Multiple imputation confidence intervals for the mean of the discrete distributions for incomplete data

Chung-Han Lee(李宗翰)*, Hsiuying Wang(王秀瑛)
Institute of Statistics
National Yang Ming Chiao Tung University

Abstract

Confidence intervals for the mean of discrete exponential families are widely used in many applications. Since missing data are commonly encountered, the interval estimation for incomplete data is an important problem. The performances of the existing multiple imputation confidence intervals are unsatisfactory. We propose modified multiple imputation confidence intervals to improve the existing confidence intervals for the mean of the discrete exponential families with quadratic variance functions. The simulation study shows that the coverage probabilities of the modified confidence intervals are closer to the nominal level compared with the existing confidence intervals. These confidence intervals are also illustrated with a real data example.

Keywords: Binomial distribution, Coverage probability, Exponential family, Missing value, Poisson distribution, Wilson interval.

On identification and estimation for sufficient cause interaction through quasi-instrumental variable

夏珮瑄*、戴安順
國立陽明交通大學

林聖軒
國立陽明交通大學

Abstract

Sufficient cause interaction (SCI) has received much attention to investigating the mechanism of causality. Under the counterfactual framework, VanderWeele and Robins (2007, 2008) provided empirical tests for SCIs. However, the previous studies only assess the lower bound of SCIs rather than estimate SCIs directly due to the lack of the degree of freedom. Moreover, such empirical tests for the lower bound of SCIs are less powerful. To address this issue, in this study, we propose a novel method to estimate the probability of individual with SCI by introducing a new factor named quasi-instrumental variable, which is necessary for the background condition of SCI. We also develop a corresponding hypothesis test and show that it is more powerful than the empirical test.

Keywords: Sufficient cause interaction, quasi-instrumental variable.

利用物聯網傳感器暨動態影像評估蚯蚓成長率

花郁婷、王豪善、游竣宇*、謝文權
義守大學生物科技學系

黃詠暉、陳泰賓
義守大學醫學影像暨放射科學系

摘要

目的：由於蚯蚓具有改良土壤之作用，但是大多數的蚯蚓對農藥殘留非常敏感；況且評估其成長率卻很困難；因此透過實驗養殖並透過物聯網傳感器(Internet of Thing Sensor, IoT Sensor)暨動態影像評估蚯蚓成長率為本實驗之目的。
材料與方法：本實驗以印度藍品系為主要實驗對象；以 10 隻為一組共計三組。IoT Sensor 為 Raspberry Pi 4 搭配溫、濕、氣體感知器，同時透過 CCD 取像量測其體長各二次。每週測量乙次，記錄其體重(g)及體長(cm)；成長率分析以迴歸曲線估計體重與體長之關連，估算其 4 週成長變化率，顯著水準為 0.05。
結果：根據 4 次量測體重與體長之迴歸方程式顯示，每增加一公克體重，其體長每週平均成長為 6.4 至 9.8 公分($0.23 \leqq R^2 \leqq 0.34$, $P<0.05$)。
結論：透過物聯網傳感器暨動態影像評估蚯蚓成長率具有可行性及未來性，未來仍需精進測量技術之提升，使蚯蚓養殖未來利用性能大幅提升。

關鍵詞：蚯蚓成長率、物聯網傳感器、迴歸方程式

Phylogenetic curved optimal regression for adaptive trait evolution

鍾冬川*、王智平
逢甲大學統計學系

Abstract

Regression analysis using line equations has been broadly applied in studying the evolutionary relationship between the response trait and its covariates. However, the characteristics among closely related species in nature present abundant diversities where the nonlinear relationship between traits have been frequently observed. By treating the evolution of quantitative traits along a phylogenetic tree as a set of continuous stochastic variables, statistical models for describing the dynamics of the optimum of the response trait and its covariates are built herein. Analytical representations for the response trait variables, as well as their optima among a group of related species, are derived. Due to the models' lack of tractable likelihood, a procedure that implements the Approximate Bayesian Computation (ABC) technique is applied for statistical inference. Simulation results show that the new models perform well where the posterior means of the parameters are close to the true parameters. Empirical analysis supports the new models when analyzing the trait relationship among kangaroo species.

