A symbolic data analysis approach to regularized sliced inverse regression for gene expression data with multiple functional categories and a phenotypic response

Han-Ming Wu (吳漢銘)

Department of Statistics, National Taipei University

Abstract

Gene expression data such as those obtained from the hybridization microarray, the serial analysis of gene expression (SAGE) and/or RNA-Seq is being used to study a phenotypic response of interest. It is often characterized by a large amount of genes but with limited samples. Also, a priori knowledge of genes such as the functional and/or curated annotations is accumulated and available over the years. This study intends to incorporate both the biological knowledge of genes and the information of a discrete phenotypic response of subjects into dimension reduction through the framework of symbolic data analysis (SDA). The proposed approach consists of two steps. Firstly, the concepts of the symbolic data analysis will be used to aggregate the genes expression levels into functional intervals according to their functional categories. For unknown genes, we perform the gene selection procedures to select fewer genes that differentiate subtypes of a phenotypic response. The selected unknown genes are further aggregated into the intervals. Secondly, the regularized sliced inverse regression for interval-valued data is applied where the information of a phenotypic response of subjects acts as the slices. We illustrate the proposed method using several public gene expression data sets for data visualization and the class prediction. The results are compared with those of the regularized PCA. The results show that the proposed method can achieve better performance in understanding biologically relevant processes of genes and subjects than purely data-driven models.

Key words: data visualization, interval-valued data, symbolic data analysis, sufficient dimension reduction, gene expression, biological knowledge.

Predicting One-day-ahead Wind Power Capacity Factor via Functional Inverse Regression

Lu-Hung Chen
National Chung Hsing University

Ci-Ren Jiang*
Academia Sinica

Abstract

Inverse regression is an appealing dimension reduction method for regression models with multivariate covariates. Recently, it has been extended to the cases with functional or longitudinal covariates. However, the extensions simply focus on one single functional or longitudinal covariate. Motivated by a real application, we extend functional inverse regression to the cases with multiple functional covariates, whose domains could be different. The asymptotical properties of the proposed estimators are investigated. The computational issues are taken care with data binning, the fast Fourier transformation and random projections on a multi- core computation platform. In addition to simulation studies, the proposed approach is applied to predict the one-day-ahead wind power capacity factors in Germany from 2016 to 2017. Both demonstrate the good performance of our method.

keywords：functional data, inverse regression, smoothing

Simultaneous confidence bands for functional regression models

Chung Chang

Department of Applied Mathematics, National Sun Yat-sen University

Abstract:

In recent years, the field of functional data analysis (FDA) has received a great deal of attention, and many useful theories and interesting applications have been reported. One topic of particular interest involves estimation of simultaneous confidence bands (SCB) for an unknown function. Degras (2011) proposed an estimator of SCBs for the mean function in a simple (no covariates) function-on-scalar regression model that relies on some assumptions on the tail behavior of the errors. In the case that such distributional assumptions do not hold, Degras also proposed a bootstrap method (sampling with replacement). We consider a more general function-on-scalar regression model that involves multiple covariates and allows the variance function of the functional responses to be dependent on the covariates (heterogeneity). In this general model, we propose a wild bootstrap method for estimating SCBs for the coefficient function. Some asymptotic results are provided for the simple case (no covariates) and simulation results for both the simple and general models.

Keywords:

Simultaneous confidence bands, wild bootstrap, functional regression

Diagnosis of Myocardial Perfusion Images for Coronary Heart Disease by 3D Convolutional Neural Network

Jui-Jen Chen#
Engineer, Department of Nuclear Medicine, Chang Gung University College of Medicine, Kaohsiung, Taiwan

Ting-Yi Su#
Institute of Statistics, National Chiao Tung University, Taiwan

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

Yen-Hsiang Chang
Attending Staff, Chang Gung Memorial Hospital, Chang Gung University College of Medicine, Kaohsiung, Taiwan

Shu-Hua Huang
Chief, Kaohsiung Medical Center, Chang Gung University College of Medicine, Kaohsiung, Taiwan

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

ABSTRACT

This study focuses on the classification of myocardial perfusion images for coronary heart disease by deep learning techniques. In these grayscale images, the central bright region contains the most important features. Therefore, the data-driven preprocessing is developed to extract out the region of interest. After removing the surrounding noise, the 3D convolutional neural network model is utilized to classify whether the patient has coronary heart disease or not. The prediction accuracy, sensitivity and specificity are 91.28%, 88.37% and 94.19% in this study based on the images collected at Chang Gung University College of Medicine in Kaohsiung. It can assist clinical experts to diagnose coronary heart disease accurately in practice.

Keywords: deep learning, 3D convolutional neural network, myocardial perfusion image, coronary heart disease

Presenter: Ting-Yi Su
Correspondence to: Henry Horng-Shing Lu
# Contributed equally

Performance of a two-sample test with Mann-Whitney statistics under dependent censoring

Jiung Huang Hsu (許竣瑝)
Graduate Institute of Statistics, National Central University

ABSTRACT

The Mann-Whitney statistics is a measure for testing the equality of two survival functions in a two sample problem. That is involved the survival function for two independent groups. According to the asymptotic inference from Dobler (*Test, 27*(3), 639-658, 2017), the two sample problem can be tested effectively under the large sample size. Traditionally, the survival functions are estimated by the Kaplan-Meier (KM) estimator. However, the KM estimator is derived under the independent censoring assumption. Therefore, the KM estimator is biased under the dependent censoring. In order to solve this problem, we use the copula-graphic method to estimate survival function in this paper. The simulated results are performed well by using the proposed method and the asymptotic inference from Dobler (*Test, 27*(3), 639-658, 2017). Finally, we analyze a real data from Chen (*Commun Stat Simul Comput, 23*(1), 1-16, 1994).

Keywords *Mann-Whitney effect • Kaplan-Meier estimator • Copula • Copula-Graphic estimator • Dependent censoring*

探討抽樣誤差於長期追蹤常態資料之一致性相關係數估計

李佳音
國立彰化師範大學統計資訊研究所

摘要

在臨床研究中，常透過一致性相關係數(Concordance correlation coefficient; CCC) 評估連續的重複判讀資料之間的可靠性。然而，在許多領域中早已發現樣本抽樣誤差的問題，其中包括臨床試驗、流行病學研究、基因組織研究和野生動物管理。因此，本文主要探討在抽樣誤差下，樣本分佈以及樣本個體數對 CCC 估計的影響，且使用三種一致性指標來判讀重複讀數間的一致性，其分別為方法內的一致性(intra-method agreement)、方法之間的一致性(inter-method agreement)和整體方法間的一致性(total-method agreement)，並將連續型的重複判讀資料建構在線性混合模型(linear mixed model; LMM)下，且利用變異數成份(Variance components; VC)以及 U 統計量(U-statistic; US)進行一致性指標的估計及推論，並從偏差、變異數估計及均誤差來比較兩者之表現。從模擬結果中顯示，當抽樣誤差造成樣本來自非常態分佈時，其估計值有高估的現象，因此，本論文提出一個重抽樣的程序，將非常態分佈的樣本經由重抽樣程序後則可更接近常態分佈，如此以降低個體分佈對一致性指標估計值的依賴性，最後，將此程序應用在學童近視成因長期追蹤判讀資料的研究上。

