MFPCA: a fast package for multidimensional functional principal component analysis with GPGPU computation

Chen, Lu-Hung (陳律閎)*、Hsu, Chai-Yung (許嘉揚)
中興大學統計所

摘要

Functional principal component analysis (FPCA) is an important tool in functional data analysis. Recently, FPCA had been extended to analysis multidimensional functional/longitudinal data observed on a general d-dimensional domain. Unfortunately, the computational burden of multidimensional FPCA increases exponentially as d. In this study we introduce a fast implementation for multidimensional FPCA based on modern GPGPU (General-purpose computing on graphics processing units) architecture. The package implemented by OpenCL, and thus it can be accelerated by most of the mainstream GPUs. The package can also be integrated with Apache Spark to deal with big data.

關鍵詞：Multivariate functional principal component analysis, GPGPU computation, big data, functional data analysis, local polynomial smoothing

Gradient-Based Approach to Sufficient Dimension Reduction for Functional and Longitudinal Covariates

Ming-Yueh Huang

Abstract

In this talk, we will propose a new sufficient dimension reduction method for functional and longitudinal covariates. Different from the existing inverse regression methods, we adopt an average derivative approach, which requires only smoothness conditions on the population parameter functions and no linearity condition is needed. The proposed estimator can be obtained by standard functional principal component analysis method, and can adapt to sparsely observed covariates. We also demonstrate the practicability through extensive simulations and two empirical examples.

Functional clustering and missing value imputation of longitudinal data

Pai-Ling Li
Tamkang University

Abstract

We propose a functional data approach to clustering and missing value imputation for incomplete longitudinal data. We adopt the notion of subspace-projected functional data clustering that each observed trajectory is viewed as a realization of a random function and is drawn from a mixture of stochastic processes, where each subprocess represents a cluster with a cluster-specific mean function and covariance function. The proposed algorithm comprises the probabilistic functional clustering (PFC) and the missing value imputation based on clustering results obtained from the PFC. The performance of the proposed method is demonstrated through a data example.

Keywords: Clustering; functional data analysis; missing value

Habitat variation in temporal beta diversity of tree species in a tropical forest

Yi-Ching Lin[1, 4], Shu-HuiWu[2], and Jonathan A. Myers[3]

1 Department of Life Science, Tunghai University, Taichung, 40704, Taiwan
2 Taiwan Forestry Research Institute, Taipei, 10066, Taiwan
3 Department of Biology, Washington University in St. Louis, MO, 63130, USA

Understanding temporal patterns of species diversity and identifying mechanisms underlying such patterns are important themes in community ecology. The contemporary theory of community assembly suggests that species diversity is determined by the net effects of deterministic and stochastic processes. These two opposing processes, however, may differ among habitats. The effects of habitat filtering may lead to habitat-specific patterns of species diversity. The aim of this study is to determine whether temporal beta diversity varied among different habitats (limestone outcrops, slopes, and valleys) in the Kenting Karst Forest. We also evaluate habitat variation in the relative importance of deterministic and stochastic processes. This study was carried out in the Kenting Karst Forest Dynamics Plot (10 ha). All woody plants with diameter at the breast height (DBH) greater than 1 cm were mapped, tagged and identified to species in 2008 and 2013. Temporal beta diversity was estimated by Bray-Cutis Dissimilarity Indices. Using a null-model approach, we evaluated effects of stochasticity on temporal beta diversity after controlling for local community size. Our results indicated that α diversity was lower in the valley habitat than the other habitats. And yet, its temporal beta diversity was significantly higher in valleys. Habitat-specific temporal changes were identified. Stochastic processes was dominated in slopes, but not in valleys or limestone outcrops. The high beta diversity in the valley habitat may be attributed to deer herbivory. This results highlights the importance spatial variation in temporal processes and provided a unique opportunity to reconcile existing theories of species diversity.

Keyword: Biodiversity, Forest Dynamics Plot, Habitat filtering, Stochasticity,

4. Author to whom all correspondence should be addressed:
E-mail: yichingtree@gmail.com

Studies of the long-term temporal variation in coastal fish assemblages in northern Taiwan

陳虹諺*
國立臺灣大學農藝學系

邵廣昭
中央研究院生物多樣性中心

摘要

Long-term time series data with consistent sampling methods are rare in marine fish communities, especially the ones of non-target coastal fishes. We describe two long-term time series fish assemblage datasets at two nuclear power plants on the northern coast of Taiwan. One is an impinged fish assemblage dataset containing fish collected once a month over 19 years. The fish killed by impingement upon cooling water intake screens at the plants were collected systematically. The other one is a long-term time series dataset of fish collected by trammel net fish sampling and observed by an underwater diving visual census near the thermal discharges at the plants. The fish assemblages were monitored four times per year over 18 years. By using these datasets, the long-term temporal variation in the coastal fish assemblages in northern Taiwan were reported.

關鍵詞：biodiversity index, community diversity, phylogeny, seasonal pattern

Using Good-Turing frequency formula to insight the statistical behavior of richness estimators

Chun-Huo Chiu (邱春火)

Department of Agronomy, National Taiwan University

Abstract

Species richness is the simplest and most intuitive concept of diversity. However, due to practical limitations, it is virtually impossible to detect all species, especially in hyper-diverse assemblages with many rare species. In almost every biodiversity survey and monitoring project, some proportion of the species that are present fail to be detected. There are many non-parametric richness estimators proposed to correct the bias, including Chao1 estimator, first order and second order of Jackknife approaches. However, there are no theoretically evaluations about these widely used richness estimators in the literature. The Good-Turing frequency formula, originally developed for cryptography, estimates in an ecological context the true frequencies of rare species in a single assemblage based on an incomplete sample of individuals. Here, we used the theory of Good-Turing frequency formulas to reveals the sufficient conditions under which the estimators are nearly unbiased and to theoretically elaborate the statistical estimator's behaviors increasing sample size.

Base on Good-Turing frequency formulas, it becomes to be intuitive and easily understanding that (1) the Chao1 estimator is unbiased not only in homogeneous model, (2) no species frequency model to fit the unbiased conditions of the Jackknife estimators, (3) Jackknife estimators are always underestimated when sample size is small and overestimated when sample size is large enough. Finally, simulation results

are reported to numerically verify the performance of the investigated estimators in this study.

Keywords: biodiversity, Good-Turing frequency formula, Chao1 estimator, Jackknife

A parametric model for wearable sensor-based physical activity monitoring data with imformative device wear

Lee, Chien-Wei(李建緯) *、Dr. Huang, Shih-Hao(黃世豪)
National Central University

摘要

Wearable devices provide the opportunity to collect information of human being's physical activity. However, there is non-negligible deviation from the subject's willingness and other potential behaviors. In practice, researchers usually utilize semiparametric panel count regression to avoid the bias from model misspecification. Parametric approach,on the other hand, can control such bias by model selection, and can provide estimation with smaller variance. In this paper, we provide simulation studies to compare the performance of the semiparametic and parametric approaches. We find that the parametric approach has smaller MSE under a moderate sample size. We apply our approach to the wearable device data from National Health and Nutrition Examination Survey is USA.

Keywords:
bias-variance trade-off,panel count regression, wearable devices.

Ordered Iterative Least Square Estimator for Extended Efficiency Model

Prof. Chih-Hao Chang (張志浩), Lim Wei Yee (林暐詒)*
Institute of Statistics, National University of Kaohsiung

Abstract ..