Keywords: adaptive trait evolution, approximate Bayesian computation, geometric Brownian motion, geometric Ornstein-Uhlenbeck process, phylogenetic comparative analysis.

Bayesian modelling integer-valued transfer function models

A. C. Pingal(高裘雷)*, Cathy W.S. Chen(陳婉淑)
Feng Chia University (逢甲大學統計學系)

Abstract

External events commonly known as interventions often affect times series of counts. This research introduces a class of transfer function models that include four different types of interventions on integer-valued time series: abrupt start and abrupt decay (additive outlier), abrupt start and gradual decay (transient shift), abrupt start and permanent effect (level shift), and gradual start and permanent effect. We propose integer-valued transfer function models incorporating a generalized Poisson, log-linear generalized Poisson or negative binomial to estimate and detect these four types of interventions in a time series of counts. Utilizing Bayesian methods, which are adaptive Markov chain Monte Carlo (MCMC) methods to obtain the estimation and prediction, we further employ deviance information criterion (DIC) and mean squared standardized residual for model comparisons. As an illustration, this study evaluates the effectiveness of our methods through a simulation study and application to crime data in Albury City, New South Wales Australia. Simulation results show that the MCMC procedure is reasonably effective. The empirical outcome also reveals that the proposed models are able to successfully detect the location and type of interventions.

Keywords: Intervention analysis, generalized Poisson, integer-valued GARCH model, Markov chain Monte Carlo method, transfer function.

Detection of change points for wind velocity

楊子賢*
國立陽明交通大學應用數學

蔡碧紋
國立臺灣師範大學數學

郭志禹
中央研究院應用科學研究中心

張文溢
國家實驗研究院國家高速網路計算中心

Abstract

Wind power is a valuable source of sustainable energy, hence understanding the regional wind properties and the changes of wind speed and wind direction at a specific wind farm location is important in wind energy applications. In this study, we develop an automatic procedure to identify such changes from the wind speed and wind direction data. Firstly, we extend the Pruned Exact Linear Time (PELT) method proposed by Killick et. al (2012) to a Weibull distribution. A simulation study is done which shows that the Weibull assumption improves the sensitivity in detecting changes in the characteristics of wind speed data when the Normal assumption is inadequate. Additionally, we consider the simultaneous multiple changes for wind speed and wind direction. We extend PELT to two-dimensional cases by considering the joint distribution of Weibull distribution and Von Mises distribution. An application to the hourly wind speed and wind direction data observed from a wind turbine at Changhua Fuhai, Taiwan, is given. The results found by our method coincide with the changes of wind patterns in different seasons in Taiwan.

This research is supported in parts under MOST, 109-2218-E-001 -003.

Keywords: change points, time series, wind speed, wind direction, Normal distribution, Weibull distribution, Von Mises distribution, PELT.

Robust inference for AR-G/GARCH model with Mallows-type model averaging estimator

Hsin-Chieh Wong(翁新傑)
Department of Mathematics
National Central University

Abstract

In this paper, we consider AR-G/GARCH models and provides the limiting distribution of the Mallows-type model averaging estimators in a local asymptotic framework. This local-to-zero asymptotic framework for linear regression is used to ensure the finite asymptotic mean squared error in model uncertainty structures. Upon different levels of noises, we need scaling of the data length T to have limiting distribution of estimator. In order to gauge the level of a noise ϵ, we use the well-known tail index λ, that is, the tail probability of the noise satisfies $P\{\epsilon > x\} \approx x^{-\lambda}$. Zhang and Ling (2015) considered four different ranges of λ and found that the tail index λ affects the scaling in the asymptotic framework. Under these four cases, we show that the averaging estimators with fixed weights are asymptotic stable-family distribution in a local asymptotic framework and then we propose a model averaging estimator based on the Mallows model averaging of Hansen (2007). We also present a robust test for our averaging estimator, this is done by dividing our sample into a fixed number of groups together with the asymptotic distribution of groups and Student's t-statistics. Through simulation experiment, our method delivers great numerical performance for the finite-sample properties of the test. This is joint work with S. Y. Shiu.

