料量大量的產生，正是所謂的大數據時代，所以本平台提供幫使用者自行找尋疾病相關文章，並且做立即的資料探勘分析，讓人們能夠省下了解疾病的相關資訊的時間，透過此平台能夠初步了解高血糖、高血壓及高血脂之相關的資訊。

關鍵詞：文字探勘、大數據、資料採礦、健康照護、生理檢測

個人化成本效益分析與效用分配爲雙峰之預測模型選擇

林建甫、吳軍毅*、張詠絜

台北大學統計學系

摘要

在藥物選擇及治療方式的選擇評估決策上，個人化成本效益分析 (Individual Cost-Effectiveness Analysis, ICEA) 的方法逐漸受到重視。個人化成本效益分析中必須同時考慮成本與效益的最適預測模型，主要爲迴歸模型，將反應變數 (成本與效果) 對解釋變數分別建構統計模型，估計不同的治療方式所造成的調整效應 (Adjusted Effect)，然後再透過成本及效用的預測模型進行計算增量成本效益比 (Incremental Cost-Effectiveness Ratio , ICER)。

在外科手術的成本效益分析的資料中，效益反應變數的邊際分配常常呈現爲雙峰且大多集中在變數範圍的兩端。本研究使用模擬產生線性及非線性雙峰邊際分配，考慮四種建構模型方法：一般線性迴歸、測試線性迴歸、拔靴法測試迴歸，次序迴歸模型合併限制立方條樣，透過交叉驗證比較四種建構預測模型對預測模型選擇的影響。

研究結果顯示，雖然效用邊際分配爲雙峰，但並不表示條件分配一定爲雙峰分配，使用在不同的情境下，四種建構預測模型的預測誤差，常受到不同的情境的影響，進而對預測模型選擇的影響。在計算個人化成本效益分析比時，對效益的最適預測模型顯擇，建議針對雙峰邊際分配，使用不同的非線性參數或非參數預測模型，進行敏感度分析，比較不同預測模型對計算增量成本效益比的影響。

關鍵詞：個人化成本效益分析、測試迴歸模型、立方條樣迴歸模型

Joint Modeling of Additive-multiplicative Hazard Model and Longitudinal Data

徐永東

國立中央大學統計學研究所

摘要

以往存活分析中，許多相關研究使用的模型，Cox 比例風險模型被廣泛討論以及應用。然而，當 Cox 模型處理多個共變量時，所有共變量均需滿足比例風險之假設，才能符合模型，一旦出現一個共變量違背時，Cox 模型即無法被使用。本篇建議使用 Cox-Aalen 加乘法模型，此模型可解決以上問題，將那些無法滿足 Cox 比例風險假設的共變量放置於加法部分，得以繼續使用模型，Cox-Aalen 爲可行之模型。本篇將以 Cox-Aalen 模型搭配計數方法，協助模擬研究與愛滋病資料之分析。然而計數方法類似部分概似法，需有完整共變量歷史之要求並且無法容忍測量誤差之影響。因此，本篇統計方法將以聯合模型方法建模，進一步以 EM-演算法做估計。

個人化增量成本效益比預測模型與計算圖表

林建甫、張詠絜*、吳軍毅

國立臺北大學統計學系

摘要

在醫學疾病治療中，每一種疾病都有不同的治療方式. 對每種治療成本考量下，在醫療經濟學發展出成本效益分析 (cost-effectiveness analysis, CEA)，計算不同治療的成本差異與效益差異的比爲增量成本效益比(incremental cost-effectiveness ratio, ICER)。
目前有許多文獻討論個人化增量成本效益比的建模方法，這些方法通常以統計檢定推論爲出發點進行分析。但較少文獻討論以預測觀點爲出發，對個人化增量成本效益比的觀點進行建模與提供個人化增量成本效益比預測值。因此本研究利用老年髖關節骨折的健保資料，探討兩種治療方法，內固定方法與半髖人工關節置換法，建構個人化增量成本效益比預測模型，繪製個人化增量成本效益比預測計算圖表 (predictive nomogram)，提供政府、醫師及病患作最佳的決策。
研究顯示對個人化增量成本效益比，必須同時考慮成本與效益的最適預測模型，且最適預測模型通常是非線性模型，因此使用一般非參數或半參數迴歸模型有助於尋找最適預測模型，但不利於得到簡單計算的預測公式。若輔助個人化預測計算圖表，同時呈現成本與效益的最適預測模型，以及個人化增量成本效益比計算圖表，有助於在緊急接受老年髖關節骨折手術時，將有利於醫師與病患間的溝通，並可以讓病患針對個人的狀況選擇符合自己需求的治療方式。

關鍵詞：成本效果分析、增量成本效益比、預測計算圖表

子宮頸癌患者復發或死亡事件之風險分析

張媛婷*、羅夢娜

國立中山大學應用數學系

蔣安仁

高雄榮民總醫院婦女醫學部

摘要

子宮頸癌爲婦女常見的癌症之一，是目前台灣婦女排名第一的惡性腫瘤。而一般無論是醫生或病人都需要瞭解，在接受手術及治療後，癌症復發或死亡的風險爲何。因此針對病人的個別狀況，做預後的風險評估有其重要性。本研究主要是探討，哪些風險因子會影響子宮頸癌患者的復發或死亡。蒐集腫瘤類型爲鱗狀細胞癌及腺癌的患者，根據其基本資料與病理資料，作爲評估哪些可能爲主要的風險因子，如年齡、是否停經、腫瘤期數、腫瘤大小等。本文可以分成三個部分，第一部分討論鱗狀細胞癌與腺癌的病人，在不同風險因子之下是否有顯著地差異。第二部分則是對不同的風險因子，進行術後復發或死亡相對風險及勝率比的比較。以及在不同風險因子的條件下，鱗狀細胞癌與腺癌患者兩者間條件勝率比的比較。最後則是藉由邏輯斯迴歸模型，討論哪些風險因子會顯著地影響病人預後狀況的評估。

關鍵詞：獨立雙樣本檢定、相對風險、勝率比、邏輯斯迴歸模型

在相關性複製數資料下決定缺失的位置之研究

林駿昇*、馬瀰嘉

成功大學統計學系

摘要

非侵入性胎兒染色體檢測 (NIPT) 是一種 (對孕婦相對安全的) 檢驗胎兒有無先天遺傳疾病的方法。因爲胎兒有一部分的 DNA 小片段會經由臍帶、胎盤，流到孕婦的血液裡。透過抽取孕婦血液取得胎兒游離 DNA，以檢測胎兒是否有先天性的疾病。本研究目的是在第 22 對染色體找到可以偵測 DNA 有缺失的統計方法，作爲醫學上在進行 DNA 有缺失之疾病檢測方法的依據。Olshen & Venkatraman (2004) 提出環狀分段法 (Circular Binary Segmentation, CBS)，將一有序的基因讀數 (reads count) 數據連接成環狀，找出平均數有顯著差異的改變點 (change point)。本研究考慮基因之間有相關的情況下，混合 Hotelling T-squared 檢定的想法，擴展 CBS 的方法至多維度的情形。最後，我們使用統計模擬方法與實例比較兩種方法的型 I 誤發生率 (Type I error rate) 及檢定力 (power)。

關鍵詞：CBS、change point、基因相關性、Hotelling's T-squared test

Effect Aliased Subset (Eas) Criteria for Characterizing Supersaturated Designs

Shyh-Kae Chou* and Frederick Kin Hing Phoa

Institute of Statistical Science, Academia Sinica

Abstract

It is common to characterize supersaturated designs (SSDs) via the $E(S^2)$ criterion for its connections to effect dependency. However, the SSDs with minimal $E(S^2)$ are not unique for most dimensions (number of runs and number of factors). In this work, we propose a new criterion, namely Effect Aliased Subset (EAS) criteria, for characterizing SSDs based on the effect dependency, which is essential for the bias on effect estimates in the analysis. The EAS criteria consist of three different versions, including the worst-case scenario, average situation and counting independency ratio. Depending on the users' requirement, these criteria successfully distinguish SSDs with equivalent $E(S^2)$ values. We further compare the results of our criteria with the traditional resolution and aberration criteria. We provide some design tables for SSDs suggested by our criteria at the end of this talk.

Analyzing Survival Data without Prospective Follow-up

Shih-Wei Chen* (陳世緯) and Chin-Tsang Chiang (江金倉)

Department of Applied Mathematics, National Taiwan University

Abstract

In survival analysis without prospective follow-up, two data structures are considered: (i) data with incident and prevalent covariates and (ii) prevalent cohort data with only covariate and truncation time. Our research aims to identify the effects of covariates on a failure time. To deal with these issues, we employ more general survival regression models and propose the rank correlation estimation methods. Different from our proposal, the existing ones only took into account some particular moment regression models of a failure time and assumed a stationary truncation time. In addition to the advantages of the developed approach, our approach can also adapt to their set-ups based on a more flexible model formulation. Further, the asymptotic properties of the estimators are well-established. Moreover, a series of simulations is conducted to assess the finite sample performance of the estimators. Data from the National Comorbidity Survey Replicate(NCS-R) are also used to illustrate our methodology.

Keywords: asymptotic normality, consistency, covariate-independent truncation, incident cohort sampling, prevalent cohort sampling, rank-correlation estimation, single-index survival model, survival time, truncation time.