In data analysis, we are generally interested in estimating the mean trend of data. The mean trend of this paper is modeled as a piecewise regression with two breakpoints, where the mean trend between the two breakpoints is stable or at an optimal status. This model can be applied to find a time period or an interval of a variable such as temperature, pressure or humidity of a system where the status between the interval is stable. For example, in the steel casting industry, the degree of superheat of molten iron shall be carefully controlled at a suitable interval where the low quality of production may be due to the degree of superheat that is too low or too high. For this purpose, we consider two breakpoints for the covariate of our data which can be univariate or multivariate and assume the mean trend of the data within the interval of the two breakpoints is a constant function; while for data outside the interval, the mean trend is polynomial linear regression functions. We also provide an estimation strategy and algorithm for estimating breakpoints. Furthermore, we establish the asymptotic properties of the breakpoints estimators under some regularity conditions.

Keywords: breakpoints, segmented regression, asymptotic properties

利用溫度、濕度和風速建立新空氣汙染指標

Establish New Air Pollution Indicators using Temperature, Humidity and Wind Speed

高于珺*、馬瀰嘉

國立成功大學統計學研究所

摘要

在臺灣，對於空氣品質與相關議題的重視逐年提升，主要因為相關領域研究的蓬勃發展與知識迅速地傳播，社會大眾逐漸意識到空氣中的汙染物會影響到人體的健康狀態，甚至是危害到其他生物的發展，進而對此議題有了更多的關心。環保署於 2016 年全面改以空氣品質指標(Air Quality Index, AQI)取代 PSI (Pollutant Standards Index)，整合懸浮微粒(PM_{10})、二氧化硫(SO_2)、二氧化氮(NO_2)、一氧化碳(CO) 、臭氧(O_3)以及細懸浮微粒($PM_{2.5}$)等主要六種污染物質副指標值，採取當中最嚴重的一項副指標值作為 AQI 的數值。

本研究參考之前的文獻改善 AQI 僅以六種污染物質最嚴重的一項副指標值作為 AQI 數值的不足，以及 RAQI 以熵函數建構修正指標的概念，延續修正 RAQI 指標，進而考慮溫度、濕度、風速以及測站的地理資訊對 AQI 的影響，提出調整後的臺灣空氣品質指標。

關鍵詞：空氣品質指標、溫度、濕度、風速、地理資訊

運用信令大數據資料與抽樣調查方法推估人口參數之研究

陳芷瑩*、王鴻龍
國立臺北大學統計學系

摘要

目前國內使用之人口統計或相關調查多以政府的戶籍登記資料為基礎，但以城市地區而言，戶籍人數可能低估實際活動人數，對於政府政策規劃或企業投資效益評估，更需要一地區之人流、活動與常住人口等資訊。而每十年進行一次的人口普查，雖然能提供常住人口的資訊，但人口變遷快速，十年更新一次的普查資料，無法即時反應人口的活動情形。

運用手機信令大數據資料，捕捉用戶不同時間點的位置與移動，推論各地人口常住與活動情形。亦加入「107 年手機持有與使用狀況調查」提供之市話、手機調查資料，採用 Skinner(1991)與洪永泰(2017)提出的雙清冊反覆加權方法，整合市話與手機樣本，並依基本資料結構做反覆加權調整，使樣本結構比例與母體一致。

結合信令大數據與抽樣調查資料，運用比率估計方法，建立人口推估模式及相關參數估計，並推估常住人口估計值與估計變異數，提供即時的常住人口推估資訊。

關鍵詞：常住人口、信令資料、雙重清冊、人口推估

基於符號型統計量和度量方法之 EWMA 管制圖於輪廓監控的比較分析

陳星達*、許湘伶
國立高雄大學統計學研究所

摘要

現今有些產品或製程的品質特性可用一個反應變數對多個解釋變數的函數關係式來表達，這種函數關係式的資料類型稱為輪廓資料(profile data) ，為現代工業中所需監控的製程資料之一，其監控方式稱為輪廓監控。而輪廓資料的函數關係式可概分為線性與非線性，部分文獻在線性模型考量下提出相關的製程監控程序，然而非線性模型可更佳的近似真實的輪廓資料特性，因而在此探討非線性輪廓資料製程變化之偵測方法。本研究透過 Sign Test 無母數統計量、Wilcoxon 無母數統計量以及 6 種度量(metrics)為指標來建構指數加權移動平均(exponentially weighted moving average, EWMA)管制圖並進行輪廓資料品質偵測評估。考量製程變化對於製程監控之影響性，以平均連串長度(average run length, ARL)作為偵測能力評估指標，透過模擬研究對所採用的品質偵測方法進行各種製程輪廓改變之分析與討論，以判斷出適用的方法及較佳的度量指標。而後，透過一組非線性輪廓資料進行數值實例的驗證。

關鍵詞：EWMA 管制圖、Sign Test 無母數統計量、Wilcoxon 無母數統計量、度量

結合氣象研究與預測模型(Weather Research and Forecast Model)及機器學習技術建置太陽及風力發電預測模型

*黃凱斌、陳雍太
財團法人台灣電子檢驗中心

摘要

綠色電力係指以再生能源所產生的電力，如風能、太陽能、水力、地熱、生質能等。因具有較傳統電力低碳與低污染排放特性，在先進國家被區別為一般電力，而收取綠色電價。為配合國家再生能源發電量分析暨查核系統發展計畫，此研究利用機器學習技術結合氣象研究與預測模型(Weather Research and Forecast Model, WRF)輸出之數值氣象預測資料，建置出再生能源發電預測模型，以利再生能憑證案場之發電量預測及案場回傳電量之比對。本研究主要以內插至憑證案場之WRF變數及憑證案場回傳之發電資料，並利用統計技術及機器學習相關智慧演算法進行氣候資料間的相關性分析及發電量預測，有助於瞭解氣象資料對發電量之影響，進而預測再生能源之發電量。本研究主要利用長短期記憶網路(Long Short Term Memory Network, LSTM)演算法進行發電量預測分析，並基於預測分析結果，開發再生能源憑證案場之氣象及發電監控地理資訊視覺化系統。相關研究成果可用於查核再生能源憑證案場回傳之再生能源發電量是否合理，有助於提高憑證機制之公信力及附加價值。

關鍵詞：再生能源憑證、機器學習、深度學習、長短期記憶網路、地理資訊視覺化監控 。

演講題目：

Model selection for semiparametric marginal mean regression accounting for within-cluster subsampling variability and informative cluster size.