Keywords: AR-G/GARCH, model averaging, heavy tails, tail behavior, stable distribution, Mallows-type model averaging.

Large-scale multi-mediator analyses under a composite null

En-Yu Lai(賴恩語)*, Yen-Tsung Huang(黃彥棕)

中央研究院統計科學研究所

Abstract

Mediation analysis aims to evaluate the effect of a hypothetical causal mechanism that is from an exposure, through mediators, to an outcome. Therefore, the effect of the exposure on the mediator and the effect of the mediator on the outcome conditional the exposure will jointly construct the mediation effect. Conventional test statistics of the mediation effect become conservative when signals are sparse. The power loss originates from 1) a poor approximation of the normal product distribution using the normal distribution and 2) the complications of a composite null hypothesis. Huang (2019) has proposed a novel test for single-mediator analyses to accurately assess the normal product distribution under a composite null hypothesis accounting for the composition of null hypotheses within a study. Here we extend the method to accommodate the setting with multiple mediators. We utilize Huang (2019)'s method to account for the composite null hypothesis and then exploit global testing procedures proposed by Sun and Lin (2019) to conduct multivariate tests. We conduct extensive simulation studies to evaluate the performance of the proposed method. In addition, we apply the method to the TCGA-LUAD dataset in order to select genes whose expression may be regulated by smoking-induced DNA methylation. We implement the method into an R package `MACtest`, which is available at `https://github.com/roqe/MACtest`.

Keywords: mediation analysis, composite null hypothesis, test statistics, multiple mediator models, mechanism evaluation.

Spatio-temporal analysis of extreme $PM_{2.5}$ in Taiwan

周秉儒、徐南蓉
國立清華大學統計所

黃信誠*
中央研究院統計科學研究所

Abstract

Exposure to high levels of fine particulate matter (PM2.5) can cause adverse health effects to humans. In Taiwan, the PM2.5 concentration has significant regional differences. Its value tends to be higher in the south and lower in the north and has apparent seasonality. To understand the temporal and spatial variability of extreme PM2.5 concentrations, we develop a simple yet flexible space-time model for the daily maximum PM2.5 employing the extreme value theory. The proposed model accounts for spatial and seasonal structures, and its parameters are estimated by maximum likelihood. We use the Environmental Protection Administration monitoring data from 2006-2018 to illustrate the proposed method.

Keywords: Bayesian information criterion, Generalized extreme value distribution; maximum likelihood; multi-resolution spline basis functions; seasonality.

Matrix autoregressive spatio-temporal models

Nan-Jung Hsu(徐南蓉)*
Institute of Statistics
National Tsing-Hua University

Hsin-Cheng Huang
Institute of Statistical Science
Academia Sinica

Ruey S. Tsay
Booth School of Business
University of Chicago

Abstract

Matrix-variate time series are now common in economic, medical, environmental, and atmospheric sciences, typically associated with large matrix dimensions. We introduce a structured autoregressive (AR) model to characterize temporal dynamics in a matrix-variate time series by formulating the AR matrices as a bilinear form. This bilinear parameter structure reduces the model dimension and highlights dynamic interaction among columns and rows in the AR matrices, making the model highly explainable. We further incorporate spatial information and explore sparsity in the AR coefficients by introducing spatial neighborhoods. In addition, we consider a non-stationary multi-resolution spatial covariance model for innovation errors. The resulting spatio-temporal AR model is flexible in capturing heterogeneous spatial and temporal features while maintaining a parsimonious parametrization. The model parameters are estimated by maximum likelihood (ML) with a fast algorithm developed for computation. We conduct a simulation study and present an application to a wind-speed dataset to demonstrate the merits of our methodology.