變異數成份於長期追蹤重複資料之一致性相關係數估計

潘麗蓉*、蔡秒玉

國立彰化師範大學統計資訊所

摘要

在臨床醫學研究上，評估連續的重複判讀資料之間的一致性 (Concordance correlation coefficient) 一直是被關注的議題，而重複判讀的資料可能是被不同測量方法或不同儀器並隨著不同時間所產生的讀數。本文除了探討被不同測量方法所得的判讀結果間的一致性外，也考慮了在不同時間點所測量的讀數間的一致性。我們使用了三個指標來判讀重複讀數間的一致性，分別爲內部一致性 (Intra-agreement)，其主要評估在同一個測量方法下，且在不同時間點所測量的讀數間的一致性；第二個指標爲方法間的一致性 (Inter-agreement)，其主要是評估不同測量方法所得的眞實讀數間的一致性；第三個指標爲整體方法間的一致性 (Total agreement)，其主要是評估不同測量方法所得的觀察讀數間的一致性。我們利用變異數成份 (variance components) 的方法進行一致性指標的估計及推論，並將此方法應用在兒童近視長期追蹤資料的研究上。

關鍵詞：變異數成份 (variance components) 、agreement、Concordance correlation coefficient

中文母音共振峰之擷取與探討

李愷聲*、王元弘

國立中興大學統計研究所

江翠蓮

中臺科技大學食品科技系

邱國欽

朝陽科技大學財務金融系

摘要

中文是一種聲調語言，不同母音音素會產生不同的共振峰頻率。本篇論文將提出一種演算法對中文母因音素從頻域上擷取第一共振峰和第二共振峰數之頻率位置，觀察在不同母音音素下，其共振峰的分佈情形。演算法如下：第一，先將聲波做快速傅立葉轉換使聲波從時域轉成頻域波形圖；第二，找出頻域波形圖中相對極值的點；第三，如果極值之能量太小，則會被去除；第四，計算相對極值之間的頻率差異並找尋頻率差異數量最多者爲此聲波的基頻；最後利用基頻及高頻能量和低頻能量比來計算共振峰可能最高頻率之位置，並在此範圍中找尋能量最大和第二大的極值頻率，作爲第一共振峰和第二共振峰頻率位置。本研究共19位受試者，共 167200 個語音資料做測試並觀察不同音素之分佈。本論文也進一步討論母音之共振峰軌跡以及母音「 一 」及非「 一 」之辨識。衣服品牌，可看出消費者的偏好方向。

關鍵詞：快速傅立葉轉換、基頻、音框、共振峰

MDS 文字雲：以 PTT 八卦版爲例

侯則瑜

國立東華大學應用數學系統計研究所

摘要

本研究以 PTT 八卦版作爲資料，將文章資料中的字詞區分成「一般常用字詞」、「文章結構字詞」以及「關鍵字詞」，探討不同類型的字詞之間的特性並從文章中萃取關鍵字詞。藉由相關係數收集與關鍵字詞有關的字詞，透過多維標度法 (MDS) 將字詞視覺化於二維平面，產生新的文字雲。這樣的 MDS 文字雲可以區分不同的話題，並保留字詞之間的關聯，有別於一般文字雲基本上爲單一話題一維呈現。

關鍵詞：文字雲、文字探勘、多維標度法

Improving Economic Predictions with Random Forests

Huan-Jui Chang* and Yi-Chi Chen

Department of Economics, National Cheng Kung University

Abstract

There has been growing interest in economic predictions with efficiently using large datasets. The traditional economic literature produces many casual models that are best in the in-sample fitting, but perform poorly at out-of-sample prediction. With the rise of big data, statistical learning algorithms become increasingly popular as a means of generating predictive analytics due to their ability to produce accurate prediction and to identify the key features that lead to such prediction from within large scale datasets. In this paper we demonstrate that the most prominent method, Random Forests, can be a powerful tool to forecast macro-economic aggregates, such as Gross Domestic Product and Consumer Price Index, and to draw particular interest in statistical learning techniques that have rarely been used in economics. We are motivated by the problem of finding practical tools that would be of use to applied econometricians or policy-makers in forecasting macro-economic aggregates with large numbers of economic predictors. Comparisons are made with a benchmark AR(1) model and an ad hoc linear model built on the most important variables suggested by Random Forests. Evaluation by forecasting shows that the Random Forests makes the most accurate forecast supporting the theory that there are benefits to using Random Forests on economic time series. In order to avoid using statistical learning as a black-box tool, we consider it as a methodology with an emphasis on its practical application for exploratory analysis and economic interpretation. This paper discusses these results and the ways in which algorithmic statistical methods like Random Forests can be useful to more accurately predict economic time-series.

Keywords: statistical learning, random forests, GDP forecast, feature selection

利用 k 最近鄰居法於不特定語者之中文單音辨識及錯誤探討

陳冠逢*、陳貽靖

國立中興大學統計研究所

江翠蓮

中臺科技大學食品科技系

李宗寶

中興大學應用數學系

摘要

本論文主要是探討不特定語者對於 1391 個中文單音、母音與子音之辨識。辨識流程主要分成三部分：首先將錄製好的語音資料進行前處理，如端點偵測、切割音框、預強調、視窗化等；接著利用梅爾頻率倒頻譜係數 (MFCC) 求取特徵值；再進一步利用 K-means 分群演算法，對訓練語音進行分群，最後以 K 最近鄰居法建構語音模型並進行比對，觀察給定不同的實驗因子，如「訓練語音分群數」、「子母音權重」等參數組合，從中選出最佳辨識組合。本次實驗的語音資料庫是由二十位不同語者所錄製的 1391 個所有中文單字進行辨識。實驗結果發現在訓練音不分群時，當子、母音權重同爲 0.5 時，單音辨識率最高爲 76.3%，母音辨識率最高爲 83.7%；然而對訓練語音進行分群，單音辨識率最高爲 76.2%，母音辨識率最高爲 83%。本論文也進行不分聲調之母音辨識，其平均辨識率爲 87.15%，有關錯誤原因也被探討。

關鍵詞：K 最近鄰居法、梅爾頻率倒頻譜係數、K-means分群演算法

Economic Design of Two Stage Control Charts with Bivariate Skew Normal Measurements

Chun-You Lu* (呂俊佑) and Nan-Cheng Su (蘇南誠)

Department of Statistics, National Taipei University

Abstract

In many instances, the cost is high to monitor primary quality characteristics called performance variable, but it could be economical to monitor its surrogate. To cover asymmetric processes for two-stage charting methods using both performance and surrogate variables, bivariate skew normal distribution is considered as the underlying distribution of process variables. When the correlation relationship between the performance variable and its surrogate is specified, a charting procedure to monitor either the performance variable or its surrogate in an alternating fashion rather than monitoring the performance variable alone is proposed. The proposed two-stage control charts are constructed under an economic design using Markov chain approach. Its advantages over the existing methods will be presented.

Keywords: performance variable, surrogate variable, skew-normal distribution, two-stage charting, economic design, Markov chain approach.

二維市場調查與模糊統計分析及在衣服品牌與消費者行爲之應用

楊顗錞*、吳柏林

國立政治大學應用數學系

摘要

研究動機與目的：由於市場調查與消費者行爲分析，消費者的感知是主要研究的對象，使用傳統問方法設計的問卷，往往落入二元邏輯的迴圈，因而無法貼近填答者眞正的意向。近幾年，全球各大流行衣服品牌大舉進駐台灣設店，因而以衣服品牌與消費者行爲的因素，作實證研究分析，得出更合理的分析，嘗試進行較貼近消費者感受之消費行爲分析。

研究方法：使用模糊層級分析法 (fuzzy AHP) 與模糊類別資料之卡方齊一性檢定。

創新與推廣：傳統的模糊層級分析 (AHP) 較爲理想化，模糊層級分析在分析中，不僅可以解決德菲法的冗長過程，可以更進一步的解決層級分析中因子權重計算困難的問題，並以卡方齊一性檢定來檢驗兩個或兩個以上的母體，各類別的比例是否齊一之統計方法。

結論：在模糊問卷情況下，利用二維模糊數矩陣發現消費者的行爲的相關因子會影響所選擇的衣服品牌，可看出消費者的偏好方向。

關鍵詞：二維市場調查、模糊統計分析、衣服品牌

偏斜分佈之分層模型研究

張漢揚*、蘇南誠

國立台北大學統計學系

摘要

我們探討來自發育神經毒的偏斜常態資料之劑量反應模型，在分層的結構下推導出反應的非條件分配，並且探討來自同一個窩的樣本的相關性，最後利用最大概似估計法去估計模型的參數，並觀察相關係數越高會對這個模型的影響。

關鍵詞：偏斜常態、分層模型、劑量反應、組內相關性

飯店需求不確定性之模糊決策分析

郭晏辰*、吳柏林

國立政治大學應用數學系

摘要

研究動機與目的：近年來兩岸的開放與國家發展、休閒觀念的改變，使得觀光產業發展逐年增長。觀光收入也成爲賺取外匯的首要來源。而光觀經濟中，飯店因此推測飯店未來的供給需求，尤其在需求方面的不確定性對經營決策者是很重要的資訊。
研究方法：影響住房率的因素，很直觀的就是來台旅遊的人口多寡，因此預測觀光人口的增長爲首要重點。現今的預測的方法都是以量化方法與質化方法爲主，本研究擬以時間序列模型預測出未來來台觀光人口的增長。再將預測出的觀光人口與住房率取比例後，依照比例將數據模糊化後，代入預測的觀光人口數，得到數據後運用模糊分析來達到判斷飯店是否有成長空間。
創新與推廣：過去相關文獻數據大多都是以定值來表示，但數據與實際上往往存在著誤差，結論常常需要再做平移、調整。本研究以模糊化分析，結果更加的直觀，如果將模糊化方法更廣泛運用的話，可以將誤差值的影響降到最低，相信這樣會使數據更加的貼近現實情況。
結論：台灣觀光業發展的歷程與台灣經濟成長、政府政策的推廣、來台旅客成長及市場競爭等因素，有著相當重要關聯。以來台旅客成長的因素來看，將刺激市場供給的增加。但單純的預測來台觀光人數並非爲影響旅館住房率的唯一因素，本研究認爲要更精確的預測，需要考慮到國內旅遊的投宿情況，民宿、汽車旅館與非登記之套房等許多因素。因此飯店的住房率、觀光人口的預測都是需要考慮到很多方面的因素，但還是有很多因素都是無法預測造成預測的不準確，例如陸客來台觀光簽證的限縮造成的影響。由本研究可得知利用模糊化分析的結果的確可以更加接近實際的結果。