關鍵詞：抽樣誤差、個體分佈、一致性相關係數、線性混合模型、變異數成份

探討抽樣誤差於長期追蹤 Poisson 資料之一致性相關係數估計

黃湣然
國立彰化師範大學統計資訊研究所

摘要

在連續型臨床研究資料中，一致性相關係數(Concordance correlation coefficient; CCC)被廣泛的用來評估兩種測量方法之間的一致性。而在臨床研究中亦需要評估離散型資料的一致性，且在不同的領域中發現，一致性相關係數會因為抽樣誤差而導致估計偏差。因此，本文將探討當使用變異數成份(Variance components; VC)和 U 統計量(U-statistic; US)兩種方法來估計數計型重複判讀資料時，在不同的抽樣誤差情形下，造成一致性相關係數的估計偏差情形。本文將長期追蹤重複數計資料建構在 Poisson 混合效應模型(Poisson mixed-effects model)上，並分別估計方法內的一致性(Intra-agreement)、方法之間的一致性(Inter-agreement)和整體方法間的一致性(Total-agreement)。在模擬分析中，當抽樣誤差導致樣本的分佈服從非 Poisson 分佈時，會造成一致性相關係數估計值的偏差。因此，本文提出重抽樣步驟將非 Poisson 分佈的樣本進行重抽樣，使樣本更接近 Poisson 分佈，如此以降低個體分佈對一致性相關係數估計值的影響。

關鍵詞：一致性相關係數、長期追蹤重複資料、Poisson 混合效應模型、變異數成份

運用 ROC 曲面下體積評估三元資料判定準則的敏感度分析

胡慧禪*、黃怡婷
國立臺北大學 統計學系

摘要

醫學診斷常需要依病人的疾病嚴重狀況進行分級，運用生物標記探討病人分級的準確性，若僅需分成兩類，文獻多採用接收者操作特徵曲線（Receiver operator characteristic curve; ROC） 與曲線下面積來評估準確性，且有詳盡的討論。若有三類分級時，目前主要使用曲面下體積（Volume under ROC surface）做為衡量的指標，但有較少的文獻探討該指標的性質。再者，這類醫學診斷分級資料，實務應用會針對每一級別收集資料，如僅有兩個類時，則分級依據採用兩級資料的最大值，但若資料有三類或是分布不平均的時候，分級依據則無法採用最大值。

李嘉泰(2018)提出三級分類資料的四種判定準則，來進行醫學診斷分級，但資料分布方式可能會影響這些判定方式的表現。在多種參數設定情境下， 本論文主要運用統計模擬來探討判定準則在多變量常態分配，狄利克雷分配(Dirichlet distribution)與多變量偏斜常態分配(Multivariate Skew Normal distribution)的表現，利用 VUS 指標來評估判定方式的表現。

關鍵詞：三元資料，ROC 曲面下體積，狄利克雷分配, 偏斜常態分配，判定準則

A meta-analysis approach to causal mediation modeling of semi-competing risks

Shu-Hsien Cho*(卓書賢), Yen-Tsung Huang(黃彥棕)
Institute of Statistical Science, Academia Sinica, Taipei, Taiwan

Abstract

We recently proposed a model-based causal mediation analysis of semi-competing risks. Despite the novel perspective of semi-competing risks under the framework of causal mediation inference, the computation cost of the method makes it not scalable to a very large data set such as National Health Insurance Research Database (NHIRD). To address the issue, we propose a meta-analysis approach dividing the large dataset into subgroups and pooling the results in each subgroup into a summary estimate. The finite sample efficacy of the meta-analysis approach is illustrated by extensive simulation. We evaluate empirical bias of the pooling estimators under various weighting schemes as well as their empirical relative efficiency and computation cost, compared with the estimator using the full dataset. We also investigate the performance under different combinations of group size and sample size in each subgroup. Numerical simulations suggest that the meta-analysis approach achieves similar statistical efficiency as the full sample-based estimators, and substantially outperforms the full sample-based estimators with respect to the computational speed. The meta-analysis approach has the practical utility in applying causal mediation analyses of semi-competing risks to a large dataset.

key words: Causal mediation model; Cox proportional hazards model; Meta-Analysis; semi-competing risk.

Bayesian Inferences of Multiple Structural Change GARCH Model with Skew Student t Errors

Bonny Y.F. Lee*and Cathy W.S. Chen
Department of Statistics, Feng Chia University, Taichung, Taiwan.

摘要

This research considers a piecewise autoregressive GARCH model with exogenous variables and skew Student t errors, which we call this model a segmented ARX-GARCH model and use it to make inferences about all unknown parameters and to identify the location of structural breaks. It fills the gap in existing literature to cover skew Student t errors. Compared with other distributions, the fat-tailed skew Student t distribution performs well at describing financial time-series datsets in financial markets. We employ the segmented ARX-GARCH model with skew Student t errors, and estimate the model parameters via Bayesian inference, in order to show the validity and reliability of the Bayesian methods. We then utilize the adaptive Metropolis-Hastings (MH) MCMC algorithm, which combines with the random walk MH algorithm and the independent kernel MH algorithm to accelerate convergence. Our goal is to know how many breakpoints and the location of the breakpoints. We first assume the number of breakpoints is prefixed and employ deviance information criterion to decide the optimal number of breakpoints. We also extend the segmented GARCH model to an asymmetric GARCH one. As an illustration, we provide a simulation study to examine the credibility of our MCMC sampling scheme. For real data analysis, we examine the impact of daily crude oil returns and gold returns on stock S&P 500 returns during 2007 to 2018.

關鍵詞：structural change, skew Student t distribution, segmented ARX-GARCH model, Markov chain Monte Carlo methods, Bayesian inference, stock market.

Semiparametric causal mediation modeling of semi-competing risks

Ju-Sheng Hong*(洪鉅昇), Shu-Hsien Cho(卓書賢), Yen-Tsung Huang(黃彥棕)

Institute of Statistical Science, Academia Sinica, Taipei, Taiwan

Abstract

Semi-competing risks are frequently encountered in biomedical research in which a primary outcome (e.g. death) may censor an intermediate event (e.g. cancer incidence) but not vice versa. We propose a semiparametric approach formulating the semi-competing risks as a causal mediation problem. We develop a mediation model with the intermediate and primary events, respectively as the mediator and the outcome. Counting process-based indirect and direct effects, respectively, are defined as an effect of an intervention on the primary outcome mediated through the intermediate event, and that not mediated through the intermediate event. We construct Breslow-type hazard estimators for direct and indirect effects, with time-varying weights. Asymptotic properties are established for the proposed estimators. Using simulations, we evaluate the finite-sample performance of the proposed estimators. The utility of our proposed methods is illustrated in a hepatitis study of liver cancer survival.

key words: Causal mediation model; Counting process; Cox proportional hazards model; Nonparametric maximum likelihood estimator; Semi-competing risks.

A non-zero mean uniform linear array with Fielder spatial correlation matrix

許哲瑋、林財川
國立台北大學

摘要

ULA is widely applied in many fields, e.g. engineer, military, communication, sonar, etc. Most of references assume the sensors are temple or spatially white. Only small part of ULA references consider a so-called von Karman spatial correlation model, which is derived from the physical philosophy and the expression is complicated. In this paper, we propose a non-zero mean ULA along with the Fielder matrix for describing the spatially coherence of model. We derive the Cramer Rao Low bound (CRLB) for the direction-of-arrival under various cases in a closed form. Theoretical find that the CRLB with non-zero mean always smaller than zero mean, the CRLB decrease as number of sensor or number of snapshots increase, and the CRLB decrease as noise decrease.

關鍵詞： antenna arrays, uniform linear array (ULA), array signal processing, spatial decorrelation CRLB

Model diagnostic procedures for copula-based Markov chain models for statistical process control

Xin-Wei Huang*(黃昕蔚) and Takeshi Emura(江村剛志)

Graduate Institute of Statistics, National Central University, Taiwan

Abstract

Investigating serial dependence is an important step in statistical process control (SPC). One recent approach is to fit a copula-based Markov chain model to perform SPC, which provides an attractive alternative to the traditional AR1 model. However, methodologies for model diagnostic have not been considered. In this paper, we develop two different approaches for model diagnostic procedures for copula-based Markov chain models. The first approach employs a formal test based on the Kolmogorov-Smirnov or the Cramér-von Mises statistics with aid of a parametric bootstrap. The second approach employs the second-order Markov chain model to examine the Markov property in the model. This second approach itself is a new SPC method. We made all the computing methodologies available in the R *Copula.Markov* package, and check their performance by simulations. We analyze three datasets for illustration.

