Robust mislabel logistic regression without modeling mislabel probabilities

Hung Hung (洪弘) and Zhi-Yu Jou
Institute of Epidemiology and Preventive Medicine, National Taiwan University

Su-Yun Huang (陳素雲)*
Institute of Statistical Science, Academia Sinica

Abstract

Logistic regression is among the most widely used statistical methods for linear discriminant analysis. In many applications, we only observe possibly mislabeled responses. Fitting a conventional logistic regression can then lead to biased estimation. One common resolution is to fit a mislabel logistic regression model, which takes into consideration of mislabeled responses. Another common method is to adopt a robust M-estimation by down-weighting suspected instances. In this work, we propose a new robust mislabel logistic regression based on γ-divergence. Our proposal possesses two advantageous features: (1) It does not need to model the mislabel probabilities. (2) The minimum γ-divergence estimation leads to a weighted estimating equation without the need to include any bias correction term, that is, it is automatically bias-corrected. These features make the proposed γ-logistic regression more robust in model fitting and more intuitive for model interpretation through a simple weighting scheme. Our method is also easy to implement, and two types of algorithms are included. Simulation studies and the Pima data application are presented to demonstrate the performance of γ-logistic regression.

Keywords: Classification; Logistic regression; Minimum divergence estimation; Mislabeled response; Robust M -estimation.

A robust RUV-testing procedure via γ-divergence

Hung Hung

Institute of Epidemiology and Preventive Medicine, National Taiwan University, Taiwan

Abstract

Identification of differentially expressed genes (DE-genes) is commonly conducted in modern biomedical researches. However, unwanted variation inevitably arises during the data collection process, which could make the detection results heavily biased. It is suggested to remove the unwanted variation while keeping the biological variation to ensure a reliable analysis result. Removing Unwanted Variation (RUV) is recently proposed for this purpose by the virtue of negative control genes. On the other hand, outliers are frequently appear in modern high-throughput genetic data that can heavily affect the performances of RUV and its downstream analysis. In this work, we propose a robust RUV-testing procedure via γ-divergence. The advantages of our method are twofold: (1) it does not involve any modeling for the outlier distribution, which is applicable to various situations; (2) it is easy to implement in the sense that its robustness is controlled by a single tuning parameter γ of γ-divergence, and a data-driven criterion is developed to select γ. In the Gender Study, our method can successfully remove unwanted variation, and is able to identify more DE-genes than conventional methods.

Key words: γ-divergence; negative control genes; robustness; RUV; unwanted variation.

Estimation and Model Checking for General Semiparametric Recurrent Event Models with Informative Censoring

Hung-Chi Ho
Department of Internal Medicine & Big Data Center, China Medical University and Hospital, Taiwan

Abstract

This research aims to explore a recurrent event process with informative censoring using more general semiparametric latent intensity regression models. When the distributions of the subject-specific latent variable and the censoring time are left unspecified, the distinct distributional features of the recurrent event times are found to be linked to the shape parameter, which, hence, merits the development of estimation and testing procedures. In light of this finding, two contrasting estimation methods are proposed for shape-dependent and -independent models. Especially, the estimation criteria are useful in building test rules to distinguish between competing rate regression models without the need to specify a significance level. Under very mild conditions, we establish large-sample properties of the estimators and test statistics. Comprehensive simulations are further conducted to assess their finite-sample performance. Moreover, our methodology is demonstrated by applying it to recurrent event samples of intravenous drug users needing inpatient care and patients with chronic granulomatous disease.

Keywords: informative censoring; model selection consistency; occurrence rate function; pseudo estimation; shape parameter

現代化CRM中的客戶行為分析

楊思(Jamie Yang)

叡揚資訊雲端及巨資事業群

摘要

隨著數位科技的迅速發展，企業可以蒐集到更豐富、完整的消費者行為資料：從線上的網站瀏覽、對行銷訊息的反應到線下實體店的消費等等。自上世紀90年代初就開始發展起來的 Customer Relationship Management(CRM) 也在技術的驅動下，不斷演進升級：融合了社交、行動化、雲端和人工智慧。本次演講，將分享業界如何透過現代化CRM系統中的客戶行為分析來了解和預測消費者的偏好，實現個人化行銷和服務；也會談到分析模組開發和維運的經驗以及遇到的挑戰。

關鍵詞：客戶關係管理，預測性客戶行為分析，分眾行銷

在設計、製造端的數位智慧化

卓嘉純

啟碁科技 工程統計資料管理部

摘要

近年來資訊技術、機器學習、人工智慧等持續快速發展，而在工業現場透過這些技術，最佳化產品設計，以因應自動化生產；在生產前優化生產排程提高生產力，在生產中能偵測異常與故障預警，以邁向智慧製造的目標。

關鍵詞：智慧製造、人工智慧、異常偵測、故障預警

統計在工業上之落地應用

林尚毅
鴻海科技

摘要

有賴於近年大量數據成長、儲存運算設備效能提升，結合演算法和軟體工程，落實了各種耳熟能詳的大數據與人工智能的場景應用，如人臉辨識、自駕車等。至於講者所處的工業領域上，也正受這波浪潮席捲，數個耳熟能詳的名詞，如工業 4.0、智慧製造、工業大數據、工業人工智能等，正是這個階段下的產物，且更甚有工業互聯網的議題出現。在這波浪潮中，除了擁有大量製造相關數據的製造本體開始積極推動這方面的轉型外，也開始有大量的上游設備商偕同網路通訊、資料儲存、資料分析及應用等項目，開發自身搭載於設備的分析工具與解決方案，藉此提高整體獲利。此一勢不可擋的趨勢也可從終端的互聯網與移動互聯網公司中發覺，這類公司嘗試跨足製造業、零售業等實體經濟，主要目的除了擴展版圖外，更可以降低虛擬經濟上的高風險。講者處於大型製造業公司內，又如何利用自身所學，協助將工業數據應用落地於製造產業上，並利用一個案例說明如何降低和磨平製造與統計兩個專業領域間的隔閡，讓落地更快速、無痛。