關鍵詞：飯店需求、模糊化分析、時間序列

齊一性價格拍賣配給探討—IPO 實例

姜堯民、鍾麗英、陳韋迪*

國立台北大學統計學系

摘要

在本篇論文中，沿用了 Back and Zender (1993) 的齊一性價格拍賣的需求函數，利用此需求函數進一步的推導出均衡價格，並引進配給 (Rationing) 的想法，探討在以下三種情況的均衡價格與供給量：(1) 當投標人知道賣方可能有截標的情形，投標人是否會把出價提高，(2) 當投標人心中認定此標的物價格比市場上還高時，投標人是否出價不變，(3) 當投標人心中認定此標的物價格比市場上還高時，是否使得投標人出價提高。在以上每一種情況下以賣家收入最大化爲目標推導出其需求函數，並探討在何種情況時的賣家收入會最大化。

此外，我們使用眞實的 IPO 市場資料代進所推導的模型以探討模型的正確性，最後針對齊一性價格拍賣的各種配給方式進行比較，藉此可幫助賣家選擇合適的配給方式。

關鍵詞：齊一性價格拍賣、配給、收入最大化、IPO 市場

企業社會責任與資本結構調整行爲之探討

陳文松*、顏汝芳

國立台北大學統計學系

摘要

本研究探討企業社會責任 (Corporate Social Responsibility, CSR) 對於公司資本結構調整行爲的影響。過去的文章指出公司的財務槓桿與 CSR 活動績效成反比關係，當低槓桿公司擁有高獲利時，公司將擁有更多的資源於未來投入於 CSR 活動中。然而，從事 CSR 究竟是否會影響資本結構調整行爲？是否會削減傳統抵換理論的執行力？文獻尚未有明確的答案。因此，本研究探討 CSR 及 CSR 的七大指標 (包含公司治理、社區、環境、多樣性、員工關係、人權及產品) 與公司資本結構調整之關聯性，並分析其對於調整速度的影響。

本研究使用一般線性迴歸模型，分析結果發現：首先，整體而言，對於過度槓桿 (不足槓桿) 的公司，從事 CSR 活動確實會加速 (減速) 資本結構調整至目標槓桿的調整速度；第二，正面的 CSR 細項活動對於過度槓桿 (不足槓桿) 的公司會加速 (減速) 資本結構之調整速度，表示過度槓桿公司擁有更多能力調整公司槓桿至最適值以及增加公司價值，不足槓桿公司則是偏好保持低槓桿，保留更多的公司資源於未來投入 CSR 活動；第三，負面的 CSR 細項活動對於公司會減速資本結構的調整速度，說明了具有風險性的公司通常傾向過度槓桿，負面的 CSR 活動致使公司能力有限無法加速調整至公司的目標槓桿。本篇研究認爲，公司經理人決定融資政策時，會考量到企業社會責任精神的落實，以提升公司正面的形象。

關鍵詞：資本結構調整行爲、企業社會責任、線性迴歸模型

Credit Risk Illustrated under Coupled Diffusion

Po-Heng Kuo

國立中央大學統計研究所

Abstract

We introduce a model to analyze credit risk where the log-monetary reserves are driven by the coupled diffusions. The default is described as the assets of firm less than the liabilities in the maturity time T. Due to the special structure of assets of banks, we apply the Naïve alternative method in order to solve the high leverage problem. Compared to KMV-Merton model, the joint default probability given by the coupled diffusions is seen as a rare event treated as systemic risk. Finally, the empirical study is also discussed using Maximum Likelihood technique.

Keywords: default, credit risk, systemic risk, KMV-Merton model, coupled diffusion model.

均一性價格拍賣配給探討：公債拍賣實例

姜堯民、鍾麗英、鍾佳軒*

國立台北大學統計學系

摘要

在本篇論文中，延伸 Back and Zender (1993)、(2001) 針對均一性價格拍賣的賣家收益最大化研究，我們加入配給的概念，假設賣方收到投標者投標的需求表之後能夠改變他的供給數量，減少投標者原本能得到的數量，期望投標者能夠提高出價，使最後賣方收入能夠大於原先的收入。

我們使用公債拍賣的眞實資料，並引用 Back and Zender (1993) 均一性價格拍賣的需求函數，在加入配給後，需求函數會有所改變，之後利用總供給等於總需求的公式算出個別的均衡價格 ($p*$)，並分別做比較，找出賣方的最大總收益。

關鍵詞：均一價格拍賣、歧視性價格拍賣、配給、均衡價格

A Copula-based Dynamic Prediction of Death According to Tumour Progression and High-dimensional Genetic Factors: Joint Cox Proportional Hazards Models for Meta-analytic Data

Takeshi Emura*

Graduate Institute of Statistics, National Central University

Masahiro Nakatochi

Center for Advanced Medicine and Clinical Research, Nagoya University, Japan

Hirofumi Michimae

School of Pharmacy, Department of Clinical Medicine (Biostatistics), Kitasato University, Japan

Virginie Rondeau

INSERM (Biostatistic), Université de Bordeaux, France

Abstract

The availability of genomic information and large-scale meta-analytic data for clinicians have motivated the extension of the traditional prediction scheme based on the Cox proportional hazards model. The aim of our paper to develop a copula-based risk prediction scheme for death according to patients' genetic factors and dynamic tumour progression histories based on meta-analytic data. To this end, we extend the existing joint Cox proportional hazards models (Rondeau et al. 2015; Emura et al., 2015) to a model allowing for high-dimensional genetic factors. Here we utilize Tukey's compound covariate predictors to reduce the dimension of genetic factors. In addition, we propose a dynamic prediction scheme to predict death given the patient's tumour progression histories over time. We also develop a tool to validate the performance of the prediction scheme by assessing the prediction error. We illustrate the method with the meta-analysis of individual patient data on ovarian cancer patients.

References

[1] Rondeau V, Pignon JP, Michiels S. A joint model for dependence between clustered times to tumour progression and deaths: A meta-analysis of chemotherapy in head and neck cancer. Statistical Method in Medical Research 2015; 24(6): 711-729

[2] Emura T, Nakatochi M, Murotani K, Rondeau V, A joint frailty-copula model between tumour progression and death for meta-analysis, Statistical Methods in Medical Research 2015, doi: 10.1177/0962280215604510

Keywords: copula, dependent censoring, semicompeting risk, survival analysis.

Optimal and Efficient Designs for Functional Brain Imaging Experiments

Ming-Hung Kao

School of Mathematical and Statistical Sciences, Arizona State Universit

Abstract

This article concerns optimal experimental designs for neuroimaging studies in which the pioneering functional magnetic resonance imaging (fMRI) technology is used to investigate functions of the human brain. We develop analytical results for identifying designs that allow precise estimates of the contrasts between hemodynamic response functions (HRFs), each describing the effect of a mental stimulus over time. We also provide systematic methods for constructing optimal and very efficient designs, and show that fMRI designs obtained by relabeling the symbols of m-sequences can be very efficient in comparing the HRFs. This is a joint work with Ching-Shui Cheng and Federick Kin Hing Phoa from the Institute of Statistical Science, Academia Sinica.

Keywords: circulant orthogonal array, De Bruijn sequence, Hadamard sequence, Schur optimality, type I criterion.

股票價格趨勢預測之研究

林碁域*、陳宜伶

高雄大學亞太工商管理學系碩專班

摘要

股票市場是一複雜多變的動態社會體系，其價格走勢直接影響著投資者的利益， 也影響和反映著國家的經濟政策，因而受到投資者們的廣泛關注。由於股價變化 無常，投資者必然會遇到投資的風險，因此投資時的規劃與預測之分析是非常重 要的。本研究將採用時間序列 (Time series)，來預測台灣上市公司之股價及走勢，以三家上市公司之股票爲研究對象， 比較不同類股之預測準確率，並模擬交易。結果顯示時間序列在預測當日股價下跌 的趨勢有其準確性，預測準確率則是股價波動較小之股票爲佳，模擬交易的部分， 其中一檔股票獲利績效良好。此預測準確之結果與模擬之績效，時間序列模式將可提供一般投資者、研究者，或公司決策者。

關鍵詞：股票、股價、預測、時間序列

原油價格預測—時間數列模型之比較分析

洪毓婷*、許玉雪

國立台北大學統計系

摘要

近期原油價格的波動起伏不定，時而增漲時而下跌，然而油價又與我們的生活息息相關，油價的變化往往也牽動著物價的改變，進一步影響著我們日常的大小所需，因此事先了解油價的趨勢，對於政策的研擬也能有所幫助。本文研究目的在於比較分析幾個時間數列模型，試圖找出一個用來預測原油價格的較佳模式，同時進行未來幾年的油價預測。而本文的研究方法，首先，整合過去國內外關於原油價格預測的方法，包含 ARIMA 模式、ARCH 模式、GARCH 家族模式和 Markov Switching 等，利用近幾年的原油價格資料，進行實證分析，比較分析各模式的預測能力，提出一個較佳的預測模式，作爲後續預測的基礎。

