產業智慧化關鍵-智慧製造實務應用及未來趨勢

蕭哲君　總經理

采威國際資訊股份有限公司

摘要

製造業面對「少量多樣化」的市場需求，很容易陷入「規模不經濟」的困境，也就是產線更忙，但並未創造更好效益。

為解決多數中小企業長期以來不易管控但又想在生產端擠出效益，

本次將講解如何透過數位化的資料蒐集、模擬及分析工具，以可視化方式真實呈現，讓管理者能掌握到現場潛在問題，找出有助提升稼動率、提高生產良率的關鍵因子。整合各式工具機 NC、PLC 系統等，將生產設備進行大量資料的蒐集、生產稼動的監控，另透過采威來進行 EPR、MES 等跨域系統的整合後，甚至可以做到設備的預防保全及耗用品的更換預警，以達到高效益的生產運作目標。

關鍵詞：跨域系統整合，大數據，自動化設備。

製程分析與最佳化技術

藍坤銘
工業技術研究院 巨量資訊科技中心

摘要

製程工程師多憑領域知識與經驗法則進行製程品質預測及製程配方參數調整，但過程耗時費力且工程師的經驗不易累積成為公司的研發資產，因此本技術建立由製程參數快速估測產品品質特性的能力，可與製程工程師透過互動協作模式，縮短研發時程，提供一個符合產業高效率的製程參數最佳化系統。本次分享內容將先針對產業問題與需求進行簡單描述後，針對目前的核心技術發展現況進行說明以及該技術在不同應用中的案例分享。

關鍵詞：
製程分析、參數最佳化、品質預測、製程參數優化

大數據下的行為分析

謝一平

思為策略股份有限公司　共同創辦人

社群網站是我們這個時代人與人交流的 DNA，我們非常的幸運，開始有電腦、網際網路的發展，從 1996 年開始，有撥接網路，我們可以看見資訊科技的便利， 訊息傳遞的方式也開始有重大的變化，Web1.0 的時代，以現在的觀點就像遠古時代，我們僅重視訊息的佈達，不太重視回饋，就是是一個單向訊息傳送，就跟電視一樣。

可以這麼說，網站、BBS 就是所謂的佈告欄，我們看見過去 20 年來在資本市場的投入下，訊息傳遞的走向更有效率的境界，那個年代有了 Google，但是基本上來說那僅是做好訊息的傳遞的基本功。

人與人的互動不該僅是如此，我們的世界走向了 Web2.0，開始有部落格的出現像是無名小站，開始有網路遊戲像是世紀帝國、爆爆王、天堂，我們可以看見那時候所有的網路互動走向更多元， 資訊科技的發展不僅是能夠讓我們可以更有效率的提供訊息，其實也更有效率收集每一個使用者瀏覽足跡，Google Analytics 存在的意義與價值。

接著包含 RSS 訂閱等服務出現，Twitter、Facebook 等社群網路也被發展出來，甚至在最近的十年發展出以人為核心的推薦系統，我們可以看見訊息走向個人化，走向更破碎化。訊息傳遞平台的轉移，也陸續插旗社群網站。

在這股浪潮中，最特別的應用莫過於政治，特別是在選舉。2008 年歐巴馬第一次競選美國總統，就開始分析 Twitter 的資料，協助制訂溝通策略。這一步，成為了大數據引入全球政治的濫觴。但台灣直到 2014 年底柯文哲選當選台北市長，才讓政治工作者或多或少注意到如此的現象，並開始採取行動。

思為策略成立五年多來，在公開數據上發展解析「行為」的分析服務，但世界越來越重視隱私權，起因於 2016 年美國總統大選的隱私權爭議，劍橋分析、以及接著制定的 GDPR，造成了一波海嘯，帶給資料分析產業最黑暗的一年。但我們並沒放棄努力，除了堅守我們自己的道德底線：不作假資料、不透過非法方式取得資料外，我們也持續開發新的分析方式和模型，以期能把「行為」分析服務，再往更前方推進。

Efficient Estimation of Linear Regression Models for Combining Internal Small Data and “External” Big Data

Wei Yann Tsai
Columbia University Mailman School of Public Health

Chatterjee, Chen, Maas and Carroll (CCMC) (2016) have identified and investigated important problems on combining information from two data resources. They proposed and studied a constrained semiparametric maximum likelihood method for regression models based on individual-level data from "internal" study while using summary-level information from "external" big data source. In this article, we proposed and studied a constrained moment type estimator of the parameters in linear regression model by combining information from "internal" and "external" data source. Asymptotic theory and variance estimators are developed. If the data distributed according multivariate normal distribution, then the proposed estimator achieves the constrained Cramer-Rao lower bound. Simulation study also showed that the proposed method is more efficient than its competitor.

結合電子健康紀錄資料與精準資料的統計分析

程毅豪

中央研究院統計科學研究所

摘要

大型電子化健康紀錄資料庫如健保資料庫分析在近年的醫學研究中日漸普及。這樣的研究具有省時省力，且可避免回溯性研究之回憶偏差等優點。然而此類研究的重要侷限之一，是這些大型資料庫的資料收集並非針對學術研究目的，因此其往往缺乏較詳盡的關於個人之干擾因子(confounder)如吸菸、飲酒、飲食習慣與職業暴露等，以及生物標記(biomarker)測量如血壓血糖等資訊，因此這些資料庫無法產出較精準的研究成果。一個解決此問題的方法是設法由一些專門的研究或調查資料庫中取得較精準的資料，其包含了干擾因子及生物標記資料。 此精準資料可提供校正了干擾因子及生物標記資訊的研究結果。但相較於前述的大型資料庫，此類專門性的研究或調查資料庫往往所收集的個案數規模小了許多，因而影響其統計上的效力 (power)。我們將介紹如何結合電子健康紀錄資料與精準資料的統計分析方法，期能將兩種類型資料截長補短，進行適當的結合，以提供更精準的分析結果。

Testing Monotonicity of Density via a Nonparametric Likelihood Ratio

Gary Chan
University of Washington

Hok Kan Ling
Columbia University

Chuan-Fa Tang*
University of Texas at Dallas

Phillip Yam
Chinese Hong Kong University

Abstract

We study the likelihood ratio test statistic for a hypothesis testing problem on whether a random sample follows a distribution with a nonincreasing density function. The obtained test statistic has a surprisingly simple asymptotic null distribution, which is Gaussian, instead of the well-known chi-square for generic likelihood ratio tests. We further suggest testing procedures based on the least favorable configuration and bootstrap to improve the power of detecting departures of the underlying density from being nonincreasing.