關鍵詞：製造、統計、大數據、人工智能、工業 4.0

Use of a two-sided tolerance interval in the design and evaluation of biosimilarity in clinical studies

姜杰*
國家衛生研究院

陳啟天
晉加股份有限公司

蕭金福
國家衛生研究院

摘要

In assessing the biosimilarity between two products, the problem is always to answer the question "How similar is similar?" Traditionally, the equivalence of the expected effects between products is the primary consideration in a clinical trial. Alternatively, a two-sided tolerance interval test with a recommended biosimilarity margin is suggested herein to assess biosimilarity such that the variations can be considered simultaneously. We derive an asymptotic distribution of the tolerance limits. Thus, the sample size can be determined to achieve a targeted level of power. A numerical study shows that the two-sided tolerance interval test is comparable to the traditional equivalence test. A real example is presented to illustrate our proposed approach.

關鍵詞： Biosimilarity; Margin; Two-sided tolerance interval; Asymptotic distribution; Sample size determination.

An Empirical Bayes Approach to Statistical Assessment of Biosimilarity with Binary Data

Yu-Chieh Cheng(鄭宇傑)* 、Hsiao-Hui Tsou(鄒小蕙)

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

Chi-Tian Chen(陳啟天)

Technical Operation, StatPlus, Inc., Taipei, Taiwan

Ya-Ting Hsu(許雅婷)、Hsiao-Yu Wu(吳小玉)

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

Abstract

Because of the expiry of patent protection for many biological products, biosimilars have attracted attention from the clinical trial sponsors and policymakers. A biosimilar is a biological medicine highly similar to another innovative biological medicine (the reference medicine). In this presentation, we provide two criteria to evaluate the similarity between a biosimilar and its reference product using empirical Bayes approach. The primary endpoint is a binary endpoint. The common prior of a biosimilar and its reference product is the historical information of this reference product. Based on this prior, we establish posterior probabilities of these two criteria to evaluate the similarity. Furthermore, we discuss the impact of design parameters (such as treatment response and limits for assessing biosimilarity) for the posterior probabilities. The sample size determination is presented with the posterior probabilities of two criteria and corresponding pre-specified limits. Finally, a real example for patients with malignant lymphoma is given to illustrate our approach.

Keywords: Biosimilar, Binary endpoint, Prior, Empirical Bayes, Sample size determination

Challenge for a statistician in a contract research organization

Sky Chi-Tian Chen (陳啓天)
StatPlus Inc.

Abstract

I am a statistician work in a local CRO. Before that, I was a postdoctoral researcher in NHRI and focused on statistical research in clinical trials. Discrepancy in job scope between academic institute and industry is the first thing I faced. Some experience of mine may be interesting to share. In last two years, I have to figure out some sophisticated and challenging statistical/clinical issues in real world of clinical trials. I will introduce a practical case of combo drug in our projects. The case involves the concept of factorial design and phase II/III study. What issues we confronted and what the role of a statistician in communication with client and authority will be my point in this talk.

Keywords: factorial design, overall type I error rate, overall power, phase II/III design

Estimation of treatment effect and sample size allocation to regions in multi-regional clinical trials

Hsiao-Hui Tsou(鄒小蕙)*、Yu-Chieh Cheng(鄭宇傑)、Chin-Fu Hsiao(蕭金福)、Hsiao-Yu Wu(吳小玉)、Ya-Ting Hsu(許雅婷)

Institute of Population Health Sciences(群體健康科學研究所), National Health Research Institutes(國家衛生研究院), Taiwan

Yuh-Jeng Wu(吳裕振)

Department of Applied Mathematics, Chung Yuan Christian University, Taiwan

Eric V. Slud

Department of Mathematics, University of Maryland, USA

Abstract

Multiregional clinical trials (MRCTs) has become a popular strategy in development of new medicines. The primary objective in an MRCT is to show the overall treatment efficacy of a new therapy. After demonstrating the overall treatment effect of the new drug in an MRCT, the evidence of consistency in treatment effects among regions is usually required for regional approval. In this presentation, we focus on the design and analysis of a two-arm comparative multiregional clinical trial and provide a statistical model to combine regional treatment effects for estimation of overall treatment effect. We further explore some approaches of sample-size allocation and evaluate the evidence of consistency in treatment effects among regions.

Keywords: Multiregional clinical trials, overall treatment effect, sample-size allocation, consistency

Approximate Maximum Likelihood Estimation of a Threshold Diffusion Process

余定宏、蔡恆修*
中央研究院

冉煥凱
Universitat de les Illes Balears

摘要

In order to estimate the parameters of a two-regime threshold diffusion process with discretely sampled data, an approximate maximum likelihood method (AMLE) based on approximating the log-likelihood function of the observations is proposed. Both the drift and the diffusion term are allowed to be either linear or non-linear. In order to choose the most appropriate between these four possibilities, three information criteria can be employed. Further, a likelihood ratio test can help to determine whether threshold effects are present. The finite sample performance of the proposed AMLE is compared to an alternative Quasi-likelihood estimator. The finite sample performance of the information criteria as well as the likelihood ratio test is studied. Finally, the efficacy of our approach is demonstrated with two financial time series.