關鍵詞：原油價格、預測模式、價格波動、時間數列模式

基於蒙地卡羅法預測共整合配對交易之預測獲利

廖俊豪

國立成功大學統計學研究所

摘要

本研究主要目的爲預測共整合配對交易之獲利，用來提供交易者選擇配對之參考。在本研究中，選取歷史資料當成訓練集，分別根據以殘差爲基礎之共整合模擬過程 (Lin 與 McCrae., 1999) 與誤差修正模型兩種方法，來建立具有共整合關係股票價格的模型，以此模型生成未來的股票進行配對交易，借此預測獲利。實證資料使用納斯達克 100 指數成分股，使用上述兩種方法對所有組合股票價格的共整合向量及程序之下的統計模型參數來估計。並用來重複模擬股票未來價格執行配對交易，計算模擬交易平均獲利。結果顯示，以殘差爲基礎之共整合模擬過程的模擬平均獲利無法有效解釋實證獲利。而在誤差修正模型模擬下完整交易平均獲利與實證完整獲利呈現正相關的線性趨勢，在線性迴歸模型下模型解釋能力達約 80%，平均來說模擬獲利高估實證獲利 1.4 倍。

關鍵詞：配對交易、共整合、向量誤差修正模型

利用整數值時間序列模型監控登革熱爆發時間點

林良靖、陳孟翔*

成功大學統計研究所

摘要

登革熱已成爲全球性的問題，而台灣近三年夏季的登革熱病例人數也逐年上升。衛生局訂定疫情失控的標準爲一個縣市單日病例數達 30 人，而本研究目的是建立新的監控標準，期望能更早監測出登革熱疫情失控的時間點。本研究使用差分後的日資料建立整數值時間序列模型，接著利用蒙地卡羅馬可夫鏈方法估計參數後建立管制圖，再以西方電氣法則來監控失控的時間點。實證上以高雄 2012 年登革熱每日發病人數爲基礎，建立管制圖來監控高雄 2013 到 2015 和台南 2014 到 2015，共五個年度的登革熱疫情。其中，高雄 2013 年和台南 2014 年屬於疫情穩定，本研究建立的管制圖與衛生局監控相同，並沒有出現過度反應的情況。而高雄 2014、2015 年與台南 2015 年爲疫情失控的年度。與衛生局的監控標準相比，本研究建立的管制圖至少能提早兩週以上監測出登革熱疫情失控。

關鍵詞：整數值時間序列、蒙地卡羅馬可夫鏈、登革熱

On the Basis Functions for Fixed Rank Kriging with Irregularly Spaced Data

ShengLi Tzeng* (曾聖澧)

China Medical University

Hsin-Cheng Huang (黃信誠)

Institute of Statistical Science, Academia Sinica

Abstract

Fixed rank kriging is a flexible and computationally efficient spatial prediction method based on the spatial random-effects models. It allows one to construct a non-stationary and/or nonparametric spatial covariance. Its success, however, depends on the selection of the basis function family and the number of basis functions. Commonly used basis functions encounter the difficulty in determining the partitions of spatial regions, centers of knot, or number of resolutions, etc. In this talk, we propose an alternative family of basis functions via an optimal linear combination of thin-plate splines, which are in a descending order of their degrees of smoothness. In this family, a smoother function corresponds to larger-scale features and a more fluctuating one corresponds to smaller-scale details. The proposed family of basis functions has several advantages over commonly used ones. First, we do not need to concern about the allocation of the basis functions, but simply select the total number of functions corresponding to a resolution. Second, only a small to moderate number of basis functions is usually required, for capturing either a stationary or a non-stationary spatial covariance structure, which facilitates computation. As a side benefit, estimation variability of model parameters can be considerably reduced, and hence more precise covariance function estimates can be obtained. Third, the proposed basis functions depend only on the data locations but not the measurements taken at those locations, and are applicable regardless of whether the data locations are sparse or irregularly spaced. Some numerical examples are provided to demonstrate the effectiveness of the proposed method.

Keywords: fixed-rank kriging, non-stationary spatial covariance, radial basis functions

Use of Bayesian Approach for the Design and Evaluation of Multiregional Clinical Trials

Yu-Chieh Cheng* (鄭宇傑)

Institute of Population Health, Sciences, National Health Research Institutes
Institute of Statistics, Nation Chiao Tung University

Hsiuying Wang (王秀瑛)

Institute of Statistics, Nation Chiao Tung University

Chin-Fu Hsiao (蕭金福)

Institute of Population Health, Sciences, National Health Research Institutes

Abstract

To speed up drug development to allow faster access to medicines for patients globally, conducting multi-regional trials incorporating subjects from many countries around the world under the same protocol may be desired. Several statistical methods have been purposed for the design and evaluation of multi-regional trials. However, in most of recent approaches for sample size determination in multi-regional clinical trials, a common treatment effect of the primary endpoint across regions is usually assumed. In practice, it might be expected that there is a difference in treatment effect due to regional difference (e.g., ethnic difference). Therefore, we propose to use a Bayesian approach to design and evaluation of MRCTs under normality. We assume that there is a difference in treatment effect due to regional difference but the population variance is assumed to be equal and known. More specifically, two different types of priors, inverse gamma distribution and half normal distribution, are used to address the heterogeneous treatment effect across regions. Methods for sample size determination for the multi-regional clinical trial are also proposed.

Keywords: multi-regional clinical trial, ethnic difference, Bayesian approach, prior

Copy Number Identification in Clonal Cell Populations Using Spatially Correlated Two-way Poisson Mixture Models

戴安順*、謝文萍

國立清華大學統計研究所

彭千華、彭士齊

長庚醫院醫學研究部

Abstract

Tumor heterogeneity describes the situation that different cells in a tumor consist of different DNA profiles and it makes cancer a complex disease. The presence of intra-tumor heterogeneity brings a difficult problem of biomarker identification. The measurements of the biomarker come from different cell populations and those populations have their own specific DNA variants. Hence, the information is mixed and should be adjusted before any genomic analysis.

The identification of copy number aberration also faces this challenge from sequencing data. If we can efficiently decompose the mixture structure of copy number states and divide the cell population into several subsets, we can not only get more accurate copy number information but also recognize the architecture in clonal cell populations.

We developed a spatially correlated two-way mixture model to detect the copy number status of different populations. This model is based on the traditional Poisson mixture model and regulates the mean structure such that it can include the information of copy number and cell population architecture in the same model. We also combine the spatial information of genome into our model and relax the independence assumption of the traditional mixture model. Finally, we applied our model to TCGA data and identified the copy number structure.

Keywords: tumor heterogeneity、copy number aberration、Poisson mixture model

Risk, Predictive Direction, and Control Dosage ($NP \neq P$)

Hsieh Hsih-chia*

Hsing-Kuo University

Pei-Gin Hsieh

National Chung-cheng University

Abstract

The target is to find all solutions for control and prediction Method:Non-deterministic polynomials (NP) compute the dynamic equilibrium $(x^*, r^*, Y^*, \lambda^*)$ which is a upper and lower boundary of smooth regularity pathways for the local and global predictions or attractors. Result: The Pareto-optimal growth or profit, $n - 1 <= x^*/r^* < n$, has a upper and lower boundary of regulation, and is a unique unbiased consistent test statistic, which always exists, detecting the location of risk σ_x^2 and a saturating probability $p(x) = 1$.
Conclusion: Below or beyond the boundary of dynamic equilibrium, the direction angle, λ, is sign-changing for optimization, image-mapping, and Hamilton cycles, reducing the risk σ_x^2. Errors v or mutants moves in a wrongdirection.

2000 AMS Mathematics Subject Classification. Primary: 62G05;
Secondary: 62G30; 39B82, 49K40, 35B37.
JEL. A1, B1, C1

Keywords: convergence probability, local and global general equilibrium, nondeterministic polynomials ($NP \neq P$), conflict theories.

探討參數化區間型象徵性資料的最大概似估計法

吳欣展*、黃怡婷、汪群超

國立臺北大學統計學系

摘要

在統計科學與資訊科技蓬勃發展的二十一世紀，全球資料總量以倍數成長，說明了大數據時代的來臨。爲處理大量數據與整合數據，Diday (1987) 提出象徵性資料分析 (Symbolic Data Analysis)，其資料特徵爲每個觀察值表示一個類別或群體，稱爲象徵性觀察值 (Symbolic observation) 或概念 (Concept)。因爲觀察的對象包含複數個觀察個體，所以象徵性資料中變數所包含的是一個較複雜的資料結構。Rademacher 與 Billard (2011) 使用參數化方式討論區間型象徵性資料與直方圖型象徵性資料的最大概似估計方法，但其參數化設定是直接假設區間型象徵性資料服從母數族，而非推論原始資料的參數。假設原始資料服從特定分配，本論文推導出傳統資料轉變成象徵性資料型式的參數分配，利用最大概似估計法來推估原始分配的參數，最後以蒙地卡羅模擬討論不同參數分配及不同轉換方式參數估計的表現。

關鍵詞：象徵性資料、最大概似估計法

On the Cluster Detection and Methods Comparison for Spatial Data

Zhixin Lun* (倫智欣) and May-Ru Chen (陳美如)

Department of Applied Mathematics, Nation Sun Yat-sen University

Abstract

Cluster detection is one of the most important topics in spatial statistics. With increasing public health concerns about environmental risks, the development of statistical methods for analyzing spatial health events becomes immediate. Firstly, we introduce the two most popular cluster detection approaches which are the spatial scan statistics and spatial autocorrelation. Bayesian hierarchical structure, which is a modern method to fit spatial data, is illustrated by using Poisson log-linear conditional-autoregressive model. Hierarchical clustering algorithm is applied for comparison since it is one of current statistical learning method to detect cluster. By comparing these methods above, we summarize their advantages and drawbacks on cluster detection for spatial data. Finally, we analyze the Dengue Fever breakout over south Taiwan in last year which as an empirical study.

Keywords: cluster detection, spatial scan statistics, spatial autocorrelation, local Moran's I, poisson log-linear CAR model, hierarchical clustering.