Keywords and phrases: Grenander estimator, Shape constraints, Uniform spacings, Renyi's lemma, Kullback–Leibler divergence, Pseudo-true density, Donsker's class, Hellinger distance, Entropy, Bracketing number, Bernstein norm, Bootstrap.

Glucose variability and risk of peripheral arterial disease in persons with type 2 diabetes

Tsai-Chung Li

Department of Public Health, College of Public Health, China Medical University

Chun-Pai Yang

Department of Neurology, Kuang Tien General Hospital

Cheng-Chieh Lin

Department of Family Medicine, China Medical University Hospital

摘要

Objective The relationship between glycemic variability and peripheral artery disease (PAD) in patients with type 2 diabetes mellitus (T2DM) is unclear. This study was to investigate whether visit-to-visit variations in fasting plasma glucose (FPG), as measured by the coefficient of variation (CV) were associated with the risk of PAD, regardless of HbA1c and other cardiovascular risk factors in T2DM patients.

Methods T2DM patients enrolled in the National Diabetes Care Management Program during the period of 2002-2004, ≥ 30 years of age and free of PAD (n = 30,932) were included and monitored until 2011. Cox proportional hazards regression models were implemented to analyze the related factors.

Results During an average 8.2 years of follow-up, a total of 894 incident cases of PAD were identified, with a crude incidence rate of 3.53/1000 person-years (3.63 for men, and 3.44 for women). After adjustment for socio-demographic factors, lifestyle behaviors, diabetes related variables, drugs related variables, FPG and HbA1c, and comorbidities, both FPG-CV and HbA1c were significant predictors of PAD, with corresponding hazard ratios (HRs) of 1.24 (95% CI 1.04–1.47), for FPG-CV in the third tertile and 1.50 (95% CI 1.10 – 2.04) for HbA1c 10%. This finding maintained consistency after excluding potential confounders in the sensitivity analysis, further validating the results.

Conclusions FPG-CV and HbA1c 10% were potent predictors of PAD in T2DM. The associations between HbA1c, glycemic variability, and PAD suggest a linked pathophysiological mechanism, which may play a crucial role in clinical managements and therapeutic goals in preventing PAD in T2DM.

關鍵詞：

HbA1c; fasting plasma glucose; glycemic variability; peripheral artery disease

Improved analysis for the means of several gamma distributions

Shu-Hui Lin
National Taichung University of Science and Technology, Taiwan
Email:suelin@nutc.edu.tw

Abstract

The gamma distribution is widely used distributions for modeling the real data set because the gamma distribution not only can be treated as the generalized exponential distribution, but also is flexible to fit the positive and right-skewed data. Statistical techniques generally have somewhat difficult to develop for the gamma distribution, partly because its parameters are not of the traditional location-scale type family, therefore the maximum likelihood estimates (MLEs) are not available in closed-form, and can only evaluated numerically. Therefore, in this paper, we focus on providing improved analysis for the means of several gamma distributions under multiple populations and small sample sizes. We derive five pivotal quantities from the methods including likelihood ratio method, signed likelihood ratio method, revised signed likelihood ratio method, modified signed likelihood ratio method, and computational approach method and then those pivotal quantities will be used to test the equality of several gamma means and further the common mean. We also applied those methods to the numerical studies to illustrate their applications, limitations and reliability.

Keywords: common mean; gamma distribution; improved analysis; mean parameter; maximum likelihood estimator; pivotal quantity.

A robust clustering approach for piecewise regression models

呂岡玶*

國立臺中科技大學應用統計系

張少同

國立臺灣師範大學數學系

Abstract

Piecewise regression models are applied in many substantial areas including econometrics, ecology, and meteorology. Change-point detection is important for knowing the changes in the data structure. Identifying change-points is like grouping data into clusters of similar individuals. Fuzzy clustering is powerful and suitable for change-point problems because segmented boundaries are often indefinite in reality. We present a robust clustering approach for piecewise regression models. We propose to connect the fuzzy change-point algorithm with the M-estimation method for robust estimations. We embed the fuzzy c partitioning into the piecewise regression models so that we can implement the fuzzy c-regressions and fuzzy c-means clustering to estimate the change-points and regression parameters simultaneously. For making the proposed method robust to outliers and heavy-tailed distributions, we adopt the M-estimation with a robust criterion as the distance measure for estimating regression parameters robustly. A robust algorithm is constructed by using the popular Tukey's biweight function. A simulation study is conducted to show the effectiveness of the proposed method. The simulation results actually show the proposed approach is robust to outliers and heavy-tailed distributions.

Keywords: change-point, regression models, outliers, robust

Robust fitting of change-point regression models

張少同*
國立臺灣師範大學數學系
呂岡玶
國立臺中科技大學應用統計系

Abstract

Change-point regression models are widely applied in many areas such as finance, medicine, and meteorology. The existing methods for fitting change-points regression often assume normal distributions for regression errors and estimate the regression parameters by the maximum likelihood estimate (MLE). The normality-based MLE are sensitive to outliers and heavy-tailed distributions. In reality, data often contain groups of observations having longer than normal tails or atypical observations, the normal assumption can considerably influence the fit of regression models. Least absolute deviation (LAD) has been widely employed in robust estimations. Because of the robust property of the LAD procedure and the connection of LAD with the likelihood approach for regression models with errors following a Laplace distribution, we propose a robust likelihood approach using Laplace distributions for change-point regression models. The change-point regression model is first converted into a mixture regression model by treating change-points as latent class variables. Thus, a sample from a change-point regression system is incomplete with the data of change-points unobserved. Converting the probability distributions of change-point collections into the memberships of data belonging to respective regimes, we then implement the Expectation and Maximization algorithm to estimate change-points and regression parameters simultaneously. The proposed Laplace-based approach is much more robust to outliers and heavy-tailed errors than the normal-based method. The effectiveness of the proposed method is demonstrated through a simulation study.