Keywords: Control chart; Copulas; Goodness-of-fit tests; Markov chain; Serial dependence; Statistical process control; Time series

Modeling Association between DNA Copy Number and RNA Expressions on Paired and Unpaired COAD Patients

Yu-Ru Liao (廖鈺茹)* and Shuen-Lin Jeng (鄭順林)
National Cheng Kung University

Abstract

The association between DNA Copy Number (CN) and RNA expressions is an important issue in cancer studies. The critical genes with strong DNA-RNA association may serve as therapeutic targets. In this study, we explore several methods to identify genes with strong DNA-RNA association in specific groups of cancer patients. We analyze the patients with the Colon Adenocarcinoma (COAD) downloaded from the Genomic Data Commons (GDC). We used the toolset "Bedtools" and the annotation "unoverlap-GRCh38" to establish a new calculation read count method. The new calculated read count method is to solve the problem that the exons of the different gene are overlapping. After obtaining the read counts, we using maximal information coefficient (MIC), distance correlation (dCor) and Multivariate Adaptive Regression Splines (MARS) to find out the two-dimensional relationship between DNA copy number and RNA expressions. The innovative algorithm in this study is called Appro2dgMIC, which is able to calculate the three-dimensional relationship between the DNA copy number and RNA expressions across genes.

Keywords：COAD, MIC, dCor, MARS, DNA, RNA, copy number

Causal mediation analysis with the mediator truncated by death in the survival study

An-Shun Tai* (戴安順)、Sheng-Hsuan Lin (林聖軒)
Institute of Statistics, National Chiao Tung University, Hsinchu, Taiwan

Abstract

The researches of causal mediation analysis with survival outcome have succeeded to develop many approaches to fit the settings with multiple mediators and different survival functions. However, the current studies are lack of discussing the truncation-by-death problem on mediators, and in that case, the traditional mediation formulation is invalid because the mediator for the subject truncated by death is undefined. To address this issue, this study makes a series of new assumptions and moreover provides a corresponding approach to identify the causal parameter from observed data. The regression-based method and inverse probability weight approach are both applied to the statistical inference in this study. We comprehensively compared the complete case analysis with the proposed method via the simulation study, and as a result, our research shows the advantage of estimating the causal effect in the case of the truncated mediator.

Keywords: Causal mediation analysis; Survival study; Truncation-by-death; Inverse probability weight.

Bayesian Approach to Genome-Wide Genetic Association Studies with Survival Time as outcome

Li-Hsin Chien[1], I-Shou Chang[2], Tzu-Yu Chen[1], Chung-Hsing Chen[1], and Chao A. Hsiung[1]

[1] Division of Biostatistics and Bioinformatics, Institute of Population Health Sciences, National Health Research Institutes, Taiwan.
[2]National Institute of Cancer Research, National Health Research Institutes, Taiwan.

In the cancer study of treatment response, it is common to consider time to progression or overall survival as the outcome variable. Since cancer is often a heterogeneous disease, treatment response may vary greatly even for a specific cancer type. For example, time to progression varies tremendously among late stage lung adenocarcinoma patients treated by TKI. It is of interest to know if there are any SNPs that can be used for outcome prediction when treated by TKI. The first step in building this prediction model is to look for SNPs that are associated with the survival time of interest (time to progression or overall survival). Current approach to this type of studies usually considers hundreds of thousands of SNPs simultaneously. We take a Bayesian approach to this association studies. In particular, we specify the prior distribution in terms of the realistic concept of heritability, widely accepted in genetics. In fact, one of the main advantages of this approach is the utilization of the concept of heritability. We use Markov Chain Monte Carlo to simulate the posterior distribution for inference and use a technique termed “small-world proposal” to improve the convergence rate of the Markov chain. Simulation studies will be provided to indicate the numerical performance of this method. We note that it is of interest to assess the computational burden in this type of problem.

Key words: Bayesian Approach, Survival, Heritability, Variable Selection

Asymptotic Normality and Bayes Factor Consistency for One-Way Multivariate Variance Components Models

Chun-Lung Su
Department of Statistics
Tunghai University

Abstract

We derive the asymptotic posterior for one-way multivariate variance components models under the assumption that the number of the main factor levels or the within sample size becomes large. The explicit asymptotics of balanced one-way multivariate variance components models for two different parametrizations under certain prior assumptions are given. Bayes factor consistency for testing variance components is also investigated.

Keywords: Asymptotic posterior, Bayesian factor consistency, One-way multivariate variance components, Sufficient reduction statistics

Efficient Estimation of a Semiparametric Zero-inflated Bernoulli Regression Model

Chin-Shang Li, Ph.D.

School of Nursing
The State University of New York, University at Buffalo
Buffalo, NY14214, U.S.A.
E-mail: chinshan@buffalo.edu

Abstract

When the observed proportion of zeros in a data set consisting of binary outcome data is larger than expected under a regular logistic regression model, it is frequently suggested to use a zero-inflated Bernoulli (ZIB) regression model. A spline-based ZIB regression model is proposed to describe the potentially non-linear effect of a continuous covariate. A spline, which can be expressed as a linear combination of *B*-spline basis functions, is used to estimate the unknown smooth function. The spline estimator of the nonparametric component is shown to be uniformly consistent and achieve the optimal convergence rate under the smoothness condition. The regression parameter estimators are shown to be asymptotically normal and efficient. A spline-based semiparametric likelihood ratio test is established, and a direct and consistent variance estimation method based on least-squares method is proposed. Extensive simulations are conducted to evaluate the finite-sample performance of the proposed method. A real-life data set is used to illustrate the practical use of the proposed methodology.

Ancestry-Informative Pharmacogenomic Loci in Global Populations

Hsin-Chou Yang* (楊欣洲), Chia-Wei Chen (陳佳煒), Yu-Ting Lin (林昱廷), Shih-Kai Chu (朱是鍇)

Institute of Statistical Science, Academia Sinica

Abstract

Ancestry informative markers (AIMs) are a type of genetic marker informative for tracing the ancestral ethnicity of individuals[1,2]. Previous studies observed AIMs enriched in expression quantitative trait loci[2] (eQTL) and associated with adverse drug reaction and drug response[3]. An example of ancestry-informative pharmacogenetic locus (PGx) is rs1045642 on Multi-Drug Resistance Gene (*MDR1*); it is an ancestry-informative eQTL associated with adverse drug reactions to amitriptyline and nortriptyline and drug responses to morphine. Identification of ancestry-informative PGx is beneficial to population precision medicine. This study analyzed more than 77 million of single nucleotide variation in The 1000 Genomes Project – Final Phase, which provided the whole-genome sequencing data of 2,504 individuals from 26 global populations. We identified AIMs and ancestry-informative PGx for global populations. The results highlight the importance of a proper consideration of population differentiation in pharmacogenetics studies. Public resources of ancestry-informative PGx have been established.

[1] Rosenberg, N.A., et al. (2003). Am J Hum Genet 73, 1402-1422.
[2] Yang, H.-C., et al. (2012). BMC Genomics 13, 346.
[3] Yang, H.-C., et al. (2014). BMC Genomics 15, 319.