Variable Selection and Spatial Prediction Under a Misspecified

Chao-Sheng Chen* (陳朝聖) and Chun-Shu Chen (陳春樹)

Institute of Statistics and Information Science, National Changhua University of Education

Abstract

In geostatistics, spatial prediction and variable selection both are important issues. If spatially varying means exist among different subareas, globally fitting a spatial regression model for observations over the study area may be not suitable. To alleviate deviations from spatial model assumptions, we propose a methodology to locally select variables for each subarea based on a locally empirical conditional Akaike information criterion, where the global spatial dependence of observations is considered and the local characteristics of each subarea are also identified. It results in a composite spatial predictor which provides a more accurate spatial prediction for the response variables of interest in terms of the mean squared prediction errors. Further, the corresponding prediction variance is also evaluated based on a resampling method. Statistical inferences of the proposed methodology are justified both theoretically and numerically. Finally, an application of a mercury data set for lakes in Maine is analyzed for illustration.

Keywords: geostatistics, information criterion, prediction variance, resampling, squared prediction error

模糊績效評估應用在雙北市交通施政滿意度

黃宇君*、吳柏林

國立政治大學應用數學研究所

摘要

動機與目的：本研究室以政策評估中的執行評估作爲切入點，政府的施政作爲能對於環境的變遷要能夠及時的回應及富有彈性，政府的施政計畫要廣泛並且深遠，其公共服務的提供要符合人民對環境急遽變化的需求，尤其是在交通方面，人民有深刻的感受，就像近期的三橫三縱、三環三線都造成許多廣大的民怨，讓他們的施政滿意度受到負面的影響，因此本研究由其著重在交通政策的績效評估。
研究方法：以模糊層級分析法 (FAHP) 與模糊德非法 (Fuzzy Delphi) 對各個層次之決策逐層評估重要性，分析結果揭露不同評估要素相對模糊權重，判斷其優先順序。
創新與推廣：釐清政府運用由企業發展出來的模糊層級分析法與模糊績效評估法可能產生的問題。希望政府必需建?有效評估政府交通施政成績效結果，對施政的總體成果進行評估。
結論：將績效評估結果的資訊公開的呈獻給社會大眾，且是其績效評估符合人民的期待且具有公信力，並應用在雙北市交通施政的情況對照人民滿意度來對應模糊績效評估。

關鍵詞：模糊績效評估、模糊德非法、模糊統計、模糊層級分析法

Session V

105 年 6 月 25 (星期六，15:00∼16:10)

V-1 (SS1001) **Invited Session : Big Data (III)**

V-2 (SS1003) **Contributed Session : Biostatistics (II)**

V-3 (SS1004) **Contributed Session : Biostatistics (III)**

V-4 (SS1005) **Contributed Session : Statistical Computing/Other**

V-5 (SS1006) **Contributed Session : Applied Statistics (III)**

V-6 (SS2002-1) **Contributed Session : Mathematical Statistics/ High-Dimensional Analysis**

V-7 (SS2003-2) **Invited Session : Internetional session (V)**

V-8 (SS2005) **Contributed Session : Financial/Econometrics (II)**

V-9 (SS2006) **Contributed Session : Financial**

V-10 (SS2012) **Contributed Session : Statistical Quality Control/ Industrial Statistics**

用 R 建立文字探勘平台應用於電視收視率預測

謝元晟*、張光昭、程美華

輔仁大學管理學院統計資訊學系應用統計所
東森 ETTODAT 新聞台協理

摘要

目前全球進入了大數據 (Big Data) 的時代，也是從資訊科技 (IT) 時代走向資料科技 (DT) 時代之際，各式各樣的儀器與技術使得人們更容易地進行資料的蒐集與累積。在大量資料中存在著結構化資料與非結構資料，一般常用來分析的都是結構化資料，然而更多的是非結構化資料，如微博、LINE、臉書、Twitter 推特、網誌、部落格等，所獲得的資料都是文字的資料。透過文字採礦技術從大量的文本中擷取傳統方法無法取得的隱含知識，以簡要的格式呈現資訊給使用者。

本研究利用 R 建構社群媒體服務平台。爬取社群網路的資料，予以匯入平台中，設計動態的互動式介面可以對資料進行篩選讓使用者快速得到想要了解的資訊，並使用收視率與社群媒體資料，結合傳統的統計方法以及文字探勘技術，形成新的收視率測量方法，並整合爬文、文字探勘、社群媒體討論熱度、收視率預測等功能，成爲一套收視率分析系統，提供「即時性的收視率預測」。

關鍵詞：文字探勘、輿情分析、資料視覺化、資料採礦、收視率

氣候及環境因子對重大心血管不良事件預測及相關因素之探討

萬擎*、廖佩珊、鄭宇庭

輔仁大學統計資訊學系應用統計碩士班
政治大學統計系

摘要

近年人們生活型態的轉變，國人健康狀況普遍不佳，造就許多疾病的產生，其中心血管疾病更是占台灣十大死因的大宗，然而也因爲氣候變遷與環境污染越來越嚴重，空氣汙染與氣候異常與心血管疾病如缺血性心臟病、腦血管心臟病等息息相關。本論文研究重大心血管不良事件之影響因素及預測模型，以供日後醫師與學者做爲參考。本研究應用 Pearson 相關分析討論氣象與環境因子之間互相的關聯性，並利用羅吉斯迴歸找出顯著的變數。另外建立資料採礦方法如隨機森林、支援向量、決策樹、類神經網路等探討不同時間點的環境氣象因子對發生重大心血管不良事件預測模型。實證發現，以羅吉斯迴歸找出影響重大心血管不良事件的顯著變數有二氧化氮、PM2.5、風速、能見度、溫差等。另外建立預測模型發現在 1 : 1 的資料集爲最好的預測模式，且最好的預測模型爲隨機森林與決策樹模型。

關鍵詞：全民健康保險研究資料庫、資料採礦、心血管疾病、氣候氣象、空氣汙染

Python 之文字探勘平臺

林柏宇*、廖佩珊、江志民

輔仁大學統計資訊學系應用統計碩士班
經國管理學院

摘要

隨著資訊科技的發展及手持裝置與社群網站越來越趨於活絡，各種電子新聞、社群網站的貼文與評論的資料量快速成長且結構複雜。一般而言，資料可簡單的分成結構化資料與非結構化資料，結構化的資料已有許多有效的方法可以運用，像是資料採礦技術，但如文字、聲音、影像等非結構化資料的分析方法，相較之下較爲少數，運用本研究的文字探勘平台，挖掘出有效的資訊，將可以快速的從資料中探討其重要意義。
本研究希望透過網路上的開源碼整合出一套平台，利用 Python 做爲後台運算，結合 HTML 撰寫網頁程式，把文字探勘的平台架在 Django 上。再將夏季旅展的新聞資料匯入平台，做文字探勘相關的分析，如詞雲分析、關聯分析、集群分析、情感分析等，討論夏季旅展資料的意義與脈絡。

關鍵詞：文字探勘、大數據、情感分析、資料採礦

Survival Analysis and Finding Cut-off Points in a Cervical Cancer Study

Wen-Yi Chang* (張文怡) and Chung Chang (張中)

Department of Applied Mathematics, National Sun Yat-sen University

Abstract

Cervical cancer is one of the most common genital tumors and ranks among the top ten leading cause of death from gynecologic cancers. We discuss how risk factors influence recurrence and death of cervical cancer. In the first part, we used Cox proportional hazards model to find out which clinicopathologic factors posed a significant risk for recurrence or death. The effect of some of the risk factors changes with time. So, we built a flexible Cox model that allowed time-varying effects, and variables with time-varying effects and those without could thus be analyzed side by side for their impact on recurrence and death. In The second part, we analyzed the correlation among risk factors. It is often necessary to determine a cutoff point to stratify patients into groups when using continuous variables in clinical studies. We provided a method of how to choose cut-off points. Given the number of cutoff points and the data type, our method offered a criterion to decide cut-off point, which may be helpful to clinical assessment and decision-making.

Keywords: cervical cancer, Cox proportional hazards model, time-varying coefficient, cut-off point

用非線性混和效應模型分析 CA125 資料

陳吉重*、張中

國立中山大學應用數學系

摘要

傳統長期追蹤資料經常使用非線性混和效應模型進行配適，且其混和效應與殘差通常都假設常態分配，但如此可能與實際資料情況不符。Lu et al. (2014) 將偏斜因子加入殘差進行配適，並且比較不同殘差假設下之結果。在本文中長期追蹤資料其混和效應或殘差有可能存在偏斜，因此我們在貝氏階層模型的架構下利用非線性混和效應模型來分析病人之 CA125 資料，考慮其混和效應及殘差存在偏斜因子，並加入對混和效應與殘差有影響之變數進行分析與討論。

關鍵詞：長期追蹤資料、非線性混和效應模型、貝氏階層模型

The Estimation of Generalized Diversity Index for Geographic Information Data

Yu-Yi Pan* (潘宥亦) and Mi-Chia Ma (馬瀰嘉)

Department of Statistics, National Cheng Kung University

Abstract

In many areas of imaging science, it is difficult to determine the phase of linear measurements. For example, in the X-ray imaging, the detector can only measure the Fourier magnitude of the received optical wave. In this setting, the goal of the phase retrieval is to reconstruct the unknown image from its Fourier magnitude data. Due to the absence of the phase information, the phase retrieval does not have a unique solution. In 2012, Albert Fannjiang proved that randomly modifying the phases of the unknown image by a mask can lead to a unique solution up to a global phase factor. Apart from considering the uniqueness of the phase retrieval problem, the simulation results show that random illuminations also significantly improve the numerical performance of the Error-Reduction (ER) algorithm, which is the most popular phase-retrieval algorithm. This talk describes the mathematical formulation of the phase retrieval problem, why the random illumination can improve the performance of the ER algorithm, and what breakthroughs about the convergence of the ER algorithm have been made compared with related works.

Keywords: geographically information system, Simpson's diversity index, Shannon's diversity index.