Keywords: robust, change-point, regression models, Laplace distributions

Goodness-of-fit Statistics for Polytomous Regression Models for Ordinal Responses with Natural Link

Wei-Hsiung Chao (趙維雄)
Department of Applied Mathematics, National Dong Hwa University

Abstract

Polytomous regression models are often fitted through the use of maximum likelihood to study the relationship between a categorical response and some covariates. To assess the fit of these models with only categorical covariates, it is appropriate to use the Pearson-Fisher's test for product multinomials. In the presence of continuous covariates, there exist ad-hoc statistics of Pearson-Fisher's type based on grouping strategies for assessing the adequacy of the fitted models. These statistics are often formed as a sum of Pearson's statistics over all groups in which the within-group observations are in fact heterogeneous. No asymptotic result was available to show that these ad-hoc statistics are chi-squared distributed when the fitted model is correct. Under certain situations, these ad-hoc statistics were found to be chi-squared distributed through simulation studies. Without using any grouping strategies, we recently proposed a Pearson-like statistic W based on pooled observations that is useful in assessing the fit of a polytomous regression model with non-natural link. The W statistic is a quadratic form in the differences between the observed totals and fitted totals over response categories. Under certain rank condition, the asymptotic null distribution of the proposed statistic was shown to be chi-squared with appropriate degrees of freedom. Because the adjacent logit link is the natural link for ordinal response, these differences are identical to zero when fitting a multinomial logistic regression model to the data. Thus, the W statistic has no value in assessing the goodness of fit of multinomial logistic regression model. To overcome this difficulty, we extend the W statistic to the W_G statistic which involves partitioning the covariate space into groups, and show that it is asymptotically chi-squared distributed.

Keywords: Goodness of fit, Pearson's chi-square test, ordinal data, adjacent logit link

Nonparametric Analysis of Activity Profiles From Physiological Data

Hsin-Wen Chang
Institute of Statistical Science, Academia Sinica

Ian W. McKeague
Department of Biostatistics, Columbia University

摘要

摘要內文

This talk develops a unified nonparametric inference framework for analyzing physiological sensor data collected from wearable devices. We introduce a nonparametric likelihood ratio approach that makes efficient use of the activity profiles to provide a confidence band for their means, along with an ANOVA type test. These procedures are calibrated using bootstrap resampling. A simulation study shows that the proposed procedures outperform competing Wald-type functional data approaches. We illustrate the proposed methods using wearable device data from an NHANES study.

關鍵詞：

Accelerometry, Empirical likelihood, Functional data

Adaptive treatment allocation for clinical studies with recurrent events data

Pei-Fang Su (蘇佩芳)
Department of Statistics, National Cheng Kung University, Tainan, Taiwan

Abstract

In long-term clinical studies, recurrent event data are collected and used to contrast the efficacies of two different treatments. The event re-occurrence rates can be compared using the popular negative binomial model, which incorporates information related to patient heterogeneity. For treatment allocation, a balanced approach in which equal sample sizes are obtained for both treatments is predominately adopted. However, if one treatment is superior, it may be desirable to allocate fewer subjects to the less effective treatment. In order to accommodate this objective, the sequential response-adaptive treatment allocation procedure is derived based on the doubly adaptive biased coin design. Our proposed treatment allocation schemes have been shown to be capable of reducing the number of subjects receiving the inferior treatment while simultaneously retaining a comparable test power level to that of a balanced design.

Keywords：Adaptive design; Doubly adaptive biased coin design; Negative binomial model; Allocation rule

The weighted concordance correlation coefficient estimated by variance components for longitudinal overdispersed Poisson data

蔡秒玉 (Miao-Yu Tsai)

彰化師範大學統計資訊所

摘要

The concordance correlation coefficient (CCC) can be used to assess agreement among multiple observers for continuous and discrete responses. We can consider not only subject, observer and time effects, but also interaction effects in extended three-way generalized linear mixed-effects models (GLMMs) for repeated measurements from longitudinal Poisson data. The variance components (VC) approach has been proposed to measure intra-agreement for each observer and inter- and total agreement among multiple observers simultaneously under extended three-way Poisson mixed-effects models. However, under VC, a model including all potential explanatory variables may lead to biased parameter estimates. To overcome this problem, the estimation of CCC using the VC approach, as well as applying the corrected conditional Akaike information criterion (CAICC) and corrected conditional Bayesian information criterion (CBICC) measures for model selection are adopted. Furthermore, we develop agreement coefficients under the VC approach for overdispersed count data. Simulation studies are conducted to compare the performance of VC with and without model selection via CAICC and CBICC for longitudinal Poisson and overdispersed Poisson data sets. The results show that performing CAICC and CBICC model-selection procedures in VC models yields small mean square errors and nominal 95% coverage rates.

Keywords: Agreement; CAICC; Concordance correlation coefficient; Generalized linear mixed-effects models; Variance components

Association Test of Copy Numbers Variations via a Bayesian Procedure in Next Generation Sequencing

Yu-Chung Wei

Department of Statistics, Feng Chia University, TAIWAN

Copy number variations (CNVs) are genomic mutations consisting of abnormal numbers of gene fragment copies. Current algorithms for CNV association study for whole genome sequencing are restricted to a specific size or common/rare CNVs. In this study, a Bayesian marker-level testing procedure to detect disease-associated CNVs is constructed. First, the absolute copy number of each window is estimated from sequencing read depths for every sample. And then the absolute copy numbers from case and control are compared to select candidate disease-associated windows. Finally, the information from neighboring windows is combined to identify the disease-associated copy number regions. We evaluate the performance and compare with competing approaches via simulations and real data.