關鍵詞：Irregularly-spaced data; Threshold diffusion process; Nonlinear time series; Stochastic differential equation; Maximum likelihood estimator.

Variable Selection for High-Dimensional Regression Models with Time Series and Heteroscedastic Errors

Hai-Tang Chiou* (邱海唐), Ching-Kang Ing (銀慶剛)
Institute of Statistics, National Tsing Hua University
Meihui Guo (郭美惠)
Department of Applied Mathematics, National Sun Yat-sen University

Abstract

Although existing literature on high-dimensional regression models is rich, the vast majority of studies have focused on independent and homogeneous error terms. In this article, we consider the problem of selecting high-dimensional regression models with heteroscedastic and time series errors, which have broad applications in economics, quantitative finance, environmental science, and many other fields. The error term in our model is not only allowed to be short- or long-range dependent, but also contains a high-dimensional dispersion function accounting for heteroscedasticity. By making use of the orthogonal greedy algorithm and the high-dimensional information criterion, we propose a new model selection procedure that can consistently choose the relevant variables in both the regression and the dispersion functions. The finite sample performance of the proposed procedure is also illustrated via simulations and real data analysis.

Keywords: Heteroscedasticity, High-dimensional information criterion, Orthogonal greedy algorithm, Short- and long-range dependence

Forecast-emphasized Principal Component Analysis for Spatial Temporal Data

Han-Yueh Lee, Nan-Jung Hsu

Institute of Statistics, National Tsing-Hua University

Abstract

In many applications in environmental science, it is of interest to identify dominant spatial patterns based on spatial data with temporal replicates. For this purpose, empirical orthogonal function is the most popular approach among others which essentially adopts the principal component analysis (PCA) for spatial data. Wang and Huang (2017) proposed a regularized PCA for handling irregularly-spaced spatial data in which both smoothness and sparseness constraints are imposed on the eigen-images. Following Wang and Huang's approach, this study extends the regularized PCA to identify dominant spatial-temporal patterns, in particular the temporal dynamics of spatial patterns are taking into considerations. A gradient descent method is adopted to solve the regularized spatial-temporal PCA, which is further used as eigen-images in a dynamic spatial random effect (SRE) model. The effectiveness of the proposed method is investigated via simulation and environmental studies. As a result, in low-rank SRE model framework, the proposed approach has a better performance on temporal forecasts, compared to the alternative basis obtained from the regularized spatial PCA approach. For implications, the methodologies can also be applied for dynamic feature extraction in high-dimensional time series.

Keywords: fixed rank kriging, orthogonal constraint, smoothing splines, spatial PCA.

中国城市 CPI 指数与中国外汇存底及中银贷款利率的关系

吕存策　　吴柏林

国立政治大学应用数学系　　国立政治大学应用数学系

摘要

论文目的：研究中国城市 CPI（consumer price index）的变化规律。研究方法：本文以中国城市居民 2014 年 1 月至 2018 年 12 月间，每月的 CPI 以及外汇存底、中国银行贷款基础利率（LPR）为原始数据，利用 MINITAB 对中国城市 CPI 建立 ARIMA 模型。再用显著检验的方式确定滞后项，踢出不显著变量，使用回归模型确定关系式。研究结果：CPI 受外汇存底的变化幅度的影响，但滞后项却是外汇存底。CPI 与 LPR 的关系不显著。结论：近 5 年中国城市 CPI 指数与外汇存底相互影响，但与 LPR 的关系无法判断。当外汇存底不变时，中国城市 CPI 会以较低的幅度缓慢增长。

关键字：中国城市；CPI；外汇存底；LPR

Large portfolio management with clustering techniques

鄧惠文*
國立交通大學資訊管理與財務金融學系

黃春僖

國立交通大學應用數學系

摘要

Large portfolio management faces many numerical problems and statistical difficulties. For instance, it is non-trivial to estimate a large covariance matrix that remains semi-positive, it is time demanding in the optimizing process when the dimension is high, and the optimized portfolio may not be stable or with high turnover rate. Instead of proposing an alternative approach to estimate the large covariance matrix, the aim of this paper is to propose a clustering method to overcome the above problems. With hierarchical clustering techniques, we partition the assets into several groups, so that assets behave similarly within groups but vary among groups. In each group, the asset closest to its centroid is selected as the candidate asset. The optimization procedure is then implemented for the selected small portfolio. With empirical analysis, we will show that the proposed method is comparable with that optimized directly from the large portfolio.

關鍵詞：large portfolio, clustering, portfolio management

Modeling financial interval time series

Liang-Ching Lin*

Department of Statistics, National Cheng Kung University, Tainan, Taiwan

Li-Hsien Sun

The Graduate Institute of Statistics, National Central University, Taoyuan, Taiwan

摘要

In financial economics, a large number of models are developed based on the daily closing price. When using only the daily closing price to model the time series, we may discard valuable intra-daily information, such as maximum and minimum prices. In this study, we propose an interval time series model, including the daily maximum, minimum, and closing prices, and then apply the proposed model to forecast the entire interval. The likelihood function and the corresponding maximum likelihood estimates (MLEs) are obtained by stochastic differential equation and the Girsanov theorem. To capture the heteroscedasticity of volatility, we consider a stochastic volatility model. The efficiency of the proposed estimators is illustrated by a simulation study. Finally, based on real data for S&P 500 index, the proposed method outperforms several alternatives in terms of the accurate forecast.