利用經驗貝氏方法估計錯誤發現率

鄭暘諭*、馬瀰嘉

國立成功大學統計學系

摘要

在多重檢定中，若不調整個別檢定之顯著水準，則整體犯錯率就會膨脹。所以當我們同時進行多個假設檢定時，重要的是如何控制型I錯誤發生率。常用的控制方法有控制整體型 I 誤發生率 (familywise error rate; FWER) 和錯誤發生率 (false discovery rate; FDR)，但無論何者，都要針對虛無假設爲眞的個數給一較精確的估計，方能改善前述問題。本文先透過集群分析將基因資料分爲獨立群和相關群，再針對各群虛無假設爲眞的個數給一較適當的估計。再者，假設基因資料呈混合型多變量常態分配，參數具先驗分配，我們利用貝氏驗後分配和 EM 演算法估計分配中虛無假設爲眞的比例，進而估計虛無假設爲眞的個數和估計 FDR。經由統計模擬，結果顯示當基因個數愈大，所提經驗貝氏之 EM 估計法在獨立群有較小 RMSE，Ma, et al.(2005) 應用 McNemar 檢定統計量之方法，在 FWER= 25% 能有效降低 RMSE，其他方法在不同參數下各有優劣。

關鍵詞：整體型 I 誤發生率、錯誤發生率、集群分析、EM 演算法

Estimation of a General Semiparametric Promotion Time Cure Model

Chi-Hua Wang* (王啓樺) and Chin-Tsang Chiang (江金倉)

Institute of Applied Mathematical Sciences, National Taiwan University

Abstract

For survival data with a cure fraction, we propose a general semiparametric model which is derived from the biological process of cancer-relapse mechanism and includes both the mixture cure models and promotion time cure models as two special cases. The pseudo integrated least squares estimators (PILSEs) of index coefficients are shown to be consistent and asymptotically normal and an efficient computing algorithm is proposed to calculate the PILSEs of index coefficients and the moment-type estimators of promotion time function. Simulation studies are conducted to examine the finite-sample properties of the proposed estimators.

Keywords: cure model, pseudo integrated least squares estimator , single index model.

台灣愛滋病實例研究— 以聯合模型探討病毒乘載量對愛滋病患存活時間之關係

林家聿

國立中央大學統計研究所

摘要

本篇論文利用病毒乘載量來預測愛滋病患的發病時間，並探討使用雞尾酒療法對愛滋病患是否有療效。此種包含長期追蹤共變數與存活時間的資料，常因爲長期追蹤資料的測量誤差與生物體本身的差異，以及共變數觀測值與存活時間有關時，導致推論上會產生偏差，因此本研究利用聯合模型來解決這樣的問題。在生物指標的部分，使用線性隨機效應模型對長期追蹤資料作配飾，並利用概似比檢定診斷長期追蹤模型的適合度；在事件時間的部分，使用 AFT 模型描述共變數與存活時間的關係。結合這兩個部分建構出聯合模型，並且使用 EM 演算法對參數做估計。

關鍵詞：聯合模型、長期追蹤資料、AFT 模型、EM 演算法。

利用經驗貝氏方法估計錯誤發現率

張家源*、黃佳慧

國立臺北大學統計學系

摘要

復發事件資料常出現在生物醫藥的研究。在觀察期間內，研究對象可能發生超過一次感興趣之事件的資料。通常這些感興趣的事件可以被分爲至少一類的類別。然而，當事件類型發生缺失時，統計分析的過程中會出現困難並且導致參數低估的情形。在本文中我們運用 Lin et al. (2013) 提出事件類型缺失機率相等的假設和事件類型機率估計的方法。透過參數化的韋伯分配模型並結合資料類型是否可以被觀測到的觀點得到參數的估計方程式。最後透過電腦上資料的模擬並做兩階段最大概似法的估計。觀察參數估計的偏誤、標準誤、樣本平均的均值標準差及覆蓋概率，評估提出的統計方法是否具有不偏性和一致性，並且是否能夠提高估計的精確度。

關鍵詞：復發事件資料、缺失事件類別、韋伯分配模型、最大概似法估計

成對重覆觀察事件之統計檢定方法與樣本數計算之研究

鍾佳華*、蘇佩芳

成功大學統計學系

摘要

本研究主要針對觀察值相依的重覆觀察資料進行分析，如觀察初次接受口腔餵食訓練的早產兒餵食前後兩段時間內發生心跳速率異常事件的分配情況，由於在同一早產兒身上觀察了餵食前和餵食後兩段時間，即爲成對重覆觀察事件 (paired recurrent events) 之一例。針對成對重覆觀察事件的資料，本研究提出檢定方法，檢定餵食前和餵食後兩段時間內事件發生的頻率是否相同；進一步假設重覆觀察事件爲混合卜瓦松過程模型，在模型的強度函數 (intensity function) 中使用脆弱因子 (frailty) 表示每個早產兒不能觀測到變異與餵食前後的相關性，計算達特定檢定力所需之樣本數公式，並利用模擬資料，嘗試不同的參數組合，計算模擬情況下之經驗型 I 錯誤率與經驗檢定力，以驗證上述統計方法之合理性。

關鍵詞：成對重覆觀察事件、混合卜瓦松過程、脆弱因子、樣本數

網絡特徵之統計推論

林聖翔*、郭錕霖

國立高雄大學統計學研究所

摘要

網絡在數學語言上的表示爲一張圖 (graph)，由多個點 (node) 和點與點之間的連線 (link) 所組成的。近年來網絡資料的呈現在各個領域廣泛被使用，例如：朋友圈的友誼網絡、生物學的蛋白質交互作用的網絡、生態中的食物鏈網絡以及網際網路中臉書好友關係等，因此社會網絡分析成爲了熱門的發展技術。而當一筆網絡資料形成後，研究者通常會想了解一些參數的數值，像是網絡的密集程度 (density)、三角形 (triangle) 的個數或參與中間度指標 (betweenness centrality) 等，這些數值如同點估計般地呈現。本研究欲提出重抽樣的方法使得我們有機會得到合理的區間估計，雖然無法提出任何理論證明，但從模擬資料與實際資料的分析結果來看，我們的方法確有潛力，期望未來有研究者可以證明我們的概念。

關鍵詞：網絡、社會網絡分析、重抽樣、區間估計

A Performance Study on Parallel Computing between CPU and GPU on Swarm-intelligence Based Evolutionary Algorithm

Frank Po-Chen Lin*

National Sun Yat-Sen University

Frederick Kin Hing Phoa

Academia Sinica

Abstract

As the amount of processing data become inconceivable as the sequential programming no longer suffices, the evolutionary algorithm (EA), which is a population-based stochastic technique, has been successfully applied to solve problems in optimization, search and machine learning with increasing difficulty and complexity. Parallel implementations significantly improve the efficiency of EA and return high-quality results in reasonable executive times. In this paper, we present different parallel models, Open Multi-Processing (OpenMP) and Compute Unified Device Architecture (CUDA) on the swarm intelligence based evolutionary algorithm. They are commonly used in large-scale systems with a large number of potentially relevant factors, which sometimes can be prohibitive in terms of time on simulating for its big amount of data. In specific, the performance on the algorithm on the search of supersaturated designs, namely SIBSSD, is served as our benchmark of EAs.
For CUDA, in our proposal, one GPU thread takes charge of one particle (or one SSD) and launches a separate kernel at each step to achieve the necessary synchronization. For OpenMP, on the other hand, the for-loops inside the kernel are parallelized. The result shows that for Swarm-Intelligence based algorithm, each cell does not perform the same calculations and makes it unable for the elements to achieve highly data-parallelism. All the calculations of the cells in one particle were carried out by a single thread in the GPU and would be eventually time consuming for the weak cores in GPU, which would be less efficient than CPU. In conclusion, we promote OpenMP for parallel programming on Swarm-Intelligence based evolutionary algorithms and emphasize the significant irreplaceable nature of CPU on parallel programming.

A Dynamic Weighting Method and Analysis

Yi-Te Chang (張奕得)

國立交通大學統計學研究所

Abstract

Markov Chain Monte Carlo method is a universal-used method in numerical integration. In this talk, we will discuss the dynamic weighting MCMC proposed by Wong and Liang (1997), which makes the Markov chain converges faster. In the decades, Metropolis Hasting algorithm is an important simulation method, but there are still some drawbacks in the simulation. For example, the movement of the process can be influenced by some tiny probability nodes. This phenomenon may directly affect to our simulated estimation. Our main work is to review the weighted MCMC and give some theoretical proof in some special cases. Through the manner, we can make the MCMC method more efficient.

Keywords: Monte Carlo, Markov Chain, Dynamic Weight, Metropolis-Hasting Algorithm.