Key words: association study, Bayesian inference, copy number variation, next generation sequencing

A Construction of Cost-Efficient Designs with Guaranteed Repeated Measurements on Interaction Effects

Frederick Kin Hing Phoa
Institute of Statistical Science, Academia Sinica, Taipei 115, Taiwan.
January 21, 2019

Abstract: This work introduced a useful class of cost-efficient designs for two-level multi-factor experiments. It provided guaranteed repeated measurements on all 2-tuples from any two factors and the number of repetitions was adjusted by the experimenters. Given the number of factors of interest, it utilized less resources than an orthogonal array while its repeated measurement provided a resistance towards outliers that a covering array failed to achieve. To bridge the wide spectrum between two extreme settings (orthogonal arrays and covering arrays) in terms of the number of repeated measures of tuples, we developed a systematic method to construct families of these designs, namely (supersaturated) repeated coverage design, with small run sizes under different number of factors and number of repetitions. This is a joint work with Dr. Yasmeen Akhtar (Arizona State University, USA).
Keywords: Repeated Coverage Designs, Orthogonal Array, Covering Array, Cost Efficiency

Stochastic search variable selection for definitive screening designs in split-plot and block structures

Chang-Yun Lin

Department of Applied Mathematics and Institute of Statistics,

National Chung Hsing University, Taichung, Taiwan, 40227

Abstract

Split-plot definitive screening (SPDS) and block definitive screening (BDS) designs have been developed for detecting active second-order effects in screening experiments when split-plot and block structures exist. In the literature, the multistage regression (MSR) and forward stepwise regression (FSR) methods were proposed for analyzing data for the two types of designs. However, there are some limitations and potential problems with the regression approaches. First, the degrees of freedom may not be large enough to estimate all active effects. Second, the restricted maximum likelihood (REML) estimate for the variances of whole-plot and block errors can be zero. To overcome these problems and to enhance the detection capability, we propose a stochastic search variable selection (SSVS) method based on the Bayesian theory. Different from the existing Bayesian approaches for split-plot and block designs, the proposed SSVS method can perform variable selections and choose more reasonable models which follow the effect heredity principle. The Markov chain Monte Carlo and Gibbs sampling are applied and a general WinBUGS code that can be used for any SPDS and BDS designs is provided. Simulation studies are conducted and results show that the proposed SSVS method well controls the false discovery rate and has higher detection capability than the regression methods.

KEY WORDS: Bayesian, effect heredity, false discovery rate, generalized least squares, Gibbs sampling, Markov chain Monte Carlo, restricted maximum likelihood, WinBUGS.

D-optimal designs for Scheffé's first- or second-degree polynomial models in multi-response mixture experiments

Hsiang-Ling Hsu
Institute of Statistics, National University of Kaohsiung

Abstract

A mixture experiment in the (q-1)-dimensional probability simplex S^{q-1} is an experiment in which the q factors are non-negative and subject to the simplex restriction, which means the sum of all factors is equal to one. In this talk, we investigate the issue of the *D*-optimal designs for the considered k responses models, consisted of the Scheffé's first- or second-degree polynomial models, in mixture experiments. Initially, we characterize the structure of candidate designs for the multi-responses mixture experimental models based on the complete classes of the class of weighted centroid designs. According to the well-known equivalence theorem, we verify that the obtained allocation measures at the support points are *D*-optimal designs. The results of *D*-optimal designs in multi-response considerations are demonstrated to be independent of the covariance structure between the k responses, but depend on the allocation of the underlying first- or second-degree polynomial models.

Keywords: complete class, design optimality, exchangeability, Kiefer ordering, weighted centroid design.

On fixed effects estimation for spatial regression under the presence of spatial confounding

Yung-Huei Chiou(邱詠惠)
Department of Mathematics, National Changhua University of Education

Abstract

Spatial regression models are often used to analyze the ecological and environmental data sets over a continuous spatial support. Issues of collinearity among covariates are often considered in modeling, but only rarely in discussing the relationship between covariates and unobserved spatial random processes. Past researches have shown that ignoring this relationship (or, spatial confounding) would have significant influences on the estimation of regression parameters. To improve this problem, an idea of restricted spatial regression is used to ensure that the unobserved spatial random process is orthogonal to covariates, but the related inferences are mainly based on Bayesian frameworks. In this thesis, an adjusted generalized least squares estimation method is proposed to estimate regression coefficients, resulting in estimators that perform better than conventional methods. Under the frequentist framework, statistical inferences of the proposed methodology are justified both in theories and via simulation studies. Finally, an application of a water acidity data set in the Blue Ridge region of the eastern U.S. is analyzed for illustration. This is a joint work with Hong-Ding Yang and Chun-Shu Chen.

Keywords: Bias; Generalized least squares; Maximum likelihood estimate; Random effects; Restricted spatial regression.

A Data Driven Spatial Sampling Approach

Ching-Ru Cheng (鄭敬儒)

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

Abstract

To predict underlying surface based on the noisy spatial dataset is an important issue in spatial statistics. Generally, different sampling schemes will lead to different prediction results. In this thesis, we focus on constructing a selection procedure among different sampling schemes. From a prediction perspective, a data driven selection criterion based on the mean squared prediction errors is proposed, and a generalized degrees of freedom is investigated to evaluate the complexity of the prediction result corresponding to the utilized scheme. Validities of the proposed method are illustrated theoretically and numerically. Finally, we demonstrate the applicability of the proposed method by analyzing the groundwater dataset in Bangladesh.

Key words: Data perturbation, Generalized degrees of freedom, Mean squared prediction error, Sampling scheme, Spatial correlation function

應用空間點過程分析石門水庫集水區崩塌地資料

鄭維晉*、黃怡婷
國立臺北大學統計學系

黃金聰
國立臺北大學不動產與城鄉環境學系

摘要

由於航照與衛星技術的提升與地理資訊系統的發展，讓現今大範圍區域之空間資料取得容易。若以造山運動尺度的觀點，崩塌區域可用空間中的一點來代表，近年來陸續有學者提出以空間點過程的統計方法來分析地質事件在空間的分佈樣式。本論文採用空間點過程方法探討石門水庫集水區崩塌事件點分佈之點樣式 (spatial point pattern)，並建構合適之強度 (intensity) 模型，利用點過程之強度探討研究區域內崩塌潛勢的強弱。運用 R 軟體之 spatstat 套件，首先以樣區計數檢定 (quadrat counting test) ，發現石門水庫集水區崩塌事件點之分佈並非完全空間隨機，再藉由 K-function、pair correlation function、F-function、G-function 等點過程之摘要統計函數，檢討崩塌事件點樣式呈現有聚集的現象。因點樣式呈現聚集的特性，本論文進而建構 Thomas 集群點過程 (Thomas Cluster Process)、Matérn 集群點過程 (Matérn Cluster Process)、對數高斯 Cox 過程 (log-Gaussian Cox Process, LGCP) 等模型。其相關之參數估計，採用兩階段參數估計方式 (Two-Step Estimation)。第一階段會假設強度與空間共變量之間具有對數線性型態的函數關係，利用最大概似估計法配適非均質卜瓦松點過程模型。第二階段以最小對比度法 (Minimum Contrast Estimation) 估計集聚之參數。最後，本論文將比較三種集群模型之表現，選擇最適切於石門水庫集水區崩塌點樣式之模型，以利未來識別研究區域內之崩塌潛勢。