作者：沈仲維、程毅豪

Abstract:

We propose a model selection criterion for semiparametric marginal mean regression based on generalized estimating equations. The work is motivated by a longitudinal study on the physical frailty outcome in the elderly, where the cluster size, that is, the number of the observed outcomes in each subject, is "informative" in the sense that it is related to the frailty outcome itself. The new proposal, called Resampling Cluster Information Criterion (RCIC), is based on the resampling idea utilized in the within-cluster resampling method (Hoffman, Sen, and Weinberg, 2001, Biometrika 88, 1121-1134) and accommodates informative cluster size. The implementation of RCIC, however, is free of performing actual resampling of the data and hence is computationally convenient. Compared with the existing model selection methods for marginal mean regression, the RCIC method incorporates an additional component accounting for variability of the model over within-cluster subsampling, and leads to remarkable improvements in selecting the correct model, regardless of whether the cluster size is informative or not. Applying the RCIC method to the longitudinal frailty study, we identify being female, old age, low income and life satisfaction, and chronic health conditions as significant risk factors for physical frailty in the elderly.

keywords: Clustered data; Longitudinal data; Resampling; Subsampling; Variable selection

Confidence interval for the difference between two median survival times with semiparametric transformation models

Y. M. Chang*, P. S. Shen & Y. H. Tang

Department of Statistics, Tunghai University, Taichung 407, Taiwan

ABSTRACT

In medical studies, we usually are interested in comparing the treatment effects of the drug according to the difference of two median survival times. In this paper, we consider the problem of constructing conditional confidence interval for the difference of two median survival times given the covariates under a general class of the semiparametric transformation models with right-censored data. We propose two methods for constructing the conditional confidence intervals. One is based on the estimating equations (EE) estimator of Chen, Jin and Ying's (2002) and the other on the nonparametric maximum likelihood estimator of Zeng and Lin (2006). Simulation results indicate that both methods provide satisfactory coverages for finite sample. We illustrate the proposed method using a real data set in a two-arm non-small cell lung cancer study.

KEYWORDS: confidence interval; estimating equation; maximum likelihood; median survival time; semiparametric transformation model; weighted Breslow

Quantile Residual Life Regression Based on Semi-Competing Risks Data

謝進見*、王健霖

Department of Mathematics, National Chung Cheng University

Abstract

This paper investigates the quantile residual life regression based on semi-competing risk data. Because the terminal event time dependently censors the non-terminal event time, the inference on the non-terminal event time is not available without extra assumption. Therefore, we assume that the non-terminal event time and the terminal event time follow an Archimedean copula. Then, we apply the inverse probability weight technique to construct an estimating equation of quantile residual life regression coefficients. But, the estimating equation may not be continuous in coefficients. Thus, we apply the generalized solution approach to overcome this problem. Since the variance estimation of the proposed estimator is difficult to obtain, we use the bootstrap resampling method to estimate it. From simulations, it shows the performance of the proposed method is well. Finally, we analyze the Bone Marrow Transplant data for illustration.

Keywords: Archimedean copula model; Bone marrow transplant data; Dependent censoring; Quantile residual life regression; Semi-competing risks data.

Selection of EM-based Semi-Parametric Mixture Hazard Models Using Validity Indices

張怡雯*、張少同

國立臺灣師範大學數學系

Abstract

The Cox proportional hazards model is commonly used in survival analysis for describing the relationship between the survival time and covariates. The model is applied in many fields such as medicine, health care, and so on. In cases that some latent variables are involved in the model, mixture regression models are more suitable for analyzing the effects of these variables.

The determination of the number of model components is an important issue when using the mixture models. Although validity indices are a vital branch of model selection, however, they are less used for deciding the number of components in mixture regression models. Thus, we propose some new indices based on the posterior probabilities and residuals by referring to the existing methods.

The Cox proportional hazard model consists of two parts: the baseline hazard function and the proportional regression model. The estimation of baseline hazard function is known to be a challenging issue. We extend the idea of kernel estimator to the baseline hazard function and develop the expectation and maximization (EM) algorithm to estimate the parameters of the mixture regression model.

In estimating the baseline hazard function, the simulation results show that the estimated model with the kernel estimator is better suited for the data set than the piecewise constant model because the fitted curve is smoother and the kernel estimator improves the stiff structure of the piecewise constant estimator as well. Moreover, the effectiveness of the new indices in selecting the number of components is verified through experiments that a high precision of number selection of components using the new indices.

Key words: Mixture regression model, Cox proportional hazards model, EM-algorithm, Kernel estimator, Validity indices

Robust supervised dimension reduction based on γ-divergence

Wen-Shao Ho*
Institute of Applied Mathematical Sciences, National Taiwan University

Chih-Hsuan Wu、Ting-Li Chen
Institute of Statistical Science, Academia Sinica

Abstract

Linear discriminant analysis (LDA) which maximizes the ratio of the between class variance to the within class variance is widely used in supervised dimension reduction. In the traditional LDA, the discriminant space can be badly affected by the mislabeled data. To overcome this issue, we propose a robust linear discriminant analysis based on γ-divergence, a more robust measure than the Kullback–Leibler divergence. In this talk, we will introduce γ-LDA and analyze its robustness by the influence function.

key words：
Linear discriminant analysis, Dimension reduction, γ-divergence

HDMV: Visualization for high-dimensional mediation effects

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

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

A class of general pretest estimators for the normal means

Jia-Han Shih　(施嘉翰)

Graduate Institute of Statistics, National Central University, Taiwan

Yoshihiko Konno　(今野良彦)

Department of Mathematical and Physical Sciences, Japan Women's University, Japan

Yuan-Tsung Chang　(張元宗)

Department of Social Information, Mejiro University, Japan

Takeshi Emura　(江村剛志)

Graduate Institute of Statistics, National Central University, Taiwan

Abstract

For estimating a large number of mean parameters, univariate analyses remain the most popular approach in real applications due to its simplicity. Such analyses typically perform some preliminary tests (e.g. *t*-tests) to reduce the number of variables and impose some sparsity assumptions to shrink the estimates. To handle these tasks simultaneously, we propose a class of general pretest estimators that include many existing pretest, shrinkage, Bayes, and empirical Bayes estimators as special cases. We adopt the idea of randomized tests to construct a class of general pretest estimators, where the randomization probability is related to a shrinkage parameter. Theoretical properties of the proposed pretest estimator such as the exact distribution, bias, and mean squared error are derived. Our new expressions for the bias and mean squared error are simpler and more straightforward than the existing ones. We illustrate the use of the proposed estimator through the analysis of high-dimensional gene-expressions.

Keywords: *Biased estimation, Mean squared error, Shrinkage estimation, Statistical decision theory, Uncertain non-sample prior information*

基於切片的充分維度縮減法於二元不平衡資料之研究

徐維澤*、吳漢銘
國立臺北大學統計學系

摘要

運用充分維度縮減法於高維度資料以找出有效的維度縮減方向，有助於在低維度空間中探索資料所隱含的內在結構，而降維後之資料可視為原始資料的特徵擷取，將之進一步應用到分類問題或群集分析中。已有文獻研究指出，對於二元分類方法，若處理的對象是不平衡資料，亦即資料中的反應變數，其兩類別的個數比例具有極大差異時，容易導致分類方法產生偏誤; 另外一方面，尚未有充分維度縮減法於不平衡資料的相關研究。因此在本論文中，我們將探討四種以切片為基礎的充分維度縮減法應用在不平衡資料上，對於維度縮減方向估計之影響。採用的方法包含切片逆回歸法(SIR)、切片平均變異估計(SAVE)、共變異數差異法(DOC)及主要海森方向(pHd)。透過模擬研究與實際資料分析，我們發現，不平衡資料在兩類別個數比值大於 4 時，充分維度縮減方向之估計會產生偏誤，同時估計的變異程度也會增加。若應用文獻上處理不平衡資料的方法，例如過採樣(Oversampling)及 SMOTE 等方法，也無法有效改善維度縮減方向之估計問題。

關鍵詞：充分維度縮減、二元資料、不平衡資料、中心子空間、切片逆回歸法

Model Selection for High-Dimensional Misspecified Time Series Models

黃學涵*、銀慶剛
國立清華大學

蕭維政
東吳大學

摘要

In the current literature, little attention has been paid to the model selection for high-dimensional misspecified models. However, methods for model selection in misspecified models can be applied to many useful models, such as interaction models or measurement error models. In this article, we investigate the behavior of the orthogonal greedy algorithm (OGA) in high-dimensional misspecified time series models. Under a weak sparsity condition, we derive the convergence rate of OGA. By further assuming the structure of true models, we show that OGA, used in conjunction with a high-dimensional information criteria (HDIC), can achieve the sure screening property (in the sense of misspecified models) under a strong sparsity condition. We propose a concept called the "degrees of misspecification" to illustrate the effects of model misspecification on variable screening. We introduce an algorithm generalized from OGA, called Multi-step OGA (MOGA), and show that it shares similar theoretical properties as OGA but has better performance than OGA in some finite sample cases. Two special cases, interaction models and measurement

error models, of misspecified models are studied. We propose a new two-stage model selection procedure for interaction models, called MOHIT-2, under hierarchical model structure. We also propose a novel method named MOHIT_bc to tackle model selection for measurement error models. Both MOHIT-2 and MOHIT_bc are shown to possess model selection consistency. Simulation studies and real data analysis are given to demonstrate the advantages of our methods.