關鍵詞：Forecast, Interval time series, Stochastic differential equation.

從學習模型的角度看投資問題

李佳蓉

東吳大學巨量資料管理學院

摘要

在線上投資問題中，投資者需要將自己的資產分散到各種投資標的，並希望在經過一段時間之後，自己最終的投資獲益能夠越大越好。此種線上投資問題其實是線上學習問題的一個特例，而為了能掌握在金融市場中各種標的的價格變動，我們將考慮兩種不同的線上學習模型，以及相對應的線上學習演算法，並分析其優劣。

關鍵詞：線上投資問題、變動環境、變異值、偏差值

Adjusting Drawdown Risk and Return Based on Bidding Fraction

吳牧恩、林聖皓

國立台北科技大學 資訊與財金管理系

中央大學 資訊工程系

Since Kelly Criterion has been proposed by John Larry Kelly Jr., there has already been thousands of research and application to the real financial market. However, in most of these research, the fraction comes out from the Kelly formula is a constant value from the first play to the last one. Also, the implication of this fraction is only to maximize our final wealth regardless of the control of risk level. Thus in this study, we introduce the concept risk to the Kelly fraction. Specifically, we try to control the maximum drawdown from the beginning to the end of play. First, we start from the simulation of simple coin tossing game and lognormal distribution, and then we extend the simulation to the real financial data (0050.tw). All of these simulation not only merge the core concept of Kelly Criterion, also apply the perception of maximum drawdown. Finally, to make the Kelly criterion more suitable for the dynamically financial market, we apply the methodology of moving window to 0050.tw and give each 10 trading days a most suitable fraction. As a result, every 10 days, there will be a revised fraction for the next 10 trading days rather than a fixed fraction through the whole trading period.

Keywords: Optimal Fraction, Kelly Criterion, Draw Down, Back-testing

Some Limit Distributions of Discounted Branching Random Walks

Jyy-I Hong
National Chengchi University

Abstract

We consider a Galton-Watson discounted branching random walk $\{Z_n, \zeta_n\}_{n\geq 0}$, where Z_n is the population of the nth generation and ζ_n is a collection of the positions on $\mathbb{R}$ of the the Z_n individuals in the nth generation, and let Y_n be the position of a randomly chosen individual from the nth generation and $Z_n(x)$ be the number of points in ζ_n that are less than or equal to x, for $x \in \mathbb{R}$. In this talk, we present the limit theorems for the distributions of Y_n and $\frac{Zn(x)}{Z_n}$ in both supercritical and explosive cases.

Keywords: branching random walks, branching processes, coalescence, supercritical, explosive

Cover Times in Reflected Simple Random Walk and Brownian motion

May-Ru Chen
Department of Applied Math, National Sun Yat-sen University

Shoou-Ren Hsiau
Department of Mathematics, National Changhua University of Education

Chong-Yi Li (李重毅)*
Institute of Statistical Science, Academia Sinica

Abstract

We concern with the problem inspired to the simple random walk $\{S_n\}$ with the reflected barriers and to the reflected Brownian motion $\{V_t\}$ in Skorokhod mapping sense. For $\{S_n\}$ (or $\{V_t\}$), define the cover time by the first time that a process takes to visit a given coverage. As a result, we give the probability generating function (or Laplace transform) of the cover time, respectively, and the distribution of the random variable of each process at the cover time. For a standard reflected Brownian motion, we find a very simple expression of the Laplace transform of the cover time.

Keywords：simple random walk, Brownian motion, cover time, Skorokhod map

On excursions inside an excursion

陳美如
國立中山大學

Ju-Yi Yen*
University Of Cincinnati

摘要

The distribution of ranked heights of excursions of a Brownian bridge is given in a paper by Pitman and Yor. In this work, we consider excursions of a Brownian excursion above a certain level. We study the maximum heights of these excursions as Pitman and Yor did for excursions of a Brownian bridge.

關鍵詞：
Brownian excursion theory, Brownian bridge

Connecting Markov Chain to Evolution

高正雄

國立中正大學數學系

摘要

Markov Chain Model reflects the probabilistic transition among certain finite number of states in an evolution process. The daily weather change in a geographical area shall be considered as a practical example for use of Markov Chain Model to analyze the sequential dependence or independence among the states. The physical meanings of temporary state and permanent state shall be explained. The long-term transition in an irreducible group (a sub Markov Chain) turns to produce the ergodic state probabilities, in light of the Chapman Kolmogorov equations. The physical essence of such ergodic state probabilities shall be explained. More examples to which Markov Chain Model may apply to solve practical problems shall also be briefly presented.

Development of a novel algorithm to identify ceRNA-miRNA triplets

Tzu-Pin Lu* (盧子彬)、Lin Wang (王琳)