關鍵詞：空間點過程、非均質卜瓦松點過程、Thomas 集群點過程、Matérn 集群點過程、對數高斯 Cox 點過程

On Covariate-dependent Spatial Covariance Model

Yen-Shiu Chin* and Nan-Jung Hsu
Institute of Statistics, National Tsing Hua University

Abstract

In this talk, we introduce a nonstationary spatial covariance model in which the spatial dependence structure is affected by time-varying covariates. The proposed model is built on the spatial-temporal random effect model framework, in which a factorization-based strategy is adopted to incorporate the timely covariate information into the covariance structure of the temporal random effects. We develop maximum likelihood estimation via an EM Algorithm for parameters in the proposed model. Simulation studies show the flexibility of the proposed model to capture the spatial patterns successfully even the model is misspecified. Some advantages of using the proposed method compared to other covariate-driven approaches are discussed. An application of our methodology to $PM_2.5$ data from Taiwan Environmental Protection Administration will be presented for illustration.

Key Words: Cholesky decomposition, maximum likelihood, nonstationary spatial covariance function, spatial-temporal random effect model.

高雄地區空氣污染對於蕁麻疹與過敏之就診人數分析

吳宗諺

國立中山大學

摘要

在台灣空氣污染一直是民眾所關心的議題。世界衛生組織(World Health Organization，WHO)在 2009 年發表的《全球健康風險》報告中，市區室外空氣污染為全球死亡人口數第十四大風險因子。

近十幾年來，空氣污染與呼吸系統疾病的相關研究如雨後春筍般地發表，然而與皮膚疾病相關之研究卻是寥若晨星。皮膚是人體防禦外來影響的第一道防線，是最容易受到外在環境而受到損害的組織。而本文將探討空氣污染對於蕁麻疹與皮膚過敏之就診人數的影響。

為分析不同的空氣污染物對於蕁麻疹與過敏的影響，我們使用了廣義加法模型(Generalized Additive Model,GAM)來作為第一步的模型建立，並以天氣為共同變數，例如：溫度、濕度、風速等，再以樣條函數(spline function)估計在不同的天氣下對於該疾病之就診人數的影響趨勢。此外由於空氣污染與就診人數之資料皆有時間序列的性質，我們新建一共同變數—星期效應，以控制資料的短期時間趨勢。另外我們也估計各別空氣污染物一到七天的滯後效應，分別以不同滯後天數加入廣義加法模型內，分析在短期內空氣污染對於蕁麻疹與過敏的就診人數增減趨勢。

關鍵詞：蕁麻疹、皮膚過敏、空氣污染、就診人數、廣義加法模型、時間序列

存活分析中的切分點模擬研究

劉佳峻

國立中山大學應用數學系

摘要

醫學上常常需要對指標性的連續變數做切分，也就是把連續型的預後因子轉換成類別型態，以利於臨床上的判斷與解釋。而切分點的研究有三個問題需要解決，其一是尋找正確的切分點個數(number of cut points)，在傳統的方法中通常只能尋找一個切分點，若要找尋一個以上的切點往往須由有經驗的相關人士作決定。其二是決定正確的切分點位置(cut point)，若是前一步驟所找到的切點數不正確，則在此階段所找到的切點位置可能與真實情況相差甚遠。而最後則是在決定切分點數與切分點位置後一連串的統計推論，如：p-value、相對風險(relative risk)等。

而傳統的方法多數只能尋找一個切分點，且存在型一錯誤率膨脹的問題。而我們提出了二摺交叉驗證結合蒙地卡羅法(two-fold cross validation with Monte Carlo method,CVM)，解決了型一錯誤率膨脹的問題，且能夠尋找一個以上的切分點，並針對後續的統計推論進行校正(Chang et al. revised)，但我

們並未針對 CVM 與其他估計切分點的方法進行比較。是以本文後續將會介紹其他三種方法，並與 CVM 一起套用在存活類型的資料上，進行模擬，比較各個方法在單變量與多變量及不同樣本大小情況下的表現。

關鍵詞：切分點、型一錯誤率膨脹、交叉驗證法、蒙地卡羅法

The related researches of statistical methods and AI applied to medical images

Tai-Been Chen
Department of Medical Imaging and Radiological Sciences, I-Shou University

Abstract

The technology of artificial intelligence of deep learning (AI-DL) is rapidly developing in various fields. Especially in the field of medical imaging, the application of convolution neuro network (CNN) method has been proved to be able to effectively undergoing segmentation, identification, classification and prediction for medical images, such as MRI, X-ray, Mammography, Computed tomography (CT), Positron Emission Tomography (PET), Single Photon Emission Computed Tomography (SPECT), and hybrid of PET and CT or PET and MRI, etc. The main interesting target of medical images include identified types, addressed locations, estimated volume or size, evaluated outcomes of prognosis of lesions or tumor. However, the successfulness of CNN applying to medical images includes understanding of statistical methods, format and imaging principles of medical tomography, designing suitable structures of layers of CNN, and setting initial conditions in AI-DL. In this work, the related researches of statistical methods and AI applied to medical images were introduced via articles reviewing and summary. It is expected to provides ideas of research and development in the field of medical imaging.