卷積神經網路分類在大腦單光子電腦斷層掃描影像於帕金森氏症診斷之應用

王博正*、羅夢娜
國立中山大學應用數學系

朱基祥、徐健欽
高雄長庚紀念醫院
臨床試驗中心、核子醫學科

摘要

本研究主要目標為使用卷積神經網路（Convolutional Neural Network, CNN）搭配整體學習（Ensemble Learning），對大腦單光子電腦斷層掃描（Single Photon Emission Computed Tomography, SPECT）影像進行分類。使用依據醫師所挑選最具代表性之 5 張醫療數位影像傳輸協定（Digital Imaging and Communications in Medicine, DICOM）儲存格式中的影像，藉由建立多個分類模型進行整體學習，得到一最終分類模型。期在未來為輔助醫師診斷之用。

關鍵詞：大腦單光子電腦斷層掃描、帕金森氏症、卷積神經網路、整體學習、醫療數位影像傳輸協定

探討ＤＮＮ、ＣＮＮ和Ｃａｐｓ Ｎｅｔ於高混合度之中文母音辨識

辛祐任

國立中興大學統計學研究所

江翠蓮

中台科技大學食品科技系

李宗寶

國立中興大學應用數學系

摘要

本論文主要是探討深度神經網路(Deep Neural Networks, DNN)、卷積神經網路(Convolutional Neural Networks, CNN)、膠囊網路 (Capsule Networks, CapsNets)等三種網路在高混合度之中文母音如<ㄛ、ㄨㄛ> 、<ㄣ、ㄥ> 、<ㄥ、ㄨㄥ>等共 15 組之辨識。資料特徵採用梅爾頻率倒譜系數 (Mel-Frequency Cepstral Coefficients,MFCC)。DNN 利用 25 個獨立 MLP 組合來進行辨識；CNN 基本上採用 LeNet-5，以及對卷積層數、超參數及訓練次數做調適；CapsNet 參考 Hinton[2017]，試驗膠囊的維度及特徵圖對辨識的影響。實驗結果顯示 CNN 跟 CapNet 準確度相近，而 DNN 的正確率較低，以<ㄛ、ㄨㄛ>為例，DNN、CNN、CapsNet 三個模型之最高的準確率分別為 80.23%、88.43%、89.22%。

關鍵字: MLP、MFCC、DNN、CNN、CapsNet、Speech Recognition

機器學習分類法在單光子電腦斷層掃描影像於帕金森氏症診斷上之應用

莊彥騏*、羅夢娜
國立中山大學應用數學系

朱基祥、徐健欽
高雄長庚紀念醫院
臨床試驗中心、核子醫學科

摘要

本研究擬探討如何依據單光子電腦斷層掃描影像(Single Photon Emission Computed Tomography, SPECT)，協助醫師對受測者是否罹患帕金森氏症之診斷。擬從左右腦海馬迴影像中依據 K-近鄰演算法(K-Nearest Neighbor，KNN)以及交叉驗證法(Cross Validation)來萃取重要特徵，作為分類之解釋變數。然後利用隨機森林(Random Forest)建立模型，並給出各特徵變數重要性之排序，以提供是否罹患帕金森氏症診斷之建議。此外我們並利用接收者操作特徵曲線(Receiver Operating Characteristic，ROC Curve)、曲線下面積(Area Under Curve，AUC)等方式，及其對應之混淆矩陣(Confusion Matrix)來評估模型診斷之可靠度。最後再加入邏輯斯迴歸(Logistic Regression)、支持向量機(Support Vector Machine，SVM)及 Boosting 等方法，且用多數決的方式，做為最後的診斷結果。期能幫助醫師迅速得到影像判讀之結果。

關鍵詞：影像辨識、K-近鄰演算法、交叉驗證、隨機森林、邏輯斯迴歸

利用 HE 預訓練之 CNN 方法於中文單音之辨識

黃莘揚

國立中興大學統計學研究所

江翠蓮

中台科技大學食品科技系

李宗寶

國立中興大學應用數學系

摘要

本文利用卷積神經網路(Convolutional neural network, CNN)來對中文單音進行學習及辨識。主要實驗方向為把單音拆成子音與母音，並在同一模型下預測出子、母音類別，最後組合出單音。其中子音總有 36 個類別，母音則有 160 的類別，單音對母音組合則有 1391 個類別。資料特徵求取方法選用梅爾倒頻譜系數(MFCC)，並以此作為模型輸入數值。本論文將實驗不同卷積層層數、特徵圖(feature map)數和全連接層(full connection layer, FC)的層數、神經元個數對辨識結果的影響。同時地，會探討不同的活化函數(activation function)、初始化方法和 BN(batch normalization)、dropout 技術的有無是否會影響分類結果。實驗結果發現在使用 4 層卷積層、3 層全連接層並且使用何初始化(He initialization)和 BN 下獲得最高的單音辨識率。子音、母音和單音辨識率分別達到: 96.49%、97.40% 和 94.49%。

關鍵字: 卷積神經網路、MFCC、活化函數、初始化、dropout、BN

探討類神經 RNN 與 LSTM 方法於小資料中文單音之辨識

楊鎧瑞

國立中興大學統計學研究所

江翠蓮

中台科技大學食品科技系

李宗寶

國立中興大學應用數學系

摘要

本論文主要是應用遞歸神經網絡(Recurrent Neural Network, RNN)、長短期記憶神經網絡(Long Short-Term Memory, LSTM)，對部分高混合母音(如：ㄛ、ㄨㄛ等)及部分單音進行辨識，資料透過梅爾倒頻譜系數(MFCC) 求取特徵，並以此數值輸入模型。遞歸神經網絡中的隱藏層輸出會參與下一次的模型輸入，這使得遞歸神經網絡有了短期的記憶性，其中遞歸神經網絡又分為兩種，Jordan RNN 及 Elman RNN，兩者的差別在計算隱藏層輸出有不同的公式。LSTM 亦是 RNN 的一種，LSTM 較 RNN 多了輸入、輸出、遺忘，三個控制閥門，此外，在結構上新增了一個貫穿所有時間步(time steps)的方程式，這使得 LSTM 有長期記憶的功能。本篇採用的是單音資料，使用一層的 RNN 或 LSTM。本方法之辨識結果，Jordan RNN 在速度上最快速，Elman RNN 在小資料時表現的較 LSTM 佳，但在較大的資料時 LSTM 表現的最好。

關鍵字: 遞歸神經網絡、長短期記憶、梅爾倒頻譜係數

深度學習在不同維度資料下的異常檢測之研究

鍾麗英、蕭子修
國立台北大學統計學研究所

摘要

異常檢測(anomaly detection)也可以稱為離群檢測(outlier detection)或新奇檢測(novelty detection)，指的是對於不符合我們所預期的資料模式(pattern)進行辨識。異常檢測最困難的地方為異常的數據占比遠小於正常數據，因此可視為一種不平衡資料。所以確切偵測出每筆異常資料顯得額外重要。異常檢測的方法包含了統計方法、機器學習以及深度學習，相較於傳統的統計方法以及機器學習，深度學習藉由自我學習得到了良好的模型表現以及靈活性。由於深度學習藉由自我學習資料特徵，所以隨著數據的規模增加，深度學習表現會優於機器學習。

在深度學習中，主要分為監督式學習、半監督式學習，非監督式學習。三者主要差異為是否需要完整的資料的標籤。半監督式學習只需要正常樣本的標籤，藉由學習正常樣本的特徵，以去辨識異常的資料。而非監督式學習則完全不需要，藉由模型自我學習以判斷正常或者異常。

異常檢測(anomaly detection)在諸多領域上皆是一個重要的研究問題，本研究將利用深度學習裡的各種模型對不同維度資料進行異常檢測。在給定異常比例相同，資料維度不同的情況下，去比較各個模型表現優劣，以便未來在面對資料要進行異常檢測時，更能選出適合的模型，提高偵測效率。

關鍵詞：異常、新奇、離群、深度學習、

Asymptotic Theory of Conditional Generalized Information Criterion for Linear Mixed Effects Model Selection

ChiHao Chang、HsinCheng Huang
National University of Kaohsiung
Academia Sinica

ChingKang Ing
National Tsing-Hua University

摘要

We consider selecting linear mixed-effects models when both dimensions of fixed-effects and random-effects models may go to infinity with the sample size. We introduce a conditional generalized information criterion (CGIC) for model selection, which is extended from the conditional Akaike's information criterion of Vaida and Blanchard (2005). We establish the selection consistency and the asymptotic loss efficiency of CGIC with respect to a conditional Kullback-Leibler loss and the squared-error loss under mild conditions. Our asymptotic theory is applicable to unbalanced data and to linear mixed-effects models that contain only a finite number of clusters so that their random-effects parameters cannot be consistently estimable.