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

Abstract

With the advancements in high-throughput genomic technologies, it has become easier to examine multiple omics data in the same individual. MicroRNA (miRNA) is the most studied non-coding RNA recently, and it is well-known that miRNA is able to inhibit the expression level of its target genes. Intriguingly, the suppression phenomenon from a miRNA to its target genes is correlated with its own expression level in a certain condition. That is, the expression level of a miRNA must be taken into consideration while its regulatory effect is discussed and such miRNA and its target genes with this association is called as a competing endogeneous RNA (ceRNA) and miRNA pair. Till now, although several published algorithms and packages have been developed to identify novel ceRNA-miRNA pairs, most methods need to define different groups based on the expression level of the miRNA in prior to the analyses. However, challenge arises; the expression level of a miRNA is a continuous variable instead of a discrete variable. A subjective and arbitrarily defined grouping of a miRNA expression level may lose the true biological meaning and thus make the analyses incomplete. To address this issue, we developed a novel algorithm to identify ceRNA-miRNA pairs. First, a random walk method was used to exclude miRNA-gene pairs without any correlation. Subsequently, for the pairs with high correlation values, a circular binary segmentation algorithm was applied to obtain the peaks of the miRNA expression levels across different samples. A simulation study with different scenarios demonstrated that our algorithm is efficient and accurate to identify true ceRNA-miRNA pairs. Lastly, two real cancer datasets from The Cancer Genome Atlas (TCGA) were analyzed by our algorithm. The results suggest that our approach not only is able to validate previous findings from other studies but also can reveal several new ceRNA-miRNA candidates for further investigations.

The Discussion of Parametric Test Process in Meta-Analysis

Jin-Hua Chen

Graduate Institute of Data Science,
Taipei Medical University

Abstract

Evidence Based Medicine(EBM) is highly valued for medical development. They usually try to collect and review a lot articles in systematic review and combine the results of each paper in meta-analysis method. Systematic review is a procedure for reviewing papers in particular topic and meta-analysis is one step in systematic review. Meta-analysis is the statistical method to combines the point and standard error estimators from each article by weighted average concept. The traditional method is a two stages testing which might appear the pseudo association. Many researchers did not care this problem when they apply to medical topics. In our study, we would propose the one stage combined method in binomial response variable in the 2x2 contingency table. We could consider the heterogeneity and homogeneity between the studies, and simulation data is analogous to several scenarios. This study would include (1) a combine statistics method and a test statistics based on likelihood function, (2) calculated the traditional method in pooling point estimator and standard error, (3) compare the results in (1) and (2). Then we could find out the shortcoming of traditional method.

利用文字探勘技術建立憂鬱症風險預測模型

王彥雯*、葉俊怡

淡江大學數學學系

摘要

憂鬱症在現代是一個被高度關注的心理疾病，根據 WHO 的統計全球大約有三億左右的憂鬱症患者，而重鬱症也會帶來許多嚴重的影響，如：自殺，因此，就公共衛生的角度而言，如何預防憂鬱症與早期偵測憂鬱的心理狀態，將是一個重要的課題。

在本演講中，我們將介紹利用文字探勘的技術，解析短文，進行文字的情感分析，並依此發展出憂鬱症的風險預測模型。我們同時將所提出的方法應用於分析具有憂鬱症的韓國創作型歌手所創作的 87 首歌詞，透過其創作的歌詞了解歌手憂鬱症狀的變化情形。

我們希望所提出的憂鬱風險預測模型，將來可以應用於短文的分析，如：日記、部落格、留言、歌詞等，了解寫作者的憂鬱情緒程度，並利用此作為憂鬱症狀變化的偵測，以達到公共衛生早期偵測、早期自殺預防的目的。

關鍵詞：文字探勘、情感分析、憂鬱症

Calibration for Computer Experiments with Binary Responses and Application to Cell Adhesion Study

Chih-Li Sung (宋治立)*
Michigan State University

Ying Hung
Rutgers, the State University of New Jersey

William Rittasec, Cheng Zhu, C. F. J. Wu
Georgia Institute of Technology

摘要

Calibration refers to the estimation of unknown parameters which are present in computer experiments but not available in physical experiments. An accurate estimation of these parameters is important because it provides a scientific understanding of the underlying system which is not available in physical experiments. Most of the work in the literature is limited to the analysis of continuous responses. Motivated by a study of cell adhesion experiments, we propose a new calibration method for binary responses. This method is shown to be semiparametric efficient and the estimated parameters are asymptotically consistent. Numerical examples are given to demonstrate the finite sample performance. The proposed method is applied to analyze a class of T cell adhesion experiments. The findings can shed new light on the choice of settings of kinetic parameters in single molecular interactions.

關鍵詞：Cell biology, Computer experiment, Kriging, Single-molecule experiment, Uncertainty quantification

Planning of Accelerated Degradation Tests Based on Tweedie Degradation Models

I-Chen Lee (李宜真)
National Cheng Kung University

Abstract

Accelerated degradation tests are widely used to access the lifetime information of highly reliable products. To obtain the more accurate prediction of lifetime information, how to design an efficient experiment under the limited budget is a critical issue for reliability analysts. Many literatures have addressed this problem and indicated that a two-level design is the optimum strategy for an ADT plan. Their strategies were proposed under the assumption of the numbers of measurements within a degradation path are equal for all stress levels. However, many real applications showed that the numbers of measurements are not necessary to be the same for all stress levels. In this study, we release the assumption of the numbers of measurements are equal, and arrange an efficient ADT plan so that the asymptotic variance of a prediction can be minimized based on the Tweedie degradation model.

OPTIMAL DESIGNS FOR NETWORK EXPERIMENTATIONS WITH UNSTRUCTURED TREATMENTS

張明中*、潘建興

National Central University、Academia Sinica

黃靖雯

National Tsing Hua University

摘要

Experiments on connected units are commonly conducted in many scientific fields. An experimental unit in these applications may connect with some others, and the treatment applied to a unit has an effect, called a network effect, towards the responses resulted in the neighboring units. Designing such experiments was rarely discussed in the literature. Parker, Gilmour, and Schormans (2017) initiated a study of As-optimal designs on connected experimental units with unstructured treatments, assuming that the network effects are unknown constants. This work investigates in a similar design problem but the network effects are assumed to be random effects, which lead to a property that the responses of two units are correlated if some neighbors of one unit and those of the other receive the same treatment. Alphabetical optimality criteria are considered for selecting good designs with high efficiency of estimating the treatment effects and/or high accuracy of predicting the network effects. We provide theoretical conditions for designs to be optimal and illustrate our theory with some numerical examples.