Keywords: AI-DL, CNN, Medical Image

運用電腦視覺以現地雨量筒影像估計降雨量之研究

Precipitation Estimation from In-situ Rain Gauge Images by Computer Vision

陳穗碧* 黃國豪 方耀民

逢甲大學地理資訊系統研究中心

尹孝元 黃效禹 林建良

行政院農業委員會水土保持局

摘要

目前台灣地區共有超過 900 個實體雨量站，受限於地形及電力通訊設備的緣故，多數雨量站分布集中於低於 500 公尺海拔地區（平地），山區的雨量站密度較小，因此雨量站的分佈有不均的情況，水土保持局自民國 94 年起培訓土石流防災專員，希望在颱風警報發布後，幫忙監控山區的雨量與土石的狀態，以提高山區雨量觀測的密度，至今防災專員已成軍 13 年，本研究運用物體辨識的演算法辨識防災專員拍攝的現地簡易雨量筒單張影像，以單張現地雨量計影像估計降雨量模型，輔助降低防災專員雨量傳送異常的問題，同時可提高地理資訊精度與時間戳記的正確性，為後續簡易雨量筒資訊系統化、發展其他相關雨量服務以及推廣公眾參與回報雨量值的基礎。

關鍵詞：現地雨量筒影像、降雨量、物體辨識

機器學習分類方法於潛在類別迴歸模型的超參數選擇

黃冠華*、冼航平
交通大學統計學研究所

摘要

潛在類別分析（LCA）是一種用於「多變數分類資料」（multi-level categorical data）的分析模型，可將母體分為多個類別。潛在類別迴歸（LCR）模型擴展了 LCA，將變數的影響結合到估計潛在類別和測量指標上，放鬆了均勻概率假設，即不同測量指標的發生機率會隨某些個體特徵而變化。我們開發了一種自動程序來選擇最佳的測量指標條件機率，而不需要人為判斷。此外，LCR 有三種超參數需要選擇，包括潛在類別的數量，與潛在類別分佈相關的變數和與測量指標的條件機率相關聯的變數。在現行的應用中，學者多使用基於信息標準的程序（例如，AIC，BIC）和/或逐步回歸來選擇最合適的超參數組合。本研究開發了一種基於機器學習分類技術的替代方法，來執行 LCR 模型的超參數選擇。

關鍵詞：分類、超參數選擇、潛在類別迴歸模型、機器學習

Joint Pose and Shape Optimization for 3D Face Reconstruction from a Single Image

Chia-Po Wei (魏家博)
Department of Electrical Engineering, National Sun Yat-sen University

摘要

Reconstructing the 3D shape of a face from a single image is a challenging task, because for unconstrained scenarios, the input image is captured under variations of pose, illumination, expression, or even disguise. Previous approaches to this problem typically require careful initialization of algorithms or segmentation of face regions. We propose to estimate the 3D face shape based on joint pose and shape estimation with a set of reference 3D models. Our method is not only able to perform person-specific shape reconstruction, but also able to recover the camera pose. Experimental results validate that our method is effective and computationally feasible for reconstructing the 3D face shape from a single unconstrained image.

關鍵詞：3D face reconstruction

應用線上健康日記追蹤氣喘發作危險因子

詹大千[1]、胡翠華[2]、朱彥華[2]、黃景祥[2]*

[1]中央研究院人文社會科學研究中心

[2]中央研究院統計科學研究所

摘要

過去的研究多專注在症狀的監測與如何使用藥物舒緩症狀，對於個人健康行為與環境暴露同時對氣喘發作的影響仍是一項具有挑戰的研究。在這個研究設計，我們特別強調個人與環境狀態對氣喘症狀發生所造成的影響，因此設計了一個線上健康日記來蒐集氣喘學童與其家長的健康行為，並蒐集其他有過敏史的成人當作我們的受試者(包含氣喘、過敏性鼻炎、異位性皮膚炎、過敏性結膜炎)。受試者只需要使用行動裝置或電腦紀錄每日與健康相關的活動，如睡眠、睡眠、飲食、對空氣品質與氣溫的感知與氣喘的症狀；同時記錄二手菸的暴露與室內、室外活動時間。結合以上這些資訊與大氣空氣品質資料來代表個人當日的健康行為、空氣汙染等環境因子的暴露量，並利用廣義線性混合模式(GLMM)來估計這些因子的影響力。在研究期間(2017 年 1 月-2017 年 6 月，與 2017 年 10 月-2018 年 9 月)共有 132 位受試者提供 25,016 筆日記資料，其中有 84 位受試者在 1,458 筆日記資料中有紀錄氣喘的相關症狀。研究結果顯示對於受試者氣喘症狀發生的危險因子有二手菸的暴露、接觸有類流感症狀的人、睡眠品質不好、高強度運動、吃過多的全穀根莖類、感知到低溫、不好的空氣品質與暴露到高濃度的臭氧。另外，多吃水果、海鮮與感知到高溫皆會降低發生氣喘症狀的風險。總結來說，改善個人健康行為、避免二手菸暴露與注意溫度與空氣品質的感受將有助於降低氣喘症狀的發生。

關鍵詞：氣喘、飲食行為、空氣汙染、天氣、運動、吸菸、感知

整合網路輿情資料與電話訪問資料的民眾態度演算法

楊雅惠*
典通股份有限公司

蔡孟君、劉炅函
浚鴻數據開發股份有限公司

摘要

2016 年美國總統大選結果跌破各界眼鏡，主流媒體對選情的預測全軍覆沒，使各界開始質疑電話民調查結果的可信度，學界及業界也紛紛投入研究造成民調偏差的原因，並尋找可行的解決方案。根據一份由斯坦福大學(Stanford University)、哥倫比亞大學(Columbia University)、微軟(Microsoft)合作發佈的研究報告顯示，比對 1998 年到 2014 年美國總統、國會和各州州長選舉中 4,221 份最接近選舉當天的民調以及最終選舉結果，發現樣本量為 1,000 人時的民調平均誤差範圍超過±7%，遠超過統計學原理上所說的±3%，這顯示現行電話民調的非抽樣誤差可能遠大於抽樣誤差。造成這個現象的原因很多元，包含「調查媒介(市內電話)能接觸到的受訪者代表性不足」、「民眾據實表態意願低」、「問卷設計不良」……等。本研究提出一套以網路輿情資料結合電話訪問資料的方法，解決傳統電話訪問年輕樣本缺少、受訪者不願表態比例提升的問題。