Conditional Generalized Information Criterion; Consistency; Asymptotic Loss Efficiency;

Empirical likelihood based summary ROC curve for meta-analysis of diagnostic studies

曾聖澧*
國立中山大學應用數學系
陳錦華
臺北醫學大學生物統計研究中心
臺北醫學大學大數據科技及管理研究所
陳春樹
國立彰化師範大學統計資訊研究所

摘要

Summarizing performance metrics is crucial in a systematic review of a diagnostic performance. There are various summary models for the performance metrics in the literature, and hence model selection becomes inevitable. However, most existing large-sample-based model selection approaches may not fit in a meta-analysis of diagnostic studies, typically having a rather small sample size. We proposed a modified empirical likelihood based method for selecting models given such small-sample problems. The selected model was then used for constructing summary receiver operating characteristic (sROC) curves, depicting the relationships between sensitivity and specificity for different cut-off points. Simulation studies were conducted by assuming different number of studies and various population distributions for the disease and non-disease cases. The performance of our proposal and other model selection criteria was also compared. We found that parametric likelihood-based model selection methods often fail to consistently choose appropriate models for summary under the limited number of studies. When the number of studies is as small as 10 or 5, our proposed method is best for several performance measurements. Therefore we recommend choosing a summary model via the proposed empirical likelihood method.

關鍵詞：meta-analysis, empirical likelihood, summary ROC curve, sensitivity, specificity

Spatial Regression Model Selection When Covariates and Random Effects Are Correlated

楊洪鼎
國立彰化師範大學統計資訊研究所

摘要

The spatial random effects model is popular in analyzing spatially referenced data. The model includes spatially observed covariates and unobserved spatial random effects, which if not deal properly with the confounding between the two components, parameter estimation and spatial prediction had been demonstrated to be unreliable. In this research, we focus on discussing the estimation of regression coefficients and the selection of covariates for spatial regression under the presence of spatial confounding. We first introduce an adjusted estimation method of regression coefficients and the consequent spatial predictor when spatial confounding exists. From a prediction point of view, we then propose a generalized conditional Akaike information criterion to select a subset of covariates, resulting in variable selection and spatial prediction that are satisfactory. Statistical inferences of the proposed methodology are justified theoretically and numerically. This is a joint work with Yung-Huei Chiou and Chun-Shu Chen.

關鍵詞：conditional information criterion, mean squared prediction error, restricted spatial regression, spatial prediction, variable selection.

Effects of statistical distributions of diseased leaves and numbers of classes in disease scales on the accuracy of estimate of mean disease severity

Jia-Ren Tsai* (蔡嘉仁)

Department of Statistics and Information Science, Fu Jen Catholic University, New Taipei, Taiwan.

Hung-I Liu (劉弘一)

Division of Biometrics, Department of Agronomy, National Chung Hsing University, Taichung, Taiwan.

Wen-Hsin Chung (鍾文鑫)

Department of Plant Pathology, National Chung Hsing University, Taichung, Taiwan.

Kuo-Szu Chiang (蔣國司)

Division of Biometrics, Department of Agronomy, National Chung Hsing University, Taichung, Taiwan.

Abstract

In agricultural research, an ordinal scale of measurement has often been used to estimate disease severity. The characteristics of the distribution of diseased leaves in the population and the number of classes in a disease scale often affect the accuracy of the estimates of mean disease severity. The purposes of this study are to compare various interval-scale estimates to nearest percent estimates and to further investigate

the effects of the number of classes in a disease scale. A simulation method was employed to execute the study. Moreover, real data from the field was used to verify the results. The criterion for comparison was the mean squared error of mean disease severity estimates for each of the different scales used for estimation. The results of this study indicate that, when preparing numeric category scales for rating disease severity, scales with grades of ≥7 are preferable as severities ≤ 50% disease is emphasized. Moreover, linear category scales with sensitivity to low disease severity are preferable to nonlinear category scales for assessing disease severity. We believe that the results of our study will be helpful in improving the accuracy of disease severity estimates in plant epidemiology and related areas of research.

Keywords: bias, disease scale; interval-scale estimates; mean squared error

Application of statistical methods in constructing stock abundance indices of Pacific bluefin tuna

Hung-I Liu* (劉弘一)
Overseas Fisheries Development Council of the Republic of China, Taipei, Taiwan

Shui-Kai Chang (張水鍇)
Institute of Marine Affairs, National Sun Yat-sen University, Kaohsiung, Taiwan

Abstract

Catch-per-unit-effort (CPUE) is frequently used as an index of relative fish abundance, and it is the main piece of information used in fisheries stock assessment. The catch and effort data traditionally obtained from commercial logbooks, however, are incomplete or unreliable in many cases. Pacific bluefin tuna (PBF) is a seasonal target species with high economic value in Taiwan's offshore longline fishery, but this fishery has few logbooks available for calculating CPUE. Therefore, several nontraditional procedures via statistical methods were performed to reconstruct catch and effort data from many alternative data sources for 2001–2015:(1) Estimating the catch number from the landing weight for 2001–2003, for which the catch number information was incomplete, based on Monte Carlo simulation; (2) deriving fishing days for 2007–2009 from voyage data recorder data, based on a newly developed algorithm; and (3) deriving fishing days for 2001–2006 from vessel trip information, based on linear relationships between fishing and at-sea days. These reconstruction data can be used to obtain a standardized series of CPUEs, and the reconstructed

CPUE has also been validated to be more consistent with the assumed PBF population dynamics and other fishery data in the assessment model. Therefore, demonstrating reconstruction data can improve the estimated abundance index.

Keywords: Catch-per-unit-effort, Monte Carlo simulation, Fishing day estimation

Quantitative Ordinal Scale Estimates of Plant Disease Severity: Comparing Treatments Using a Proportional Odds Model

K.S. Chiang*, H.I. Liu, Y.L. Chen
Division of Biometrics, Department of Agronomy, National Chung Hsing University, Taichung, Taiwan
C.H. Bock
USDA-ARS-SEFTNRL, 21 Dunbar Road, Byron, GA 31008, USA

ABSTRACT

Studies in plant pathology and plant breeding requiring disease severity assessment often use a certain type of ordinal scale based on defined numeric ranges, which can be termed a quantitative ordinal scale – with plant disease this special form of the ordinal scale is generally based on the percent area with symptoms [e.g. the Horsfall-Barratt (HB) scale]. We used a parametric proportional odds model to analyze directly the ratings obtained from disease scales, without converting ratings to percentages based on class midpoints of quantitative ordinal scales (currently a standard procedure). This useful feature of the proportional odds model also renders it amenable to comparing estimates from studies using different response scales. The purpose of this study is to evaluate the performance of the proportional odds model for the purpose of comparing treatments (e.g. varieties, fungicides, etc.) based on ordinal estimates of disease severity. A simulation method was implemented to perform the study. The parameters of the simulation were estimated using actual disease severity data from the field. The proportional odds model was compared with the model using midpoint conversions of ordinal intervals. The criterion for comparison was the power of the hypothesis test. Our results show that the performance of the proportional odds model is never inferior to using the midpoint of the severity range at severity <40%. Especially at low disease severity (≤10%), the proportional odds model is superior to the midpoint conversion of the interval method. Thus, for early onset of disease, or for comparing treatments that happen to share severities <40%, the proportional odds model is preferable for analyzing quantitative disease severity estimation data based on ordinal scales when comparing treatments, and at severities >40% is equivalent to other methods.