關鍵詞：Measurement error models; Model misspecification; Model selection consistency; Greedy algorithm; High-dimensional information criteria; Interaction models; Sure screening; Time series.

Classification of AMD and PCV Color Fundus Images by Deep Learning for Imbalanced Data

Yu-Bai Chou[#]
Attending Physician, Vitreo-Retinal Section, Department of Ophthalmology, Taipei Veterans General Hospital

Umi Tri Ruhana[#]
Institute of Statistics, National Chiao Tung University, Taiwan

Wei-Shiang Chen
Institute of Statistics, National Chiao Tung University, Taiwan

Yi-Ming Huang
Attending Physician, Vitreo-Retinal Section, Department of Ophthalmology, Taipei Veterans General Hospital

De-Kuang Hwang
Attending Physician, Vitreo-Retinal Section, Department of Ophthalmology, Taipei Veterans General Hospital

An-Fei Li
Attending Physician, Vitreo-Retinal Section, Department of Ophthalmology, Taipei Veterans General Hospital

Shih-Jen Chen
Director, Vitreo-Retinal Section, Department of Ophthalmology, Taipei Veterans General Hospital

Catherine Jui-ling Liu
Chairperson, Department of Ophthalmology, Taipei Veterans General Hospital

Henry Horng-Shing Lu
Institute of Statistics, National Chiao Tung University, Taiwan

ABSTRACT

This study classifies age-related macular degeneration (AMD) and polypoidal choroidal vasculopathy (PCV) color fundus images. It aims to investigate the impact of the imbalanced data on the classification performance of deep learning and

compare the results of distinct methods. Imbalanced data typically occurs in medical data. For example, there are usually much more AMD data than PCV data. To investigate the consequences of imbalanced data for classification, this study compares three main categories of methods for this issue: resampling methods (including undersampling and oversampling), synthetic methods (including the synthetic minority oversampling technique, SMOTE, and the adaptive synthetic sampling approach, ADASYN), and cost-sensitive learning (including the weighted binary cross entropy). To evaluate the performance, this study uses the Matthews correlation coefficient (MCC). Based on the evaluation results using the images collected in Taipei Veterans General Hospital, this study concludes that the weighted binary cross entropy has the best performance which has the false negative rate of 0.089, the false positive rate of 0.063, and the MCC value of 80.51% for test data. To test the performance further, we then added the PCV (minority class) data from Thailand. Moreover, this study also highlights the important local regions in the fundus image for classification using the gradient-weighted class activation mapping (Grad-CAM). Thus, clinical experts can investigate which local regions in the fundus image are important for the classification of AMD and PCV.

Keywords: imbalanced data classification, deep learning, resampling method, synthetic method, cost-sensitive learning, Matthews correlation coefficient (MCC), gradient-weighted class activation mapping (Grad-CAM), age-related macular degeneration (AMD), polypoidal choroidal vasculopathy (PCV)

Presenter: Umi Tri Ruhana

Correspondence author: Henry Horng-Shing Lu

# Contributed equally

Bayesian Analysis of Multivariate-*t* Nonlinear Mixed-effects Models with Missing Responses

Wan-Lun Wang* (王婉倫)

Department of Statistics, Feng Chia University, Taichung, Taiwan

Luis M. Castro

Department of Statistics, Pontificia Universidad Católica de Chile, Santiago, Chile

Abstract

The multivariate-*t* nonlinear mixed-effects model (MtNLMM) has been shown to be a promising robust tool for analyzing multiple longitudinal trajectories following arbitrary growth patterns in the presence of outliers and possible missing responses. Owing to an intractable likelihood function of the model, this paper presents a fully Bayesian estimating procedure to account for the uncertainties of model parameters, random effects, and missing responses via the Markov chain Monte Carlo method. Posterior predictive inferences for the future values as well as missing values are also investigated. A simulation study is conducted to demonstrate the feasibility of our Bayesian sampling schemes. The proposed techniques are illustrated through an application to the AIDS study.

Keywords: MCMC algorithm, Multivariate longitudinal data, Nonlinear mean profiles, Posterior distributions, Taylor series expansion.

A Bayesian Approach to Factor Screening for Multivariate Responses

俞一唐
東海大學

摘要

The modified Box-Meyer method (MBMM) has been proposed to identify active factors in an unreplicated screening experiment. This work aims to introduce the MBMM into the analysis of a screening experiment with multivariate responses. We propose an empirical Bayes approach to estimate the hyper-parameters instead of integrating them out. By doing so, the computational complexity is reduced. We illustrate the proposed approach by analyzing an examples, the dyestuffs experiment and all the active factors are identified successfully.

關鍵詞：Bayesian model averaging, Box-Meyer method, designed experiment, empirical Bayes estimation, screening experiment.

Bayesian Reliability Analysis on Trend-Gamma Process

Yi-Fu Wang and Wan-Chun Tsai

Department of Mathematics, National Chung Cheng University, Chiayi, Taiwan

Abstract

For highly reliable products, it is difficult to obtain the lifetime information. Thus, if there exists a quality characteristic whose degradation over time can be related to reliability, then the approach based on the degradation path is called the degradation analysis. For the stochastic process model, if there are ageing phenomenon in the degradation dataset, one may consider the Trend Gamma Process (TGP) model to make the reliability inference. In this work, we will develop a Bayesian approach to the TGP model, involving the Markov chain Monto Carlo (MCMC) method, Bayesian sequential updating and the corresponding reliability analysis. Finally, under the TGP model, the Bayesian approach will be applied to the fatigue crack growth dataset.

Bayesian Estimation for an Item Response Theory Tree Model

張育瑋*、涂俊曄
逢甲大學統計學系

Abstract

We face non-response data while analyzing questionnaire or educational testing data. While the non-response / response is also recorded, instead of simply correctness / incorrectness for every item, we can use an Item Response Theory (IRT) tree model with four end nodes (TR4) in the literature to analyze testing or questionnaire data. Many special cases of TR4 are also of interest since they describe meaningful mechanisms. However, current maximum likelihood estimation for TR4 does not work well for some of the special cases. In the current study, we propose to estimate the TR4 model using Bayesian approach with the data augmentation technique. In addition, the Markov chain Monte Carlo technique is adopted. Simulations are conducted to demonstrate the estimation performance and the efficiency of selecting the right special case. The model is further applied to an entrance examination data set for illustration.