關鍵詞：輿情分析、電話訪問法、滿意度、唯手機族

國家認同概念的裂解與重組：

使用網路調查資料於探索式研究的嘗試與發現

劉正山

國立中山大學政治學研究所

csliu@mail.nsysu.edu.tw

摘要

傳統上使用民意調查進行敏感政治議題的分析，多偏重於使用有「代表性」的樣本進行變數的描述或推論，或是將民意調查問卷的題目，用於西方文獻中假設的驗證。這個做法明顯地忽略兩個事實：當樣本的代表性降低，描述與推論便會失真；以建構理論為名的假設其實並不一定來自本土的民意脈絡，導致問卷題的設計過於偏重測量。本研究從厚數據或厚資料（thick data）的視野出發，以探索的方法結合極富彈性的網路調查，使用收集於 2018 地方選舉期間的網路調查資料，呈現當前年輕選民在國家認同上的分布，以及不同國家認同概念之間的潛在關聯。這個發現也將為國內設計民意調查及使用民意調查資料的學者及從業者提供看待網路調查這個工具的新視野。

關鍵詞：民意調查、厚數據、厚資料、國家認同、探索式資料分析

Abstract for STS 2019

Title: Sequential Designs for Biomedical Applications

(joint work with Seongho Kim, School of Medicine, Cancer Biology Program, Wayne State University and Julie Zhou, Department of Mathematics and Statistics, University of Victoria)

Speaker: Professor Weng Kee Wong,

Department of Biostatistics

Fielding School of Public Health

University of California at Los Angeles

This talk presents nature-inspired metaheuristic and mathematical programming algorithms for constructing sequential designs for biomedical studies. We first discuss use of particle swarm optimization to construct an extended version of Simon's 2-stage designs for a Phase II trial subject to multiple type 1 and 2 error rates. The sequential design first recruits a number of subjects in Stage 1 and responses from them are ascertained. Depending on the quality of the responses from the first stage, the second stage design then tests one of three possible true response rates from the treated subjects. To solve the constrained optimization problem, we need to find 10 optimal positive integers representing the number of patients required at stage 1, how many responses are needed in the first stage before the trial is abandoned or moved forward to stage 2, and the numbers of additional subjects and responses required in stage 2 to test one of the 3 hypotheses that posits different success rates of the drug.

We also discuss use of a free and very flexible engineering software called CVX to solve convex optimization problems. In particular, we show the program can be easily amended to find different types of optimal designs under a convex optimization setup.

Combine Expectation-Maximization with Active Learning for Linear Discrimination Analysis

Ray-Bing Chen

Department of Statistics and Institute of Data Science

National Cheng Kung University

We study an active learning procedure for the linear discrimination analysis based on the mixture normal model assumption. Here the expectation-maximization (EM) algorithm is used to integrate the information from the unlabeled subjects into account. To select the next subjects, an AUC-based influence function is adopted as a selection criterion. Several numerical results show the advantages of the proposed active learning procedure.

Adaptive design + Sequential method = Active Learning

Yuan-Chin Ivan Chang

Institute of Statistical Science, Academia Sinica

Abstract:

In Wikipedia, it says that “Active learning is a special case of machine learning in which a learning algorithm is able to interactively query the user (or some other information source) to obtain the desired outputs at new data points. In statistics literature it is sometimes also called optimal experimental design.”

Basically, this kind of methods is used in the scenarios where “unlabeled data is abundant but manually labeling is expensive.” In such a scenario, learning algorithms can actively query the user/teacher for labels. The number of examples to learn a concept can be selected by users/trainers and often be much lower than the number required in normal supervised learning. From a statistical perspectives, such an active learning terminologies in machine learning literature refer to model building problems without complete response information and the responses of samples can be revealed when there are selected as training samples. Therefore, this type of problems will usually involve the concepts of “sequential sampling,” “sequential experimental design,” and sequential analysis methods. Due to the adaptive sampling schemes used during learning courses of active learning, the ideas of stochastic regression/control can be useful in analyzing this type of data. In this talk, we will present several model-based classification methods under such an active learning scenarios to illustrate how the sequential analysis and experimental design are integrated in this kind of problems.

Functional Tail Dependence Coefficients for Copula

Zhiruo Liu, Da-Hsiang Lien, and Keying Ye

University of Texas at San Antonio

Researchers and practitioners are interested in the associations between variables at extreme values in which large amount of profits or losses in financial industry are considered. Sibuya (1960) proposed a conditional dependence structure, called tail dependence coefficient, to measure the asymptotic dependency between variables. This coefficient has become a standard measurement of associations between variables at extreme values.

In this research, we propose a functional tail dependence structure between two variables at their extremities in the sense that the rates approaching to extremities are functionally different. We obtain definite solutions of such proposed tail dependence coefficients for six commonly used copulas under mild assumptions. In addition, empirical studies are carried out for financial data.

Dependence Properties of B-Spline Copulas

Xiaoling Dou, Waseda University, Japan

Abstract:
We construct by using B-spline functions a class of copulas that includes the Bernstein copulas arising in Baker's distributions. The range of correlation of the B-spline copulas is examined, and the Fr¥'{e}chet--Hoeffding upper bound is proved to be attained when the number of B-spline functions goes to infinity. As the B-spline functions are well-known to be an order-complete weak Tchebycheff system from which the property of total positivity of any order (TP${}_¥infty$) follows for the maximum correlation case, the results given here extend classical results for the Bernstein copulas. In addition, we derive in terms of the Stirling numbers of the second kind an explicit formula for the moments of the related B-spline functions on $[0,¥infty)$.
This is a joint work with Satoshi Kuriki (The Institute of Statistical Mathematics), Gwo Dong Lin(Institute of Statistical Science, Academia Sinica) and Donald Richards(Pennsylvania State University).

Particle rolling MCMC

Yasuhiro Omori
University of Tokyo, Japan

Abstract:
An efficient simulation-based methodology is proposed for the rolling window estimation of state space models. Using the framework of the conditional sequential Monte Carlo update in the particle Markov chain Monte Carlo estimation, weighted particles are updated to learn and forget the information of new and old observations by the forward and backward block sampling with the particle simulation smoother. These particles are also propagated by the MCMC update step. Theoretical justifications are provided for the proposed estimation methodology. The computational performance is evaluated in illustrative examples, showing that the posterior distributions of model parameters and marginal likelihoods are estimated with accuracy. Finally, as a special case, our proposed method can be used as a new sequential MCMC based on Particle Gibbs, which is the promising alternative to SMC2 based on Particle MH.