三種時間相依的接受者作業特徵曲線下面積估計方法之比較與修正

張雅玟

國立中央大學統計研究所

摘要

在醫學診斷中，通常會記錄病患回診所測量的共變數值，即爲時間相依的共變數值，有了長期追蹤資料的性質，即不適用一般的接受者作業特徵曲線下面積 (AUC) 來判斷生物指標對於疾病預測能力的程度，因此根據 Heagerty 和 Zheng (2005) 和 van Houwelingen, Putter (2012) 以及 Blanche, Dartigues 和Jacqmin-Gadda (2013) 所提出的方法，皆爲估計時間相依的 AUC 方法。由於這三種方法主要是根據 Heagerty 和 Zheng (2005) 的架構再分別透過不同的估計方法去計算時間相依的 AUC，因此主要針對 Heagerty 和 Zheng (2005) 的方法並進一步透過模擬和實例分析來探討隨著時間的不同，AUC 對於生物指標預測疾病能力的程度。由於 Heagerty 和 Zheng (2005) 是使用部分概似函數，因此本論文預期使用聯合模型可以解決部分概似函數的缺失問題，使得時間相依的 AUC 可以有更精確的估計結果。

關鍵詞：接受者作業特徵曲線、接受者作業特徵曲線下面積、時間相依接受者作業特徵曲線下面積、附帶型敏感度、動態型特異度、生物指標、部分概似函數、聯合模型

模糊績效評估—以台南市對登革熱疫情防範之應用

李唐榮*、吳柏林

國立政治大學應用數學系

摘要

研究動機與目的：本研究探討台南市政府對登革熱疫情防範措施之績效評估，根據相關研究調查，台灣每年都有登革熱疫情的案例，尤其以台南較爲嚴重，爲了瞭解政府是否有有效的預防登革熱，並且了解疫情是否受到控制，使大眾感到安心，故筆者對台南市政府之措施進行相關研究評估。
研究方法：本篇採用最大隸屬度函數估計法，並且以隨機抽樣的形式發放問卷，由於有些樣本會有「誇大」之可能性，此研究方法可排除離群值，使整體評估更完善，且更具有公信力。
發現：看各準則及方案的滿意度及重要性，發現政府實施的一些措施與大眾實際感受到的層面有些差異，但整體來說大眾還是可以接受的。
結論及貢獻：(1) 解政府對於登革熱防範措施是否對疫情的減緩有所幫助。(2) 從市民的角度來看，探討政府機構績效評比要點各構面之妥適性。

關鍵詞：模糊績效評估、模糊統計、登革熱

人臉辨識唐氏症患者之分類方法

李偉鈴*、羅夢娜

國立中山大學應用數學系

謝凱生

高雄長庚醫院兒童內科部

摘要

唐氏症候群病人也被稱爲「國際人」，由於各國患者的面容有相似的特徵，例如：頭部長度較一般人短，面部起伏較小，鼻子與眼睛之間的部分較低，耳朵整體看上去呈圓形而且位置較低等等。但是沒有一項單項的表徵一定會的出現在每一位唐氏症患者的身上。本文探討如何利用臉部影像資料，以及常用之機器學習方法來辨識唐氏症患者。我們收集一些唐氏症患者與一般非唐氏症患者之照片，將臉部影像資料經過資料處理後，找出五官之比例與五官面積之所佔比例。再使用相關的資料探勘方法，如邏輯回歸 (Logistic regression)、交叉驗證 (Cross-validation) 及 ROC 曲線 (receiver operating characteristic curve) 等，期望能找出最有效之分類方法，配合醫生的經驗法則來建立影像中唐氏症病徵的主要規則。

關鍵詞：影像處理、邏輯回歸、交叉驗證、ROC 曲線

電力負載量之短期預測

徐朮*、羅夢娜

國立中山大學應用數學系

盧展南

國立中山大學電機工程學系

摘要

許多電力系統的操作都仰賴於負載預測，包括發電機組排程、安全電網分析等等，代表負載預測是電力系統有效運行與規劃的重要工作。電力需求的高估將會造成過於保守的運轉，導致啓動許多機組去供應非必要的備轉容量或過多的能源採購以及在電力設施的大量投資浪費；另一方面，低估可能會帶來無法達需求量及運轉上風險，即時備轉容量的準備不足，使電力系統暴露在較爲危險的運轉範圍內操作。

本研究將依據系統歷史負載資料，氣象歷史資料 (體感溫度)，氣象局氣象預測資料 (體感溫度) 來進行負載預估，並考慮負載序列具有日內時段週期及週內週期兩種循環，建立半參數迴歸模型，且在模型的殘差考慮時間序列效應。評估模型的依據爲，預測未來 24 小時的負載量，與眞實值比較計算誤差。在建立模型的過程中，考慮有特殊假日的影響 (農曆春節與國定假日)，又可分成不同的處理方式。利用預測日的前三週作爲訓練樣本，計算出在 2012 年至 2015 年的平均絕對誤差比例，以呈現此短期負載預測方法的有效性。

關鍵詞：平均絕對誤差比例、半參數迴歸模型、週期性函數基底

Dependence Measures and Competing Risks Models under the Generalized Farlie-Gumbel-Morgenstern Copula

Jia-Han Shih* (施嘉翰) and Takeshi Emura (江村剛志)

Graduate Institute of Statistics, National Central University

Abstract

The first part of this paper reviews the properties of dependence measures (Spearman's rho, Kendall's tau, Kochar and Gupta's dependence measure, and Blest's coefficient) under the generalized Farlie-Gumbel-Morgenstern (FGM) copula. We give a few remarks on the relationship among the dependence measures, derive Blest's coefficient, and suggest simplifying the previously obtained expression of Kochar and Gupta's dependence measure. The second part of this paper derives some useful measures for analyzing dependent competing risks models under the generalized FGM copula. We obtain the expression of sub-distribution functions under the generalized FGM copula, which has not been discussed in the literature. With the Burr III margins, we show that our expression has a closed form and generalizes the reliability measure previously obtained by Domma and Giordano (in Statistical Papers 54: 807-826, 2013).

Keywords: Blest's coefficient, competing risk, FGM copula, Kendall's tau, Spearman's rho

探討細格數爲零之多變量二元分配估計問題

梁丞智*、黃怡婷

國立臺北大學統計學系

摘要

在一階馬可夫鏈假設下，葉麗芬 (2014) 建構了一個多變量二元聯合機率分配，以數值分析方法求得最大概似估計值，並以費雪訊息矩陣之反矩陣來估計最大概似估計值之標準誤。在其模擬結果中，在小樣本時會有細格數爲零的問題，因而使反矩陣不存在而無法求得標準誤估計值的情況。本論文嘗試使用 Firth (1993) 所提出懲罰最大概似函數 (Penalized Likelihood function) 以及在細格數加上 0.5 等方法來解決此問題，並與數值解來進行解析比較，討論所提方式的可行性與不同估計方法之精確度。

關鍵詞：多變量二元聯合機率分配、懲罰最大概似函數

A Review and Comparison of Continuity Correction Rules; The Normal Approximation to the Binomial Distribution

Yu-Ting Liao* (廖昱婷) and Takeshi Emura (江村剛志)

中央大學統計研究所

Abstract

In applied statistic, the continuity correction is useful when the binomial distribution is approximated by the normal distribution. In the first part of this thesis, we review the binomial distribution and the central limit theorem. If the sample size gets larger, the binomial distribution approaches to the normal distribution. The continuity correction is an adjustment that is made to further improve the normal approximation, also known as Yates's correction for continuity (Yates, 1934; Cox, 1970). We also introduce Feller's correction (Feller, 1968) and Cressie's finely tuned continuity correction (Cressie, 1978), both of which are less known for statisticians. In particular, we review the mathematical derivations of the Cressie's correction. Finally, we report some interesting results that these less known corrections are superior to Yates's correction in practical settings.

Keywords: binomial distribution, continuity correction, normal approximation

Ratios of Stirling Numbers and Efficient Computation of Minimum Variance Estimators

Sara Kropf* and Hsien-Kuei Hwang

Institute of Statistical Science, Academia Sinica

Abstract

For several parameters of distributions, the minimum variance unbiased estimator can be expressed as the ratio of consecutive terms of a combinatorial sequence like Stirling numbers. In applications, it is necessary to efficiently compute these ratios. However, that is non-trivial: Both the numerator and the denominator grow exponentially fast, but the quotient is small. So it is preferable to directly compute the quotient instead of computing numerator and denominator separately.
We give a general procedure to compute an asymptotic expression for such ratios directly. The advantages of our approach are:

- It is easily applicable to a wide range of combinatorial sequences.
- The direct computation of the ratio circumvents cancellations occurring when computing numerator and denominator separately.
- Our asymptotic expression is uniformly valid for the whole range of the two parameters. Thus it can be used without thinking about different growth rates of the two parameters.

Keywords: minimum variance unbiased estimator, ratios of Stirling numbers, asymptotic expansion.

Universal Sampler for Truncated Univariate Density Kernels

Yuchung J. Wang

Rutgers University, Camden, New Jersey 08102 USA

Abstract

Generating samples from a probability kernel, density function without its normalizing constant, plays a critical role in computational statistics. This talk proposes a new sampling method, called universal sampler (US). US uses a uniformly scattered quasi-random numbers as the candidates' pool, from which a batch of samples are selected, then randomly refresh the pool for the selection of next batch. After the sampling stops there are n batches of samples, and one sample from every batch constitutes an IID samples. US uses n-out-of-n-batches bootstrap (n-o-nB) to approximate the sampling distribution and calculate the Monte Carlo error (MCE). While nonparametric bootstrap, based on one sample, can hardly does bias correction, n-o-nB does this in a straightforward fashion. It is a direct method because it does not require a proposal distribution, and burn-in is not needed either. The sampler is compared with seven random number generators of the R package. Without exception, the proposed algorithm is as accurate and consistent as the R functions. The performance of the sampler is also found to be on par with the benchmark in sampling mixtures of two normal distributions. When compared with MCMC algorithms like stepout slice, interval slice, Metropolis adoptive rejection method, and univariate Metropolis-Hastings algorithm, US is the only sampler that consistently has decent biases, and its MCE is close to the theoretical value. I will report the numerical comparisons.

Keywords: accept/reject method, bootstrap, Monte Carlo error, sampling importance resampling, slice sampler.