關鍵詞：Network effect, social network, mixed-effect model, A-, D-, E-, (M.S)-optimality

De-aliasing in two-level factorial designs: a Bayesian approach

Ming-Chung Chang
National Central University

Abstract

Given limited resources for conducting follow-up trials, the inability to separate aliased factorial effects hinders the ubiquitous practicality of regular fractional factorial designs in the analysis of experiments. Wu (2015) proposed a frequentist remedy for “de-aliasing” aliased effects by using conditional main effects. Although Su and Wu (2017) systematized the remedy to make it easy to implement, it might miss truly active effects. Missing active effects can be a severe drawback if the purpose of experimentation is to determine the mechanism of a process rather than to make predictions. In this paper, we propose a Bayesian remedy for de-aliasing in two-level regular factorial designs. Through numerical studies, we show that our method can yield desirable model fittings and reliable de-aliasing results.

Key words: conditional main effect; conditional model; regular design; Gaussian process

Application of Bayesian Spatiotemporal Varying Coefficients to the Mortality from Ischemic Stroke in Taiwan, 2004 – 2012

Kuo-Jung Lee

Department of Statistics, National Cheng Kung University

A Bayesian spatiotemporal generalized linear regression with varying-coefficient model is proposed to examine geographic variation of the medical prescription use for ischemic stroke and to have a clearer understanding of association of the relevant risk factors such as comorbidities, medication and environmental society with mortality rate in ischemic stroke. The spatial heterogeneity of coefficients for important factors that may affect the mortality in ischemic stroke across 349 townships in Taiwan. By applying spatial-temporal models, we can understand the spatial variation in risk to ischemic stroke. It then turns out that, based on the findings, we can properly arrange medical resources and reduce the life-threatening damage caused by the uneven distribution of medical resources. The data was collected from a retrospective cohort study using 2004-2012 National health insurance research database.

Implicit Copulas from Bayesian Regularized Regression Smoothers

Nadja Klein

School of Business and Economics, Humboldt-University of Berlin, Germany

Michael Stanley Smith*

Melbourne Business School, University of Melbourne, Australia

ABSTRACT

We show how to extract the implicit copula of a response vector from a Bayesian regularized regression smoother with Gaussian disturbances. The copula can be used to compare smoothers that employ different shrinkage priors and function bases. We illustrate with three popular choices of shrinkage priors- a pairwise prior, the horseshoe prior and a g prior augmented with a point mass as employed for Bayesian variable selection, and both univariate and multivariate function bases. The implicit copulas are high-dimensional, have flexible dependence structures that are far from that of a Gaussian copula, and are unavailable in closed form. However, we show how they can be evaluated by first constructing a Gaussian copula conditional on the regularization parameters, and then integrating over these. Combined with non-parametric margins the regularized smoothers can be used to model the distribution of non-Gaussian univariate responses conditional on the covariates. Efficient Markov chain Monte Carlo schemes for evaluating the copula are given for this case. Using both simulated and real data, we show how such copula smoothing models can improve the quality of resulting function estimates and predictive distributions.

Key words: implicit copulas, regularized regression smoother, Markov chain Monte Carlo

Bayesian Variable Selection of Heteroskedastic Models with Realized Exogenous Variables

Feng-Chi Liu (劉峰旗)
Department of Statistics, Feng Chia University

摘要

While high frequency data with intraday information have been demonstrated that are useful for forecasting daily risk, we incorporate realized variances (RVs) in GARCH models to capture more intraday information for accurate volatility forecasts. The RVs are calculated from intraday information, such as 5-minutes intraday returns, of different exogenous variables. This is the first purpose of this study. The second purpose is that we focus on the variable selection of GARCH models with incorporating a number of RVs of different exogenous variables. A Bayesian stochastic search variable selection (SSVS) method of George and McCulloch (1995) are used for the variable selection of GARCH models. Thus, the estimation of model parameters and the best subset of GARCH model are simultaneously obtained by the designed Markov chain Monte Carlo (MCMC) sampling. The proposed method is evaluated by some simulation studies and a real application of Taiwan stock market.

關鍵詞：realized variance, GARCH model, stochastic search variable selection (SSVS) method, Markov chain Monte Carlo (MCMC)

Using a generalized area-based truncated model to resolve Fisher’s paradox when extrapolating biodiversity

Prof. Youhua Chen

Chengdu Institute of Biology, Chinese Academy of Sciences, China

Abstract:

Why are so many tropical tree species hyper-rare? Do they really have only one or two individuals on Earth? This question, the so-called Fishers paradox, was put forward by S.P. Hubbell when applying Fishers logseries to estimate tropical tree diversity. Herein, we developed an area-based truncated Fishers logseries model to partially, if not completely, resolve Fishers paradox by assuming that the occurrence of too-rare species is impermissible globally while being possible at local scales due to limited sampling efforts. An empirical test showed that alternative truncated models were indistinguishable at the local forest-plot scale, but they could be told apart at the regional scale. By comparison, a protracted speciation neutral model had similar behaviors. However, the exceptional merit of the truncated model is that by using a small truncation threshold, the prediction of regional species richness was similar to the value predicted by the original Fishers logseries, while completely excluding the possibility of the occurrence of too-rare species. Given the issue of the inability to distinguish among alternative models at the local scale, the truncation threshold might be pre-set by referring to real-world population sizes of trees. Alternatively, the threshold can be estimated if sufficient local biodiversity data are provided.