Direct sampler with computational algebra for toric models

Shuhei Mano, Institute of Statistical Mathematics, Japan

Abstract:
For a toric model (hierarchical log-linear model) we usually consider sampling with MCMC, but it is not always easy to construct irreducible Markov chain. Diaconis and Sturmfels (1998, Ann. Stat.) showed that a Groebner basis of a toric ideal of the polynomial ring provides a basis of moves, and such bases are called Markov bases. However, by considering holonomic ideal of the ring of differential operators, the speaker showed that a direct and i.i.d. sampling from a toric model is possible and provided the algorithm (2017, Electron. J. Stat.). In this talk, I will introduce some of toric models which have been considered to hamper direct sampling. The examples include non-null models, non-decomposable graphical models, non-graphical models, and Young diagrams of non-exchangeable partitions. Then, I will introduce the direct sampling algorithm and explain how it works for such models. Finally, I will explain mathematical challenges emerging in replacing an MCMC sampler by the direct sampler. This talk is based on collaboration with Professor Nobuki Takayama at Kobe University.

帕金森氏症單光子斷層掃描影像分類系統

黃士峰、同悅誠*

國立高雄大學統計學研究所

摘要

本研究建立一套透過單光子斷層掃描（single-photon emission computed tomography，簡記為 SPECT）影像判斷受檢者是否罹患帕金森氏症的系統，以協助提升診斷效率與節省醫療資源。該系統首先透過提取帕金森氏症影像特徵，並使用支持向量機將醫生可以初判的影像資料歸類為正常或是罹病；對於醫生無法初判的影像資料，則提出一主動式學習的方法，建議優先進行進一步臨床檢測以判別是否罹病的影像資料順序，再將檢測結果用於更新分類器。實證研究方面，共收集 634 張受檢者的 SPECT 影像資料，包含醫生無法進行初判的 204 張影像，經過所提出方法依序建議其中 10 張影像進行進一步檢測以辨識其類別後，數值結果呈現更新後的分類器可提升分類準確度。

關鍵詞：主動式學習、醫療影像、帕金森氏症、支持向量機

應用淨重新分類增進法衡量臺灣自殺企圖者訪視意願預測模型

Applications of Net Reclassification Improvement to Evaluate Models of Suicide Attempt in Taiwan

報告同學：李韶凱
指導教授：劉力瑜 博士

摘要

我國自殺企圖通報系統，得以第一手獲報每一筆自殺企圖者的特徵資料，同時再委派社福單位進一步慰問及訪視。本研究將使用此通報系統之通報紀錄，經由變數選擇方法篩選具統計意義之變數以建模，預測企圖者是否願意接受訪視。最初以羅吉斯迴歸建立預測模型時，其 AUC 值為 0.65，並未能有效區分企圖者接受訪視的意願。在變數選擇時，發現「通報單位別」並未納入羅吉斯迴歸模型中，因此將其加入建立新的風險預測模型，以 NRI (Net Reclassification Improve) 方法比較模型表現差異;風險預測模型的判定閾值則由事發率以及 ROC 方法決定之。由切點型 NRI 顯示陰性正確結果有 0.7%的提升，連續型 NRI 則呈現 75%的不願接受訪視者其預測事發機率降低，兩方式有一致的結果；而加權 NRI (weighted NRI, wNRI) 能提供不同陰性與陽性的權重比例，會有不同的判定閾值設定，若陰性結果重要性為陽性結果的 2 倍時，該變數的納入對模型預測有幫助，當重視陽性結果過於陰性時，wNRI 則為負值，代表新變數的納入反倒使模型預測效果降低。由於目前的防治政策為：所有受通報企圖者都將派有專員訪視，本研究期望能有更佳的陰性預測結果，作為後續研究訪視對象心理或通報處理之參考。

關鍵字：自殺企圖、預測模型、通報單位、ROC、NRI

Market tracing through portfolio optimization

Hsiang-An Hsu
National Central University (NCU)

Abstract

Tracing index in the market is now a crucial and popular topic in finance. Usually, they use technical analysis to forecast the direction of index through the study of past market data. In this paper, we construct a model to trace an index based on the technique of portfolio optimization problem and prove the finiteness of this model using the dynamic programming principle and the corresponding Hamilton–Jacobi–Bellman equation. Also, the sensitive analysis is illustrated through the numerical study. Finally, we examine the proposed model is by real data included S&P 500 and several individual stocks in the U.S.

Key words

Market tracing, portfolio optimization, dynamic programming principle, and Hamilton–Jacobi–Bellman equation.

應用象徵性資料分析法於世界盃足球賽球隊得分之預測

陳彥辰*、吳漢銘
國立臺北大學統計學系

摘要

運動數據是包含各球隊及其球員們的資料紀錄，這類資料的收集與研究分析在職業競技場上及對相關產業發展，具有重要的參考價值。本研究即是針對世界盃足球賽各球隊歷年參賽資料，應用象徵性資料分析法預測球隊在下一屆比賽之可能得分。傳統上的運動數據是以球員基本資料及其表現紀錄為收集及分析單位。本研究中，我們以球隊為分析單位，將屬於同一球隊之隊員資料變數，彙整成象徵型資料，例如: 年齡以區間表示、國籍以長條圖表示、或得分數以直方圖表示，接著採用象徵型迴歸分析法來預測球隊之得分表現。同時我們也會進行探索式象徵性資料分析，並與傳統分析方法相比較。

關鍵詞：象徵性資料、世界盃足球賽、象徵型迴歸分析、探索式資料分析。

高雄市房屋價格預測系統與視覺化介面

黃士峰、彭筱雅*
國立高雄大學統計學研究所

摘要

本研究透過收集實價登錄網上 2013 年 7 月到 2018 年 6 月間高雄市的房價資料，結合分群技術與位置離差（location-dispersion）模型，提出一個半參數的房價估計量，並建立一個高雄市房屋價格預測系統。數值研究顯示，所提出模型的配適與預測能力優於加權最小平方估計法所建立的迴歸模型。並以 R 程式語言的 Shiny 套件建立視覺化介面，協助使用者了解房價趨勢，提升房價預估系統在操作上的便利性。

關鍵詞：位置離差模型、加權最小平方法、視覺化