Keywords: Comparing treatments; quantitative ordinal scales; proportional odds model; midpoint conversions of ordinal interval; simulation.

Optimal two-level choice designs for any number of choice sets

Rakhi Singh, Feng-Shun Chai* (蔡風順)

Indian Institute of Technology Bombay, India,
Institute of Statistical Science, Academia Sinica, Taipei, Taiwan

Ashish Das
Indian Institute of Technology Bombay, India

Abstract

For two-level choice experiments, we obtain a simple form of the information matrix of a choice design for estimating the main effects, and provide *D*- and *MS*-optimal paired choice designs with distinct choice sets under the main effects model for any number of choice sets. It is shown that the optimal designs under the main effects model are also optimal under the broader main effects model. We find that optimal choice designs with a choice set size two often outperform their counterparts with larger choice set sizes.

Key words：Choice design; Choice set; Factorial design; Hadamard matrix; Main effect.

Optimal Two-Level Designs under Model Uncertainty

Pi-Wen Tsai
Department of Mathematics, National Taiwan Normal University

Abstract: Two-level designs are widely used for screening experiments where the goal is to identify the few active factors which have major effects. Most work on two-level designs focuses on level-balanced and/or orthogonal main-effect designs based on the effect hierarchy assumption, estimations for the lower order effects being more important than those of higher order effects. In this talk we study two-level designs based on the model-robust Q_B criterion which aims to improve the estimation in as many models as possible by incorporating experimenters' prior knowledge along with an approximation to the A_s criterion. By relaxing the restrictions on level-balance and pairwise orthogonal, we find a smooth relationship between the choice of designs and the experimenters' prior beliefs on the importance of each effect with the use of the Q_B criterion. Additionally, we extend our study to the case when completely randomization is not feasible and develop a new version of Q_B criterion for block designs. We show that the standard minimum aberration criteria for block designs is a special case of this block-Q_B criterion and the criterion can lead to more appropriate designs reflecting experimenters' prior beliefs.

Detection of Location and Dispersion Effects from Partially Replicated Two-Level Factorial Designs

Shin-Fu Tsai*（蔡欣甫）and Chen-Tuo Liao（廖振鐸）
National Taiwan University

Abstract

Screening active location and dispersion effects is an important issue at the early stage of a quality improvement process. In practice, an active location effect can be used to adjust the system response toward a target value, and an active dispersion effect can be employed to control the system variation. In this talk, we will introduce a new testing procedure for identifying active location effects from partially replicated two-level factorial designs. A two-stage procedure will be introduced for integrating the analyses of location and dispersion effects. Some numerical results will be presented to demonstrate that the proposed method is a promising alternative for real-world applications.

Keywords: Dispersion effect; Factorial Design; Quality improvement; Screening experiment.

Two-Stage Maximum Likelihood Estimation Procedure for Parallel Constant-Stress Accelerated Degradation Tests

Cheng-Hsun Wu[1], Tzong-Ru Tsai[2], Ming-Yung Lee[3*]

[1]Department of Financial Engineering and Actuarial Mathematics, Soochow University, Taipei City, Taiwan.

[2]Department of Statistics, Tamkang University, New Taipei City, Taiwan

[3*] Department of Data Science and Big Data Analytics, Providence University, Taichung City, Taiwan

ABSTRACT

Parallel constant-stress accelerated degradation test (PCSADT) is a popular method used to assess the reliability of highly reliable products in a timely manner. Though the maximum likelihood (ML) method has been commonly utilized to estimate the parameters in the PCSADT, the explicit forms of the ML estimators and their corresponding Fisher information matrix are usually difficult to obtain. In this article, we propose the two-stage ML (TSML) estimation procedure for the time-transformed model, and all of the TSML estimators not only have explicit expressions but also possess consistency and asymptotic normality. Hence, this method is tractable for reliability engineers. Furthermore, the TSML estimators can provide some constructive information about the unknown accelerated relationship law. We also apply our method to analyze the light-emitting diodes data and compare the performance of our estimation with the ML method via simulation.

Keywords: Maximum likelihood estimation, two-stage ML estimation, parallel constant-stress accelerated degradation test, accelerated relationship law, time-transformed model, Wiener process.

Mediation Analyses of Ultraviolet, Air Pollution, and Structural Variations in 559 Human Genomes from the Taiwan Biobank

En-Yu Lai* (賴恩語), Hsin-Chou Yang (楊欣洲), and Yen-Tsung Huang (黃彥棕)
中央研究院統計科學研究所

Wan-Ping Lee (李婉萍)
The Jackson Laboratory for Genomic Medicine

Wen-Chi Pan (潘文驥)
陽明大學環境與職業衛生研究所

Ming-Wei Su (蘇明威) and Chen-Yang Shen (沈志陽)
臺灣人體生物資料庫

摘要

摘要內文

Structural variation is a DNA region that shows changes in copy number, sequence orientation or chromosomal location. Previous studies have suggested a link between air pollution and genetic variation in animal experiments and longitudinal studies, but the sample size is rather limited. It is imperative that a population-based study is conducted to document the potential hazard of environmental exposures such as air pollution and ultraviolet to the human genome and health. The Taiwan Biobank has been collecting biological specimens and conducts the whole-genome sequencing in order to build the reference genome of the Taiwanese population. In this study, we aim to characterize the causal relationship between ultraviolet, air pollution and structural variations. We applied a mediation model to describe the influence of ultraviolet toward structural variants through air pollution. The preliminary results showed a strong effect from ultraviolet to structural variants mediated by air pollution. Validation studies are needed to confirm this interesting finding.

關鍵詞：structural variation, Taiwan Biobank, population genetics

Blood Multiomics Reveal Insights into Individual Resistance against Diabetes, Dyslipidemia and Hypertension

Min-Wei Su,1* Chung-ke Chang,1* Chien-Wei Lin,1 Shiu-Jie Ling,2 Chia-Ni Hsiung,1,3 Hou-Wei Chu,1 Pei-Ei Wu,1 Chen-Yang Shen1,4#

[1]Institute of Biomedical Sciences, Academia Sinica, Taipei, Taiwan
[2] Wego Private Bilingual Senior High School, Taipei, Taiwan
[3]Institute of Bioinformatics and Structural Biology, National Tsing Hua University, Hsinchu, Taiwan
[4]College of Public Health, China Medical University, Taichung, Taiwan

Abstract

Rationale: Diabetes, dyslipidemia and hypertension pose heavy disease burdens in most populations, but certain individuals exhibit resistance against the three diseases. However, few have studied the molecular factors which convey resistance to the three diseases in the population.

Objective: We sought to identify the phenotypic, genomic and metabolomic characteristics of disease-resistant subjects to gain insights into the mechanisms of diabetes, dyslipidemia and hypertension.

Methods and Results: We performed k-means cluster analysis of 16,792 subjects using anthropometric and clinical biochemistry values collected by the Taiwan Biobank. Subjects were assigned to four clusters, with each reflecting different dominant disease pathways. One cluster, Cluster 2, had exceptionally low disease prevalence. Genome-wide association studies found that APOA5 was significantly associated with Cluster 2, and lesser associations included HIF1A, LIMA1, LPL, MLXIPL, and TRPC4. Nuclear magnetic resonance spectra-based metabolome analysis was carried out for 144 Cluster 2 subjects and 73 controls with normal body mass index ($18 < BMI < 24$), good exercise habits (at least 3 times per week, $\geq$30 min per session) and healthy lifestyles (non- smokers and not addicted to alcohol). Blood plasma of Cluster 2 subjects exhibited lowered levels of very low-density lipoprotein and low-density lipoprotein cholesterol, triglycerides, valine, leucine, and acetate compared to controls. The genes and metabolites identified are known to be involved in lipid metabolism and inflammation pathways.