關鍵詞：Bayesian analysis, Item Response Theory Models

結合集群分析與長短期記憶網路模型於短期電力負載預測之應用

李孟軒*、袁子倫[1]、江典聲[1]、羅夢娜[1]
[1]國立中山大學應用數學系
盧展南[2]、吳進忠[3]
國立中山大學電機工程學系[2]、台灣電力調度處[2]

摘要

本研究討論結合非監督式學習中的群集分析方法(K-medoids)與長短期記憶(Long Short-Term Memory, LSTM)神經網路模型對短期電力負載預測之成效。研究使用資料為 2015 至 2018 年的歷史負載資料，以及中央氣象局提供之預測溫度和實際溫度。利用集群分析對電力與溫度資料進行分群，並參考過去的遞迴神經網路模型(Recurrent Neural Network, RNN)使用的週期性基底函數與超參數組合做為模型的輸入與變數。嘗試利用自我相關函數(autocorrelation function, ACF)與偏自相關函數(partial autocorrelation function, PACF)尋找較佳的時間步數對未來 8 天電力負載進行預測建模。模型評估準則為預測結果與真實負載量的平均絕對比例誤差(Mean Absolute Percentage Error, MAPE)，計算 2018 年每日負載預測的平均絕對誤差並與之前遞迴神經網路效果進行比較。

關鍵詞：K-Medoids 分群法、平均絕對比例誤差、長短期記憶網路模型

運用 R Shiny 建立互動式視覺化平台之校務資料分析

黃士峰、李政葦*

國立高雄大學統計學研究所

摘要

本研究結合學生在學成績與畢業流向調查結果，應用 R 程式語言中的 Shiny 套件建立一個互動式介面，內容包括關聯分析、集群分析、和動態圖表等。由於 Shiny 套件為動態且互動式的應用程序，允許使用者觀察在不同的變數設定下，反應變數的變化結果，不僅可節省大量運行及修改程式的時間，使用者亦能透過互動的過程有效地了解數據，培養對數據的敏感性與對研究主題的直觀看法，從校務資料中攫取重要訊息。

關鍵詞：校務資料分析、互動式介面、視覺化

譜分群在向量分群之應用

陳寶寧*、李孟峰
國立臺北大學統計學系

摘要

近年來科技的進步與電子商務的發展，使得宅配業、線上叫車等服務的需求日漸增加。這些行業的增長使路線規劃變得更重要，有效率的把路線分群，除了讓業者營運成本減少外，還能改善因車輛的碳排放所導致的環境問題。目前大部分的業界上，對於路線的分群都是透過對地區的認知或其他自我經驗去進行劃分。可如果面對龐大的資料時此方法會非常耗時且效率低，對業者而言是浪費人力成本的舉動。因此本研究使用譜分群法針對向量型態的資料找出其潛在的特徵進行分群，最後與 K-means 分群法的結果做比較。

關鍵詞：

機器學習、非監督式學習、譜分群、K-means、向量分群

卜瓦松迴歸模型及遞迴神經網路於車流量之預測研究

黃詩芸*[1]、陳鼎文[1,2]、羅夢娜[1]
國立中山大學應用數學系[1]、金屬工業研究發展中心[2]

王志綱
高雄市智慧運輸中心

摘要

在城市路網的規劃當中，推估交通壅塞路段的車流量，以提供用路人相關道路之交通狀況，是發展智慧化交通路網的重要議題。本研究擬利用高雄市中正交流道附近的平面道路之車流量，依據卜瓦松迴歸(Conway-Maxwell-Poisson)及遞迴神經網路(Recurrent Neural Network, RNN)來建構預測車流量之模型，並比較此二方法之預測成效。我們使用的資料為 2016 年 8 月至 2018 年 10 月的資料，將其分為非特殊日及特殊日兩類型，分別建立模型。利用階層式分群(Hierarchical Clustering)，觀察車流量的相似性，且以週期性基底函數(Periodical Basis function)及群集分析之結果，作為模型之解釋變數，預測當天以及未來 7 天的車流量。並以平均絕對比例誤差(Mean Absolute Percentage Error, MAPE)及均方根誤差(Root Mean Square Error, RMSE)作為評估模型預測表現之依據。

關鍵詞：卜瓦松迴歸模型、平均絕對比例誤差、均方根誤差、週期性基底函數、階層式分群、遞迴神經網路

Mediation and interaction of age, follicle stimulating hormone (FSH) and anti-müllerian hormone (AMH) on in vitro fertilization pregnancy

Han-Chih Hsieh(謝函芝)*、Jia-Ying Su(蘇家瑩)
Institute of Statistical Science, Academia Sinica, 128 Academia Road, Taipei 11529, Taiwan

Yen-Tsung Huang(黃彥棕)
Institute of Statistical Science, Academia Sinica, 128 Academia Road, Taipei 11529, Taiwan

Shunping Wang
Women and Infants Hospital in Rhode Island, United States

Abstract

Both follicle stimulating hormone (FSH) and anti-müllerian hormone (AMH) are widely used to assess the ovarian reserve in women for in vitro fertilization (IVF). However, studies also showed that both AMH and FSH are significantly associated with age: as age increases, AMH decreases and FSH increases. This study aims to investigate the mechanism of age effect on IVF live birth rate, particularly through mediation and interaction by AMH and FSH. We conducted a retrospective cohort study of 13970 IVF cycles collected by eIVF from 2010 to 2016. A series of logistic mixed models were used to estimate the association of live birth and AMH (or FSH). The mediation effects and proportion mediated, were quantified by our previously proposed mediation analyses. We further investigated the FSH-modified mediation effects on live birth rate through AMH, accounting for the nonlinear age effect. Our analyses showed that age effect on live birth was mediated more by AMH than by FSH (18 vs. 6 %). The mediation effect through AMH can be further modified by FSH level regardless of women's age. In summary, mediation model provides a new perspective elucidating the mechanism of IVF successful rate by age.

Keywords:
anti-müllerian hormone, follicle stimulating hormone, in vitro fertilization, mediation analysis

具有工作故障與工作休假排隊系統之系統績效評估

楊東育、鍾淇翔*
國立臺北商業大學資訊與決策科學研究所

摘要

本研究考慮一個具有工作休假和工作故障的 M/M/1 排隊系統，顧客抵達系統服從參數 λ 之卜瓦松過程 (Poisson process)，系統內僅有一位服務者負責提供顧客服務，在任何時間，服務者一次只能服務一位顧客，服務者在一般忙碌期間的服務時間服從參數為 μ_b 之指數分配。當系統內所有的顧客被服務完畢後，服務者不會立即離開系統而是在系統等待一段時間，此段時間服從參數 ξ 之指數分配。此段時間若有顧客進入系統，則服務者會馬上提供服務，反之，服務者會離開系統進行休假。當服務者進行休假時並未完全停止服務，而是改以不同的服務率提供顧客服務，在休假期間的服務時間亦服從指數分配，其參數為 μ_v。服務者在服務過程中，會不預期地發生故障，服務者處於一般與休假狀態之故障率分別假設為 α_0 與 α_1。服務者一旦發生故障，會立即進行修復，服務者處於一般與休假狀態之修復時間分別服從參數為 β_0 與 β_1 之指數分配。當服務者處於故障期間並非完全停止服務，而是改以較低的服務率繼續提供顧客服務，在故障期間的服務時間服從指數分配，其參數分別為 μ_{bd} (一般狀態) 與 μ_{vd} (休假狀態)。我們利用幾何矩陣法 (matrix geometric method) 計算系統內顧客數的穩態機率分配，並且發展一些用來評估系統績效的指標。最後，我們提供數值結果說明不同的參數對於系統績效的影響程度。本研究所提出的排隊模型將可以應用於通訊網路、生產製造系統和資訊系統等領域。