A New Metric for the Analysis of the Scientific Article Citation Network

Junji Nakano
Institute of Statistical Mathematics, Japan

Abstract:
Citation plays an important role in the bibliometrics analysis since the introduction of the impact factors, but traditional measures focused only on the direct citations between articles. In this work, we introduce a new metric, namely Article Network Influence (ANI), to measure quantitatively the influence of an article in the whole Web of Science or its own research community. We demonstrate the use of this new metric on the analysis of article citation network in statistics research community. We identify the main differences between the new metric and several traditional measures, including the impact factor, pagerank and FWCI.

On inference for the generalized Pareto distribution

Hideki Nagatsuka
Chuo University, Japan

Abstract:
Statistical modeling of the largest or smallest values (extreme values) of certain natural phenomena, e.g., waves, floods, earthquakes, winds and temperatures, is of interest in various practical applications. The generalized extreme value distribution (GEVD) is the only possible limiting distribution for normalized extreme values, derived by the extreme value theory (EVT), and considered to be an appropriate model for the block maxima. However, there has been some criticism since using only maxima leads to the loss of information contained in other values. This problem is remedied by considering some largest values in the given period instead of the largest value, that is, exceedances over the threshold. The generalized Pareto distribution (GPD), which is the limiting distribution for exceedances over the threshold, and offers a unifying approach to the modelling of such values. Although the GPD is useful and frequently used in the extreme value statistics, it is well known that the inference for the GPD is a difficult problem. In this talk, we address some challenging problems in inference for the GPD and introduce a new general framework of the inference for the GPD, based on a likelihood function.

Recent developments of the volume-of-tube method

Satoshi Kuriki
Institute of Statistical Mathematics, Japan

Abstract:

The tube method (the volume-of-tube method) is the methodology to approximate the upper tail probabilities of the maxima of Gaussian random fields. It is equivalent to the expected Euler-characteristic heuristic applied to Gaussian random fields. We overview the idea, history, and its applications briefly. It has many applications in multiple comparisons. In particular, two topics on recent developments are highlighted: (i) Approximations of the distributions of the largest eigenvalues of random matrices; and (ii) Construction of simultaneous confidence bands in various regression models.

運用文字探勘分析政府政策推動與民意發展－以限塑政策為例

林彩玉、許書華*

逢甲大學應用數學系

摘要

社群與自媒體的蓬勃發展，透過社群網站以隱匿身分的方式討論社會議題，致使人們更加願意表達對議題的看法，進而提供不同面向的思維。日積月累之下，不論任何議題或事件，這類型的文字評論大量的增加。然而，這樣的民意對於政府政策的推動，到底有何影響？本研究以台灣歷史最為悠久的佈告欄網站（BBS）中的批踢踢實業坊之Gossiping版與限塑政策相關議題之文章為例，探討政府政策推動與民意發展。期望透過文字及情緒詞彙分析，清楚地描繪出民眾的態度與意見。本研究首先透過文字探勘的詞頻分析，找出民眾在限塑議題最常討論的詞彙；其次，透過關聯分析找出與限塑議題關聯性較高的詞彙；接著，透過主題建模找出限塑政策議題潛在的16個主題，並經由脈絡分析找出以「環保」一詞為例的主題網絡，做出詳細的探討。最後，由情緒詞彙分析來探討民意，發現在此議題得到的正面回應比例居高。藉此研究期望能夠透過民眾對政策意見態度表示，提供政府溝通政策推動等相關建議。

關鍵詞：主題建模、脈絡分析、詞頻分析、情緒詞彙分析。

Semi-parametric version of a tolerance interval strategy for pair trading of multiple assets

Fong-Yi Syu*(許鳳儀) Tsai-Yu Lin(林彩玉)
Department of Applied Mathematics, Feng Chia University

Cathy W.S. Chen(陳婉淑)
Department of Statistic, Feng Chia University

Abstract

Pair trading is a widely-used market neutral strategy and also a statistical arbitrage method that provides investors the ability to select two assets with similar trends in their historical data in order to gain low-risk profits. Based on this concept, the present paper extends two assets into multiple assets and applies them to the semi-parametric version of a tolerance interval. This method combines volatility forecasting via the EWMA (exponentially weighted moving average) model with the traditional nonparametric tolerance interval. This study selects five AI (artificial intelligence) stocks in the U.S. equities market to target profitability through a pair trading strategy of multiple assets from 2017 to the first half of 2018 with each month as the starting point, for a total of eighteen out-of-sample periods within a six-month time.

Keywords: nonparametric tolerance interval; volatility forecasting; EWMA (exponentially weighted moving average) model

Comparing GOF Tests for Degradation Models across Multiple Time Points

Tan Yung Hui and Shuen-Lin Jeng

National Cheng Kung University

Abstract

The purpose of this study is to compare the powers of the goodness of fit (GOF) tests for the non-homogeneous discrete compound Poisson (NHDCP) models under degradation processes. The degradation events are assumed to follow a non-homogeneous Poisson processes and the degradation increments may have Bernoulli or Poisson distribution. Three GOF tests are considered: Watson (W), Cramér–von Mises (CM), Anderson-Darling (AD). We use a real data set from software reliability experiment to illustrate the process of the comparisons. We explore the powers of the hypothesis testing for the cases that the specified model is placed as the null hypothesis or as the alternative hypothesis. The GOF tests are performed at multiple time points of the degradation process. The interesting discovery of this research is that under the types of the real data set, the power of the GOF tests is higher when the specified model with Poisson increment is placed as the null hypothesis than the case that the specified model with Bernulli increment is placed as the null hypothesis.