Conclusions: Subjects resistant to diabetes, dyslipidemia and hypertension share a common genetic background, which may translate to better blood plasma lipid profiles and reduced levels of inflammation-inducing metabolites. These genes and metabolites may be good targets for therapeutic development against the three diseases.

Genome-wide association analysis using Taiwan biobank data identified novel loci for fasting glucose

Ren-Hua Chung

Division of Biostatistics and Bioinformatics, Institute of Population Health Sciences,

National Health Research Institutes, Taiwan

Abstract

Elevated glucose level in nondiabetic individuals is a strong predictor for type 2 diabetes (T2D). Genome-wide association studies (GWAS) have identified many genetic loci associated with fasting glucose, using samples with mainly European ancestry. In this study, we performed GWAS analyses based on three cohorts, including Taiwan biobank, the Healthy Aging Longitudinal Study in Taiwan, and the Stanford Asia-Pacific Program for Hypertension and Insulin Resistance. A meta-analysis based on the results from the three cohorts was then performed. Our meta-analysis identified several significant and independent SNPs passing the genome-wide significance threshold of 5×10^{-8}. We also constructed genetic risk scores (GRS) based on the significant SNPs for T2D. The odds ratio of diabetes for the individuals above the 80^{th} percentile relative to those below the 20^{th} percentile of the GRS was 1.33 with a 95% confidence interval of (1.03, 1.70). In conclusion, we identified novel SNPs for fasting glucose using samples with Han Chinese ancestry. The GRS will also be useful for risk assessment of future diabetes.

關鍵詞：Genome-wide association studies, fasting glucose, Taiwan Biobank

Variables Selection and Classification in Linear Mixed-Effects Models

Chih-Hao Chang(張志浩)、Chien-Chung Wang(王建中)*
Institute of Statistics, National University of Kaohsiung

Abstract

We consider linear mixed-effects models for clustering data, where the number of clusters is allowed to go to infinity with the sample size, and the within-cluster sample sizes are balanced. In literature, the statistical inference on linear mixed-effects models such as parameter estimation or model selection are based on an assumption that the explanatory variables are correctly specified in advance for fixed-effects and random-effects models, respectively. In our study, we consider selecting and classifying the explanatory variables for the fixed-effects and the random-effects models. We apply the generalized information criterion (GIC) for variable selection and classification of linear mixed-effects models. We show the consistency of GIC under some mild conditions.

Key words：*Linear Mixed-Effects Models, GIC, Asymptotic Theory*

EWMA管制圖於偏斜常態分佈下的平均連串長度計算效益之研究

沈冠青、蘇南誠
國立臺北大學統計學系

摘要

Lucas 和Saccucci (1990) 利用馬可夫鏈方法算出母體爲常態分配之EWMA 的平均連串長度(ARL，Average run length)，即表示偵測到觀測值超過控制界線所需要的平均次數。他們將EWMA的管制區域切割成數個間隔，透過各個狀態所對應的間隔來給出轉移機率矩陣，再利用起始機率向量給出ARL的計算式。
然而根據論文的描述，我們在常態假設下所得出EWMA管制圖的相關數值與他們的數值卻有一些出入，因此我們將透過模擬和科學計算來研究是否在不同的起始狀態與切割間隔數會影響ARL的大小。我們也將進一步在偏斜常態的母體下,研究起始狀態和切割數對EWMA管制圖的ARL計算情況,並與常態母體的情況比較，以期探討計算ARL方法的精確性和穩健性。

關鍵詞：
馬可夫鏈方法、EWMA管制圖、平均連串長度、偏斜常態分配

相關性連續資料之共變數相關 ROC 曲線的估計與檢定

顏振庭*、黃怡婷
國立臺北大學統計學系

摘要

醫學領域會需要對個體作病情診斷的檢驗（diagnostic test），診斷方式一般會檢驗個體的生物指標數，來判別個體是否患病，而操作者接收曲線（ROC curve）與其線下面積（area under the curve, AUC）是最常用來評估診斷檢驗準確度的方法。檢驗方式的準確性有可能會與一些因素相關，此影響可能會導致操作者接收曲線變動，進而影響檢驗的準確度，Tosteson 與 Begg (1988) 提出與操作者接收曲線相關的順序迴歸模型（ordinal regression model），加入可能影響生物指標數準確性的因素。

重複測量資料是另一種醫學領域常見的資料收集型態，Tolendano 與 Gatsonis (1996) 將 Tosteson 與 Begg (1988) 所提出的模型架構拓展至相關性資料，使診斷檢驗準確度的評估能被應用在重複測量或是重複評估的資料，此模型依然使用順序迴歸模型，並採用廣義估計方程式（generalized estimating equation, GEE）方法（Liang 與 Zeger, 1986）估計參數。參考 Tolendano 與 Gatsonis (1996) 及 Tosteson 與 Begg (1988) 所提出的模型，本研究考慮生物指標數為重複測量的連續型反應變數，並對其建構迴歸模型，使用限制的最大概似函數（restricted maximum likelihood function, REML）來估計參數，並提出檢定來評估影響檢驗準確度因子。

關鍵詞： ROC 曲線、有限制最大概似估計式、重複測量資料

Economic design of two-stage control chart with skew-normal components

Fu-Hsin Hsu (許馥欣)*、Nan-Cheng Su (蘇南誠)
Department of Statistics, National Taipei University

Abstract

In many instances of quality control, the cost is too high to monitor the performance variable, but it could be economical to monitor its surrogate. In this study, we consider a two-stage control chart by using both the performance variable and its high-correlated surrogate variable in an alternative fashion. The components of the performance varialbe and the surrogate variable are assumed to follow the Azzalini's skew-normal distribution with different skewness parameter. This in turn implies that the distribution of the performance variable and the surrogate variable is an extention to the Azzalini's skew-normal distribution. We will study properties of the new class of the extented skew-normal distribution. Furthermore, we will investigate properties of the two-stage control chart under this extended model by the expected net income per unit time.

Keywords:
Economic design; Two-stage control charts; Markov chain approach; Skew-Normal distribution; Surrogate variable

探討偏斜常態分配運用於監控變異係數之效用

鄭雨函*、蘇南誠
國立臺北大學統計學系

摘要

本篇論文主要利用適應性休哈特控制圖中變動樣本數 (VSS, variable sample size) 管制圖來監控變異係數。由於推導變異係數的精確分配是困難的，所以傳統上都利用非中心 t 分配去逼近樣本變異係數的累積分佈函數。然而此方法在計算控制界限較為複雜，因此近年來有學者試圖用對數常態分配來給出控制界限。根據 Castagliola et.al (2015) 在常態母體的假設下討論的樣本變異係數之監控流程設計，我們將取代非中心 t 分配而利用偏斜常態分配去逼近，進而從觀察在不同情況下的平均連串長度評估指標等數值表現。我們發現其與文獻中的表現相似，因此若取代傳統作法，將可讓監控變異係數的流程更為迅速和便利。另外，在真實情況下，大多數資料不具常態分配，故我們將進一步延伸探討母體為偏斜常態分配時，我們所提供的監控變異係數流程之表現。

關鍵詞：偏斜常態分配、變異係數、VSS 管制圖。

Fractural Vertebral Body Discrimination by Deep Learning Classification and Segmentation

Po-Hsin Chou [#]
Department of Orthopedics and Traumatology,
Taipei Veterans General Hospital, Taipei, Taiwan

Yi-Chu Li [#]
Institute of Data Science and Engineering,
National Chiao Tung University, Hsinchu, Taiwan

Hung-Hsun Chen
National Center for High-performance Computing, Hsinchu, Taiwan

Ming-Chau Chang
Department of Orthopedics and Traumatology,
Taipei Veterans General Hospital, Taipei, Taiwan

Hung-Ta Hondar Wu
Department of Radiology,
Taipei Veterans General Hospital, Taipei, Taiwan

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

ABSTRACT

Orthopedists have to spend much time on measuring the vertebral fracture localization for grading in X-ray image. Therefore, this study integrates the techniques of medical image processing and deep learning to assist orthopedics to discriminate vertebral fracture localization and grade with accuracy.