簡化的 ARMA-GARCH 風險值估計法及其在投資組合上的應用

林士傑*、俞淑惠

國立高雄大學統計學研究所

摘要

巴塞爾協定提出以風險值 (value-at-risk) 衡量市場風險，然而風險值若是估計的過於保守或是不夠精確，將會導致銀行資金無法活用或遭金管會介入，因此，風險值的估計，是一個重要的議題。考量財務資料具有時間相關的特性，本文中我們將以時間序列模型估計風險值，由此產生了一個新的問題。由於資料每日更新，而決定模型中的階數無法完全自動化，需耗費相當的人力與時間，因此本文發展出一套簡化的 ARAMA-GARCH 的風險值估計法，以簡化模型配適的複雜度，並利用巴塞爾協議中的懲罰表，以風險及資金是否靈活運用這兩個不同的觀點，對新的風險值估計方法做風險及財務方面的比較。在實證分析中，我們以新的估計方法分析 86 筆資料，包含美國大盤、亞洲大盤、台股、黃金、石油 ...等，並與歷史模擬法做比較，結果顯示，新的估計方法可將合理超限數 (介於 0 到 7 的超限數) 的次數增加約 36%，而平均所需資本準備金則比歷史模擬法降低約 9% 至 14%，由此說明，新的估計方法可以得到不錯的結果。本文亦以前述風險值估計法運用於投資組合的問題上，用來作爲挑選資產的標準，並找出最佳化投資組合策略。

關鍵詞：風險值、時間序列、ARMA-GARCH、異質性、投資組合

A Study on the Optimal Order Execution Problem for Stochastic Market Depth Models

Kuan-Chich Huang* (黃冠智) and Mei-Hui Guo (郭美惠)

Department of Applied Mathematics, National Sun Yat-sen University

Abstract

Optimal order execution problem is an important issue faced by institutional traders, i.e. how should a trader splits a large order into small market orders over time to minimize his execution cost? Chen, Kou and Wang (2016) proposed a partition algorithm to solve this problem for a limit order book model. They assume the market depth is stochastic and is governed by a Markov chain, which ?ts into the framework of Markov decision processes. We revisit their study and investigate performance of the partition algorithm using stock market high-frequency transaction data and simulated data with market depth satisfying the geometric Ornstein-Uhlenbeck processes.

Keywords: optimal order execution problem, Markov decision processes, market depth, Markov chain, Ornstein-Uhlenbeck processes, high-frequency transaction, partitioning algorithm

Analysis of Indonesia Stock Market Volatility

Bakti Siregar* (錫誠嘉) and Meihui Gou (郭美惠)

Department of Applied Mathematics, National Sun Yat-sen University

Abstract

Liberalization and economic integration become topics of discussion and research in recent years. Indonesia is one of the countries that actively participates in the achievement of liberalization and economic integration, especially in the ASEAN region. Indonesia stock market has a high degree of volatility which can be used to produce high investment returns, which is one of the reasons to attract foreign investors to enter Indonesia stock market. Volatility plays an important role for market participants to control and reduce their market risk of financial assets.

In this study we establish the volatility models for the stocks listed in the Indonesia stock market index LQ45. The models we considered include the Autograssive Conditional Heteroskedasticity (ARCH) proposed by Engle (1982), Generalized Autograssive Conditional Heteroskedasticity (GARCH) by Bollerslev (1986), the Stochastic Volatility Model (SVM) by Jacquier, Polson and Rossi (1994), and Autoregressive Moving Average (ARMA) by Box, Jankins, and Reinsel (1994). We use the daily closing stock prices to establish the models and predict the future volatility.

We also apply machine learning methods such as the k-mean method to find the volatility clusters of Indonesia stocks. Finally, I will investigate profitable portfolios in Indonesian stock market.

Keywords: volatility, ARCH, GARCH, clustering, portfolio

Managerial Overconfidence and Capital Structure Adjustment

Ting-Yen Wu (吳亭諺)

國立臺北大學統計學系

Abstract

Studies show that the managers will be conditioning the financing strategy on the target capital structure, and adjust the capital structure toward its target. This study aims to figure out how overconfident CEOs manage leverage toward a target through time. We find that managerial overconfidence will lead to an increase in adjustment speed for over-levered firms, while no evidence showing the association between overconfidence and adjustment for under-levered firms. We further find that over-levered firms with overconfident CEOs will speed up capital structure adjustments only when firms have rich firm-specific growth opportunities or during prosperous periods, and they tend to use equity issuance to reduce leverage rather than debt retirement due to the tendency to invest despite that equity issuance is costly for them. This result suggests that overconfident CEOs will take into account the optimal leverage ratio when making financial strategies, especially for over-levered firms.

Keywords: overconfident CEO, capital structure adjustment, growth opportunity, financial crisis.

Capital Structure Adjustments around the Financial Crisis of 2007-2009

Ju-Fang Yen (顏汝芳) and Yu-Ju Hsiang* (項鈺茹)

National Taipei University Department of Statistics

Abstract

Recent research has found that macroeconomic conditions are important factors in analyzing firms' financing choices as well as the adjustment behaviors of the capital structure. This study analyze whether the most recent financial crisis of 2007-2009 has impacts on the implementation of trade-off theory for capital structure. We find evidence that firms adjust their capital structure toward target leverage ratio slower during the crisis years, in particular for the over-levered firms. Furthermore, we find that over-levered firms prefer reduce leverage by debt retirement. Hence, the detrimental effect of financial crisis on adjustment speed is more significant for over-levered firms with lower cash holdings and worse operating performance during crisis years. This study thus sheds light on the importance of financial flexibility during the crisis years through capital structure adjustment mechanism.

Keywords: financial crisis, target leverage, capital structure adjustment

偵測股票市場中具有影響力的交易

陳元豪*、郭美惠

國立中山大學應用數學系

摘要

我們使用紐約證券交易所的高頻交易資料，探討股票市場中具有影響力的交易。第一部份，我們將高頻交易資料，依據每筆交易資料對於市場長短期交易 (如：交易方向、交易數量等變數) 的影響，定義市場中具有影響力的交易。我們將交易的報酬率以及交易量作爲一組變數組合，交易方向以及有資訊交易機率作爲另一組變數組合，對兩組變數組合分別建立其聯合分佈，使用異常點偵測的方法將交易資料分成兩類：具有影響力交易以及一般交易。第二部份，我們考慮三個市場反應變數對具有影響力交易的反應變數建立羅吉斯回歸以及支持向量機模型，並探討這兩種模型對高頻交易資料類別的預測能力。

關鍵詞：高頻資料、異常點偵測、具有影響力交易、有資訊交易機率、羅吉斯回歸、支持向量機

Machine Learning Pairs Trading

Yu-Jin Lai* (賴俞瑾) and Mei-Hui Guo (郭美惠)

Department of Applied Mathematics, National Sun Yat-sen University

Abstract

Pairs trading is a comparative-value form of statistical arbitrage designed to use temporary random departures from equilibrium pricing between two stocks. In the first part, we use the spreads of cointegrated pairs and pre-chosen thresholds to perform pair trading for daily data. We investigate the effects of several selected covariates (e.g. EPS, strength of cointegration and etc.) on the pairs trading profits. We use principal component analysis and sliced inverse regression to find risky covariate zones which result in unprofitable pairs. In the second part, we conduct high-frequency pairs trading for intraday data. We use several high frequency covariates (e.g. money flow, relative strength index and etc.) as input features for support vector machine classification to set up trading signals of entering positions. We investigate the performance of the proposed pairs trading strategies for stocks in S&P 500 index.

Keywords: cointegration, principal component analysis, profit, sliced inverse regression, spread, support vector machine.

多變量管制圖在監控多階段系統製程品質上之應用研究

許竣幃*、潘浙楠

國立成功大學統計學研究所

摘要

傳統的統計製程管制 (Statistical Process Control, SPC) 在監控與改進製程品質的方法上，通常僅針對單一階段具相關品質特性的製程品質進行監控與改善。Pan, Li and Wu (2015) 提出監控多階段製程具單一產品品質特性之管制圖，但適合監控多階段系統製程品質的多變量管制圖仍付諸闕如。因此在考慮相關品質特性的情況下，本研究利用多變量線性迴歸模型 (Multivariate Linear Regression Model) 來描述跨階段製程的相關性，再利用各階段模型的殘差建構殘差 MEWMA 及 MCUSUM 管制圖。此外，本研究將以整體連串長度 (Overall Run Length, ORL) 的想法作爲各種多變量殘差管制圖偵測能力之評估與比較基準。最後我們以一個級聯資料 (cascade data) 爲例進行數值實例的驗證與說明，研究結果可提供實務工作者在監控多階段系統製程品質上之參考。

關鍵詞：多階段系統、多變量線性模型、整體連串長度

Threshold Degradation

Chien-Yu Peng (彭健育) and Ya-Shan Cheng* (鄭雅珊)

Institute of Statistical Science, Academia Sinica

Abstract

Accelerated degradation tests (ADTs) are widely used to assess the lifetime information for highly reliable products. One restrictive assumption with a conventional ADT model is specifying which parameter depends on explanatory variables in advance. The assumption can lead to misuse of physical/chemical mechanisms and unreasonable extrapolation of the product's lifetime at the normal-use conditions. This study proposes a two-stage approach (named threshold degradation) as an alternative model with regression structures that accommodate explanatory variables. Several real examples are performed to show the differences between the conventional ADT model and the threshold degradation model and to demonstrate the advantages of the latter.

Keywords: first passage time, Gaussian process, goodness of fit, random effects.

半競爭風險資料下的韋伯迴歸模型

鄭絜文*、黃佳慧

國立臺北大學統計學系

摘要

本篇論文探討韋伯迴歸模型的半競爭風險資料之風險函數與概似函數估計法。半競爭風險資料分爲非終端事件與終端事件，非終端事件會受到終端事件所影響造成相關設限。我們假設病人疾病復發爲非終端事件，而病人死亡爲終端事件。將資料依照死亡前是否發生疾病復發分爲兩組且機率與羅吉斯有關，死亡前復發組和復發前死亡組之風險函數皆服從韋伯分配。我們依照上述條件模擬生成半競爭風險資料，並使用概似函數估計法探討參數的偏誤、標準誤、均值標準差及覆蓋率。

關鍵詞：半競爭風險資料、韋伯迴歸模型、最大概似估計法