關鍵字：矩陣幾何法、排隊系統、工作故障、工作休假

Lamination Design of Photovoltaic Solar Panels

蔡志群
淡江大學數學學系

摘要

Solar power is inexhaustible and has become the most appreciative choice in the world. With development stage of solar modules, solar panels are conducted by the relevant reliability tests to ensure long lifetime and power generation efficiency. During the lamination process of solar modules, the performance of the solar panels has been greatly relevant with the degree of crosslinking for EVA sheet. The degree of crosslinking for EVA sheet is obtained by using the extraction method to measure the gel content of EVA sheet. Motivated by lamination tests on solar panels, this study first constructed the statistical model with extreme value residuals to describe the relationship between the degree of cross-linking for EVA sheet and lamination time. Then, under the specification upper and lower limits of the degree of cross-linking for EVA sheet, the optimal lamination time of solar panels will be derived, and the optimal sample allocation for measuring EVA sheets destructively will be addressed.

關鍵詞：EVA sheets, the degree of crosslinking, chemical extraction method, optimal lamination time, specification limits.

Cost considerations for group testing studies with an imperfect assay and a gold standard

Shih-Hao Huang

Department of Mathematics, National Central University, Taiwan.

Abstract

We focus on optimal group testing design problems when a cheap imperfect assay and an expensive perfect assay (gold standard) for a target trait are both available. The primary goal is to accurately estimate the prevalence of the trait in a given population, where the testing error rates of the imperfect assay are treated as nuisance parameters. Budget constraints are used to reflect the relative costs of performing the two assays and of collecting the individual samples. A mixed design strategy can be adopted, where individual samples are tested as a group sample by one of the three procedures: only the imperfect assay, only the perfect assay, and both assays. We characterize the optimal designs within the class with a tight upper bound on the number of distinct group sizes for each testing procedure. Based on this information, we provide an efficient algorithm to obtain an optimal budgeted design. (Work jointly with Prof. Mong-Na Lo Huang and Prof. Kerby Shedden.)

Key words and phrases: budgeted design, gold standard, mixed design, prevalence estimation, sensitivity, specificity.

Stability and Performance Analysis of a MEWMA Controller for MIMO Processes Subject to Metrology Delay

Lin, Chien-Hua(林建華)
Department of Data Science and Big Data Analytics, Providence University

Abstract

To maximize the competitiveness of semi-conductor manufacturers, they try to increase wafer sizes and reduce measurement devices. Metrology delay is a natural problem in the implementation of run-to-run (R2R) process control scheme in semi-conductor manufacturing processes. In literatures, several authors have provided the means of using metrology delay data and Virtual Metrology (VM) values on the transient behavior and asymptotic stability of Exponentially Weighted Moving Average (EWMA) controllers. However, many semi-conductor manufacturing processes have multiple-input and multiple-output (MIMO) variables naturally. To overcome this difficulty, based on the criterion of minimizing asymptotic mean squares errors (AMSE), we show that how to choose the optimal discount factor for various combinations of metrology delay. In addition, we discuss the ability of virtual metrology (VM) applied in delay MIMO processes

Keywords:
Run-to-run control, metrology delay, MEWMA controller, MIMO

廣義近乎保序迴歸及其應用

黃士峰、方奕婷*

國立高雄大學統計學研究所

摘要

本研究提出一個廣義近乎保序迴歸（Generalized Nearly Isotonic Regression，簡記為 GNIR）的無母數統計方法，主要用來描述資料的主要趨勢變動，GNIR 具有自我調整以捕捉資料大趨勢時而向上、時而向下的能力。本研究亦提出 GNIR 方法的演算法，並證明其收斂性。在數值研究方面，我們採用 2006 年至 2017 年間全球 13 個金融市場的每日風險值（Value-at-Risk，簡記為 VaR ）序列進行實證分析，數值結果顯示 GNIR 不僅可以有效地捕捉 VaR 序列的變動趨勢，也能透過所配適的 GNIR 定義出每一個市場的風險狀態序列，接著運用關聯分析建立 13 個金融市場間的關聯性，實證結果發現應用所建立的金融市場關聯性可有效地提升對任一金融市場風險狀態的預測準確度。

關鍵詞：關聯分析、近乎保序迴歸、風險值

日內高頻配對交易與台灣股市之實證研究

林育詩
國立臺北大學統計學系

摘要

本文透過擴散過程均值回歸模型觀察配對組合是否出現異常定價,當異常價差出現時則交易觸發，此時買入低估的股票,賣出高估的股票，根據歷史資訊我們期待價差重新回到均衡狀態,當異常被修正則進行相反的策略平倉以獲取報酬。本文將交易期限制在當日完成，以每 5 分鐘的報價作為判斷依據，利用的極為短促的市場變化尋求可獲利的交易。極度頻繁的交易和微小的股價變化是高頻交易獲利的途徑，但也因此高頻交易下交易次數較傳統交易來得頻繁，如何在每筆交易平均獲利較小的情形下扣除交易成本仍能獲取超額報酬是本文努力的方向。

關鍵詞：高頻交易，配對交易，日內交易，均值回歸模型

滯後自迴歸條件異質變異模型之應用

黃士峰、張益誠*

國立高雄大學統計學研究所

摘要

本研究透過滯後自迴歸條件異質變異模型(簡記 HAR-GARCH)建立全球金融市場景氣狀態網絡圖，利用 HAR-GARCH 對金融時間序列資料進行建模，並採用蒙地卡羅馬可夫鏈演算法估計模型參數，同時獲得市場景氣狀態的估計。實證研究採用 13 個全球金融市場主要指數於 2008 年 8 月 1 日至 2018 年 8 月 30 日間的資料進行模型配適，以及為該 13 個金融市場建立逐月市場狀態網絡圖，並透過各種網絡特徵指標，進一步了解各市場在調查時間內所發生之金融事件中的相互影響關係。

關鍵詞：滯後自迴歸條件異質變異模型、蒙地卡羅馬可夫鏈法、網絡

The Effects of Social Media on Firm Performance: Evidence from Consumer Industry

Ju-Feng Yen (顏汝芳)、Wei-Jen Chang (張維仁)*

Department of Statistic

National Taipei University

Abstract

With the rapid development of electronic commerce, more and more firms are aware of the importance of network marketing. In recent years, business model changes from traditional mode to new system that combines consumers and firms via social platform. Amount of users on facebook, the most popular social media, has exceeded one billion. Following by more importance on social media, the degree of public opinion has risen. Especially, B2C (business to consumer) are industries facing consumer directly. In addition to creating a convenient shopping platform, paying attention to consumer opinions is more important. Therefore, performances of users in Facebook are especially important. For example: Like, Share, Comment. These data are combined a very big data sets for long time.

This study focus on consumer industries of S&P500, including Consumer Staples and Consumer Discretionary. This study examines whether and how engagement of official facebook fanpage of a firm affects its firm performance, including accounting and stock performance. This study further examines whether great engagement of social media can diminish firms' information asymmetry. The data period was from 2011 to 2018, and about one hundred companies were collected.