Keyword: Power, GOF Tests, Degradation Model, Non-homogeneous Discrete Compound Poisson

雙眼分視視知覺電腦化測驗之研發與應用

[1] 李紀蓮* [2,4] 黃碧群 [1,2,4] 鄭兆堅 [3] 劉峻正 [1,3,5] 杜旻育 [1,2] 張壹婷 [1] 林信宏#

[1] 國軍高雄總醫院岡山分院航訓中心　[2] 國立成功大學心理系 [3] 國軍高雄總醫院岡山分院
[4]國立成功大學心智科學與應用博士學程，[5]弘光科技大學生物醫學工程系

摘要

前言：配合國軍阿帕契飛行部隊的成軍，因應該機型雙眼分視特殊飛行作業的環境，希冀研發合宜的教育訓練輔助裝備，以提升教育訓練成效。

目的：進行雙眼分視特殊作業情境模擬實驗裝備的設計，並開發電腦化測驗軟體，規劃適宜的實驗流程，建立測驗的信效度，以及探討不同優勢眼在雙眼分視作業上表現的差異。

材料及方法：以視知覺理論為基礎，規劃「光點移動方向」及「E」字母開口方向辨識為測驗內容，雙眼作業模式則設計左眼看遠方(106 公分)大螢幕，右眼看近處(16 公分)小螢幕，實驗程序每人實施四次不同測驗，並於測驗三時在左眼前方加置減光片，大螢幕亮度由 20.68cd/m^2 降低至 0.13cd/m^2，小螢幕亮度維持 186cd/m^2 模擬雙眼競爭情境。

結果：(1)樣本計 60 人。男性 55 人(91.7%)，女性 5 人(8.3%)。平均年齡 26.4±2.6 歲。裸視及戴鏡矯正後視力均在 0.6 以上，優勢眼為右眼者 38 人(63.3%)，為左眼者 22 人(36.7%)。(2)不同優勢眼(左眼或右眼)對大小螢幕表現的影響：運動方向題型，測驗一至四，大小螢幕平均正確率，不同優勢眼無顯著差異，平均正確率為 92%；「E」字母題型，測驗二、四，大小螢幕平均正確率，不同優勢眼亦無顯著差異，平均正確率為 96%，但在測驗三，左眼前方加置減光片降低亮度時，優勢眼為左眼者較右眼者為佳，達顯著差異($p<.05$) ($\underline{M}$=77.1% vs. $\underline{M}$ =63.8%)。(3)個人四項測驗重複量數分析，運動方向題型，測驗一至四，大小螢幕平均正確率，分別為 93.9%、94.7%，無顯著差異。「E」字母題型，測驗二、四，大小螢幕平均正確率在 96%以上，但在測驗三時大螢幕則降為 68.7%。(4)在測驗的反應時間，整體而言，優勢眼為左眼或右眼，二者多無顯著差異，平均時間在 1.06 秒～1.35 秒之間。惟在測驗三亮度改變時，大螢幕反應時間略升至 1.88 秒~2.35 秒之間。

結論：順利完成雙眼分視「電腦化測驗軟、硬體」的研發，受試者在測驗的表現，正確率多在九成以上，測驗難度不高，可有效體驗左右眼切換的作業模式；而優勢眼為左眼或右眼，無顯著差異；另外模擬雙眼亮度差異大時的情境，左眼正確率降低至 68%，可明顯體驗左右眼亮度差異過大時，雙眼競爭效應的感受。研發成果可納入阿帕契飛行員先期教育訓練課程規劃，提升飛訓教育成效。

New insights on “A semiparametric model for wearable sensor-based physical activity monitoring data with informative device wear”

Weng, Hsu-Wei(翁許瑋)* 、Dr. Huang, Shih-Hao(黃世豪)
National Central University

摘要

In this article we indicate and solve two problems found in Song et al.(2018). First, we characterize the time-scale invariant property of the semiparametric model they utilize, and show that their R package provides different result after time scaling. Second, we find the setting of their main simulation does not follow the assumptions of the utilized model, which may lead to an unreliable conclusion. We provide corrected R code for practical use, and give an appropriate simulation setting to illustrate the behavior of the utilized model in a real-data example.

Keywords:
augmented estimating equation, time-scale invariant, semiparametric regression

房價外溢效果-以台灣和中國城市為例

廖苡淳* 李美杏

台北大學

摘要

本文探討台灣和中國房地產市場是否具有外溢現象，採用 Diebold and Yilmaz(2010)提出的方法以 VAR 模型以及預測誤差變異數分解計算外溢指數，分別探討兩個地區房價是否有外溢連動性。台灣選取台北市、新北市、桃園市、新竹市、台中市、高雄市的房價指數；中國地區則選取一線城市的上海、北京、深圳，新一線城市的重慶、南京、西安，二線城市的廈門、大連共 8 個城。實證結果顯示台灣以台北市最具影響力，新北市以及台中市較容易受影響，又以新北是受外溢影響最為顯著；中國則以南京和上海影響力大，廈門和北京易受影響。

關鍵詞：

外溢效果、向量自我回歸模型、房價指數、衝擊反應函數、預測誤差變異數分解