In this study, we use the method of YOLO (You Only Look Once) to detect the location of vertebral body. Then, we use transfer learning to train a classification model to decide whether it is fracture or not. The prediction accuracy, precision and recall are 87.5%, 87.5% and 87.76%, respectively. In order to further detect fracture localization and compute its grade at the same time, we utilize the technique of U-Net for image segmentation to calculate the reduced height ratio of fractural vertebral body. The IoU (Intersection over Union) of segmentation is 73%.

Keywords: Deep Learning, Transfer Learning, YOLO (You Only Look Once), U-Net

Presenter: Yi-Chu Li

Correspondence to: Henry Horng-Shing Lu

# contributed equally

A Novel Study Design for Three-Arm Equivalence Clinical Trial with Binomial Distributed Outcomes

Yi-Kuan Tseng and Chia-Yun Wang

Graduate Institute of Statistics, National Central University, Jhongli, Taiwan

Ken-Ning Hsu

WuXi AppTec, Shanghai, China

Abstract

The assessment of equivalence in a clinical trial may be conducted through a three-arm trial (test drug, reference drug, and placebo). The three-arm equivalence trial consists of three hypothesis tests in practice, where two hypothesis tests demonstrate the superiority of test drug and reference drug against placebo, and the other one demonstrates the equivalence of test drug and reference drug. When designing a three-arm equivalence clinical trial, the practitioner should minimize the chance of that test drug is found to be equivalent to the reference drug but to be non-superior to placebo. To minimize the chance at the design stage for the three-arm equivalence trial using binary outcome as the primary endpoint, one way is to test the two superiorities and the equivalence simultaneously through a single set of null and alternative hypotheses based upon the ratio of difference of the proportions. In this research, we derived the test statistic and the power function for the single set of hypotheses. The required sample size for achieving the desired power at the given significant level can be obtained by solving the power function. In this article, I will illustrate the proposed design through an example, and I will demonstrate the required sample sizes for various conditions.

Keywords: Binary outcome, Equivalence test, Sample size, Superiority test, Three-arm clinical trial.

A gorup sequnetial Holm procedure for multiple comparisons in confirmatory clinical trials

Yi-Kuan Tseng (曾議寬)
Graduate Institute of Statistics, National Central University, Jhongli, Taiwan

*Ken-Ning Hsu (許根寧)

WuXi AppTec, Shanghai, China

Abstract

Biomarkers of prognosis in oncology studies have become increasingly important in recent years. They are not only used to predict the time to event, but also used in the design of phase III trial. Investigators may perform a systematic search using the log-rank test for all possible cut-off points for a continuous biomarker based on the survival data of phase II trial, and they choose the cut-off point associated with the minimum *p*-value of log-rank test. The patient population can be divided into two groups by the chosen cut-off point, and the target population of phase III trial would be the group with longer survival. However, some statistical issues for the minimum *p*-value method have been pointed out by several authors. To address the issues, we would like to introduce a likelihood approach under the extended hazard model which includes the accelerated failure time model and the proportional hazards model as the special cases. A systematic search of the Wald statistics for all possible cut-off points would be performed to choose the cut-off point. We will illustrate the proposed approach by a numerical simulation study.

Key Words：accelerated failure time model, biomarker, dichotomization, extended hazard model, proportional hazards mode, survival analysis

空間交互作用模型參數的貝氏估計量之統計評估

Statistical Evaluation for Bayes' Estimator of Parameters in Spatial Interaction Model

吳建霖、馬瀰嘉

成功大學統計學研究所

摘要

全民健康保險自開辦以來，雖使全體國民在醫療照顧方面上有了明顯的提升，但對於台灣空間醫療資源分配不均的問題，一直是須改善的重點問題之一。以往在探討台灣各區域醫療資源上的分配問題時，往往只看該區域的病床數及醫療人員數目，若將此兩項變數作為衡量該區域醫療資源是否充足的指標，其背後的假設為醫療機構與人口是平均分配在該區域上，但此假設與實際情況卻是大相逕庭。若要更公正客觀的去評估該區域醫療資源是否有分配不均的問題，應將就醫移動距離與所需花費時間考慮進去。

本研究利用 Huff 模型(又被稱為空間交互模型)，加入就醫移動距離與所需花費時間來探討就醫可近性對於就醫流動人次的影響。在統計方法上，本研究期望透過貝氏分析的方法，能適度的估計出模型中的參數。並利用一實例來說明如何應用。最後，利用統計模擬方法來比較貝氏分析混和 Huff 模型和隨機效應混和卜瓦松迴歸模型的平均絕對相對誤差值，以評估不同模型方法擬合就醫流動人次的優劣。

關鍵詞：醫療資源分配、貝氏分析、Huff 模型、空間交互模型

Computing entropy rates of partially observed Markov chains and Lyapunov exponents for products of random Markov matrices by the Fredholm integral equations

Chen, Chun-Ying
Department of Finance, National Taiwan University

Abstract

In this paper we investigate partially observed Markov chains (POMCs) that follow partially observed Markov models (POMMs), including the famous hidden Markov models (HMMs) as special cases. We develop transition calculus to derive recursive Bayesian filters for POMMs. Under mild ergodic conditions, the recursive filter equation asymptotically turns into a random eigen-equation of a random Markov matrix. There exists a unique eigen-distribution by the random Perron theorem, and it can be solved numerically by the Fredholm integral equation. The entropy rate of a POMC is equal to the maximal Lyapunov exponent of the random Markov matrix. Numerical and Monte Carlo results for two- and three-state POMM are provided.

Keywords：
POMC, POMM, transition calculus, recursive Bayesian filter, random eigen-equation, random Markov matrix, eigen-distribution, random Perron theorem, *n*-dimensional Fredholm integral equation, entropy rate, Lyapunov exponent

Multiple Acceleration on Reversible Markov Chain

Chen-Wei Hua*
Department of Mathematics, National Taiwan University

Ting-Li Chen
Institute of Statistical Science, Academia Sinica

Abstract

Chen and Hwang (2013) proposed to improve a reversible Markov chain by adding an antisymmetric perturbation on a cycle. Since the perturbed Markov chain is no longer reversible, one can not iteratively apply this antisymmetric perturbation method on different cycles. Chen and Hwang (2013) also showed that the method works on disjoint cycles. In this talk, we further investigate the case of two cycles sharing the same vertex. We will show that the method can work on two cycles under some additional conditions. In addition to the theory, we implement the antisymmetric perturbation method on the Ising model.

keywords：Markov chain Monte Carlo, rate of convergence, reversibility, asymptotic variance, antisymmetric perturbation

A new model for dependent competing risks data in reliability

Yin Chen, Wang
Graduate Institute of Statistics, National Central University

Abstract

We consider a new model based on Copula and frailty for dependent competing risks data. The purpose of our model is making the model more widely applicable. Models are no longer limited to independent situations. A commonly used copula is in Archimedean copulas, for example (Clayton copula, Gumbel copula, …). In this paper the frailty term follow the one parameter Gamma distribution. The Frailty model also has a good description of the dependent of observations. We derive data generation methods and Kendall's tau under the new model.

Keywords *: Joint model ,Copula ,Frailty Survival Analysis*