Keywords：Social Media、Facebook、S&P 500、Compustat、CRSP、Financial Statistic、Regression Analysis、Two-stage least squares regression

經濟衰退期間違約損失率之模型探討

鍾麗英、葉家齊

國立台北大學統計研究所

摘要

違約損失率（Loss Given Default,簡稱 LGD）是新巴塞爾協定進階內部評等法中用以衡量損失和估計計提資本的重要變數之一，而經濟不景氣時的違約損失率會高於一般時期，所以新巴賽爾資本協定(Basel II)要求銀行必須使用經濟不景氣時的違約損失率估計，以得到一個較為保守的計提資本。但是銀行不會在經濟不景氣時才估計違約損失率，必須用透過平常時期估計的違約損失率去轉換。相較於經濟衰退期間之違約機率（Conditional Probability of Default,簡稱 CPD），Basel II 中並沒有明訂關於經濟衰退期間違約損失率的轉換函數，這也是很多學者研究的目標。

經濟衰退期間違約損失率的轉換函數通常是透過推導違約機率的分配，並給定一個較為極端的分位數（表示經濟狀況不佳）之後求得。關於違約損失率的分配，多數的研究是假設為常態分配之後再進行後續的推導。我們認為違約損失率的分配不一定是常態分配，也可能是 T 分配、Normal Inverse Gaussian 分配、Variance Gamma 分配等。我們透過不同的分配假設，並比較不同的分配假設之下轉換函數的表現,探討哪一個分配推導出的轉換函數為違約損失率合理估計值。

關鍵詞：新巴賽爾資本協定、風險管理、違約損失率、風險計提資本

不同距離測度下的配對交易之獲利比較

薛愛蓉*、白惠明

國立台北大學統計學系，台灣 台北

摘要

配對交易是一種市場中性的投資策略，尋求具有高度相關性的股票，當配對股票的價格差（Spreads）偏離歷史均值時，可以從中獲利。Gatev 等人 (2006) 定義了配對交易中所使用的距離方法，而距離方法的基本概念是計算 MSD 值，選取較小的值作為我們的配對股票。雖然同樣是計算 MSD 值，但是利用不同的方式對股價做調整，所配對出的股票也會有所不同。本篇論文以 2013 年 1 月 1 日起至 2016 年 12 月 31 日止為樣本期間，利用台灣 50 指數成分股，去除部分缺漏值過多個股，共計 43 檔股票，進行一對一配對交易策略。利用不同的配對股價路徑距離測度選出較佳的配對股票做投資組合，並比較各種測度方法所得的報酬差異。

關鍵字: 配對交易 , 距離方法 , 統計套利

Self-Updating Process for Functional Data Clustering

Han-Chieh Chen(陳涵傑)* and Shang-Ying Shiu(須上英)
Department of Statistics, National Taipei University, Taipei, Taiwan

Abstract

The self-updating process (SUP) performs clustering on the basis of samples' movements according to between-sample associations. It has been shown that SUP is competitive in clustering data with a large number of clusters and data with noise. In this talk we will present an extension of SUP to functional data clustering. We represent functional data by eigenfunctions from functional principal component analysis (FPCA). At the initial iteration, each sample curve is represented by a set of eigenfunctions. In the end of the iteration, samples that are represented by the same set of eigenfunctions are considered to be in the same cluster. A comparison of this procedure with some existing clustering methods is presented by simulations.

Keywords：Functional data, Clustering, Functional Principal Component Analysis

基於階層式分群與主動式學習法進行軌跡分類

楊婷穎*、許湘伶
國立高雄大學統計學研究所

摘要

近年來，用於蒐集位置資訊的技術快速發展，各類軌跡易於取得且資料量增長迅速，使得軌跡資料能有多方應用，此處聚焦於軌跡標記與分類研究。對於軌跡分析，特徵萃取是一重要分析程序，於研究所需的初步特徵萃取問題上，使用 Lee 等人（2008）提出的特徵生成架構，其概念是基於探索區域及軌跡群集，進而擷取軌跡特徵，前者欲解析出具有相同軌跡類別之區域；後者則根據物體移動模式挖掘同類別軌跡，使得在原有軌跡特性外，捕捉局部軌跡片段額外產生的資訊。於研究特徵萃取部分，基於上述方法並引進粒子群優化算法（Particle Swarm Optimization）於區域的特徵提取過程，觀察並討論粒子群優化算法帶來的效益。然而，對於精確的軌跡資料標記常見的處理模式可能使用大量人力來解決類別資訊缺失之問題，由於收集程序與成本因素，多數軌跡為無標記資料，故而於研究上將所得軌跡特徵資訊導入主動式學習法於分類模型之建立，評估此類特徵萃取策略對於模型建構之幫助。

關鍵詞：主動式學習、軌跡分類、軌跡標記、粒子群優化算法、階層式分群

機器學習在類別資料分群應用

An application of machine learning on categorical data clustering

詹雯琦*、李孟峰

國立台北大學

摘要

資料分群的因素在於相似性度量，對於連續型變數通常使用歐式距離作為相似度測量，對於離散型資料在每一個變數的相似度一般以 0、1 兩個不同的測度來測量。分群的方法不是以統計模型為依循，而是以演算法達到最佳化；因此，機器學習的方法被廣泛應用於資料分群上。有別於連續型資料的 k-means 分群法，對於類別型 k-modes 是一個較常被使用的分群法。

在機器學習的方法中，初始值的選取都是使用隨機的方式。因此，容易造成收斂的不一致性。本文除了使用隨機方法外，將利用 HUANG(1998)、CAO(2009)這兩種尋找初始值的方法來進行比較，並對不同型的資料提供較適合初始值決定方式。

關鍵詞：機器學習、類別資料、k-modes

基於基因演算法之區間雙支持向量迴歸的延伸與比較

劉炳男*、許湘伶
國立高雄大學統計學研究所

摘要

支持向量迴歸為機器學習眾多方法之一，藉由最佳化技術建立數據的分析模型。Peng 等人於 2015 年提出應用於區間資料（interval data）之雙支持向量迴歸（interval twin support vector regression, ITSVR），其以兩組非平行的超平面對區間資料的上下界各自建立迴歸模型。本次報告將進行二種 ITSVR 改進方法於區間資料配適之特性比較，其一改善法考量到計算複雜度，導入加權最小平方法以減少建模之時間成本；另一改善法為引入 Xu 等人於 2018 所提非對稱 ν 雙支持向量迴歸至區間資料建模，藉以處理不對稱噪聲之影響進而改善建模表現。在此考量的方法均有特定參數需調整，故於本研究中導入基因演算法以縮減尋找最佳參數組合的運算時間成本，並以模擬實驗評比 ITSVR 與其改進方法對於不同區間資料建模之優劣。

關鍵詞：ν 支持向量迴歸、加權最小平方法、基因演算法、區間數據資料、區間雙支持向量迴歸

利用卷積神經網路提升單張圖片解析度且不損失像素

許巍瀚
國立臺北大學統計系

摘要

超解析度成像(Super-resolution imaging，簡稱 SR)，是一種提高影像解析度的技術，其中影像可以分為影片以及圖片，而這項技術會用於圖像處理已移籍超高解析度顯微鏡，在本篇論文中，我們謹對於單張的圖片進行超解析成像的處理。

而目前對於提高解析度的方法普遍為最鄰近內插法 (Nearest - neighborhood interpolation)、雙線性內插法 (Bilinear interporlation)以及雙立方內插法 (Bicubic interpolation)。在本篇論文中我們將使用不同於目前傳統的方法來進行超解析度成像的技術。

在C. Dong 等人提出的CNN(Convolutional Nueral Networks)架構可以完成單張圖片的超解析，但是在解析後會有損失像素的問題存在，在本篇論文將用 Zero-Padding 來解決這樣的問題。而損失像素即為輸入端的圖像經過卷積層的計算後輸入及輸出的像素不相等，通常是輸出端的像素小於輸入端的像素，而

Zero-padding 即是在輸入端的圖像上在圖像周圍補 0，進而維持著輸出端的大小不會與輸入端產生差異。

關鍵詞：

超解析度成像、卷積神經網路、損失像素、Zero-Padding

含外生多變數之時間數列門檻模式模型分析與預測

王治鈞
政治大學應用數學系

吳柏林
政治大學應用數學系

摘要

研究目的: 探討含外生變數之時間數列門檻模式及其應用。 研究方法: 利用隱性變數找出模型之門檻值，並考慮系統內能變化修正預測。 研究發現: 含外生多變數之模糊時間數列門檻模式模型分析與預測。 研究創新: 提出以含外生多變數之門檻模式架構方法。 研究價值: 提出用模糊熵來做預測修正，增加預測之準確度。 研究結論: 本研究建構之模式，均優於傳統的模式分析與預測。
關鍵詞: 外生多變數　時間數列　門檻模式　預測

Simulation system of high-frequency trading data

Cheng-Hsun Wu* (吳承勳)、 Mei-Hui Guo (郭美惠)

國立中山大學應用數學系

摘要

High-frequency limit order book is made up of a large number of high-frequency limit orders and transaction data received by the stock exchange. It also provides important details of market information, which implies that one can use data mining techniques to better understand the high-frequency trading phenomena and constructing useful trading strategies.

In this study, we analyze the limit order data from the LOBSTER website (https://lobsterdata.com/). Use the limit order data to construct a simulator for market trading behavior. One of the objectives of the simulator is to predict the future profit of transaction request which the user sent. The predictive model is divided into an action-based model and a time-based model. We use XGBoost as an action-based model. And the time-based model includes the ARMA time series model and other neural network models such as LSTM and GRU. We will use ensemble learning method to combine the advantages of both models.

關鍵詞：High-frequency limit order book, simulator, XGBoost, ARMA, LSTM.

廣義偽吉氏分佈的循環與分群

蕭惇中*、郭錕霖
國立高雄大學統計學研究所

摘要

在偽吉氏取樣的研究中，已發現不相容的條件機率模型下不同取樣順序所得的偽吉氏分佈將不盡相同，具有某些特殊性質，如:循環。在偽吉氏取樣的研究中有一部分的研究不同於傳統系統性的抽樣，而是採用隨機取樣的方式來進行，用此方式進行採樣最後所得之唯一平穩分佈與傳統系統性抽樣之偽吉氏分佈在二維上已經被證明兩者是有關係的。

本研究基於三維變數的偽吉氏取樣，並考慮多樣化的取樣順序，其對應的平穩分佈稱為廣義偽吉氏分佈，我們將對這些廣義偽吉氏分佈的循環與分群之性質進行探討。

關鍵詞：偽吉氏分佈、吉氏取樣、平穩分佈

高維度資料分析癌症的基因分子改變量之 shiny 應用程式

宋展毓
國立中山大學應用數學系

摘要

本研究主要探討卵巢癌以及膀胱癌，此兩個癌症分別都是台灣女性以及男性常見癌症的前十名。醫院治療病患除了參考臨床數據之外，有些醫院也會參考病患的基因表現量、蛋白表現量以及基因突變的變化，來判斷病人的存活風險。因此本篇研究想要探討基因表現量和突變是如何影響病人的存活、復發以及臨床狀態。研究資料來自研究綜合分析癌症基因和臨床資料開放平台(cBioPortal for Cancer Genomics)，透過蕭涵(2018)提出的正規化 Cox 模型與邏輯斯迴歸模型分析高維度資料方式找出重要基因以及模型預測能力，並且提供篩選出的重要基因給生醫所當作實驗參考。

但由於要進行上述的正規化統計分析需要一些程式語言背景，因此本研究也開發了 shiny 互動式平台提供不會程式語言的人，可以只透過點選的方式就可以進行高維度的基因資料分析。

關鍵詞：卵巢癌、膀胱癌、基因表現量、蛋白表現量、基因突變、分析癌症基因和臨床資料開放平台、正規化、高維度資料、shiny

Robust ridge regression: Applications of Nikkei data

Ting-yu Lin and Takeshi Emura
Graduate Institute of Statistics, National Central University

Abstract

In multiple linear regression models $\boldsymbol{y} = \boldsymbol{X\beta}$, where $\boldsymbol{y} = (y_1, y_2, \dots, y_n)^{\mathrm{T}}$and $\boldsymbol{X}$ is the n by $(p+1)$ design matrix with intercept and $\boldsymbol{\beta} = \left(\beta_0, \beta_1, \dots, \beta_p\right)^{\mathrm{T}}$ are unknow parameters and $\boldsymbol{\varepsilon} \sim N_n(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n)$, the least squares estimator $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X^T X})^{-1}\boldsymbol{X^T y}$ is usually used, but this method is not suitable for the models with multicollinearity. Therefore, Hoerl and Kennard (1970) proposed the ordinary ridge regression estimator $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X^T X} + k\boldsymbol{I})^{-1}\boldsymbol{X^T y}$, $k \geq 0$ is used here to solve the problem of the least square estimators under multicollinearity. However, the application of ridge becomes difficult if the linear model has the intercept term. This paper considers the robust regression estimators and the special class of robust ridge estimators (Askin and Montgomery, 1980 & Pfaffemberger and Dielman, 1984, 1985). Then we use the Influence functions to check the outliers, refer the new method to estimate $\boldsymbol{\beta}$ apply it to the Nikkei NEEDS data [1,2] to study the relationship between $\boldsymbol{y}$(dividend) and x_1(capital)、x_2(income)、x_3(retain).

Keywords ***Influence function*** **·** ***Multicollinearity*** **·** ***Outlier*** **·** ***Robust estimator*** **·** ***Ridge estimator***

具 ED 過程之兩因子加速衰退試驗建模

洪嘉妤*、樊采虹
國立中央大學統計研究所

鄭順林
國立成功大學統計學系

摘要

具有一因子的加速衰退試驗(ADT)常用來推論高可靠度產品的壽命,但可能影響產品壽命的不只一個因子。本文考量了具 ED 過程之兩因子恆定應力加速衰退試驗(CSADT)的隨機過程模型，Tweedie ED 過程為 ED 過程一種的重要類別，常見的三種隨機過程維納(Wiener)、伽瑪(Gamma)、逆高斯(Inverse-Gaussian)過程都是 Tweedie ED 過程的特例，且以數值方法推估產品壽命。

並考慮兩因子試驗可能有交互作用影響，將交互作用列入模型的考量。

最後探討兩組兩因子加速衰退資料，考慮兩因子是否有交互作用及不同的加速模型，與維納、伽瑪、逆高斯及 Tweedie ED 過程組合不同的隨機過程模型，比較在選模上 Tweedie ED 過程模型具有一定的優勢。

關鍵詞：兩因子加速衰退試驗、ED 過程、交互作用 、加速模型