Keywords: Bilinear autoregression, Dimension reduction, Matrix-variate time series, Maximum likelihood, Multi-resolution spline basis functions.

Model selection with an anisotropic nested spatial correlation structure

Chun-Shu Chen(陳春樹)
Graduate Institute of Statistics
National Central University

Abstract

In spatial regression analysis, a suitable specification of the mean regression model is crucial for unbiased analysis. Suitably account for the underlying spatial correlation structure of the response variables is also an important issue. Here, we focus on selection of an appropriate mean model in spatial regression analysis under a general anisotropic nested spatial correlation structure. We propose a distribution-free model selection criterion which is an estimate of the weighted mean squared error based on assumptions only for the first two moments of the response data. The simulations under the settings of covariate selection reveal that the proposed criterion performs well for covariate selection in the mean model regardless of the underlying spatial correlation structure is nested/non-nested, isotropic/anisotropic. Also, the proposed criterion accommodates both continuous and count response data. Finally, a real data example regarding the fine particulate matter concentration is also analyzed for illustration. (This is a joint work with Chung-Wei Shen and Yi-Hau Chen)

Keywords: Anisotropy, Generalized method of moments, Information criterion, Variable selection.

An intraday range model with leverage effect in intraday volatility pattern

PING CHEN TSAI(蔡秉真)*
National Sun Yat-sen University

CHEOLJUN EOM
Pusan National University

CHOU WEN WANG
National Sun Yat-sen University

Abstract

An intraday range model is specified by extending and modifying the Conditional Autoregressive Range (CARR) model of Chou (2005) and this fills in a void in the literature. Both range and squared range are considered in the modelling, and the most general framework includes features such as an endogenized intraday volatility patterns, a leverage effect due to past intraday return, a leverage effect due to past overnight return and a plausible leverage effect in intraday volatility pattern. Empirical estimation is made on 2,839 days of hourly, 30- and 15-minute KOSPI index data, with an application of intraday bid-ask spread estimation for order data.

Keywords: Range, Intraday volatility pattern, Leverage effect, Overnight return, Effective spread.

公司治理如何地影響公司績效：台灣產險公司的分量迴歸分析實證

古裕彥
實踐大學風險管理與保險學
國立台灣大學會計學系

李承謙*
實踐大學風險管理與保險學系

摘要

這篇研究從實證上提供台灣產險產業其公司治理與公司績效之間的非線性和異質性的關係證據。對於 2008 年至 2019 年的 126 家台灣產險公司的數據集，我們使用分量迴歸分析（QR）方法來探討。得到的結果表明：(1）董事會會議次數/公司成立的年數都和公司績效（ROE）有穩健負向顯著關係（2）公司成立的年數與公司績效（ROA）之間存在穩健負向顯著關係（3）審計委員會會議次數/公司規模都和公司績效（EPS）有穩健正向顯著關係，然而審計委員會中獨立董事人數對於保險公司的績效則具有負向影響。此外，QR 方法的使用可能比估計應變數的平均效果有更充分的訊息(OLS)。

關鍵詞：分量迴歸、公司治理、公司績效

CEO 薪酬與公司績效：台灣財產保險公司的分量迴歸分析實證

古裕彥*
實踐大學風險管理與保險學系
國立台灣大學會計學系

江庭
實踐大學風險管理與保險學系

摘要

這項研究使用台灣財產保險公司 2012 年至 2020 年的數據並藉由分量迴歸模型（QR）來探討 CEO 薪酬對公司績效非單一性效果的影響。實證結果表明，CEO 薪酬對於公司績效是有正面的影響，不論公司績效處在低分量、中分量、或是高分量的情況下。此外，負債比率對公司績效有顯著的負面影響；然而，公司規模對公司績效沒有顯著影響。更進一步，在控制年度影響之後，CEO 薪酬對公司績效的影響效果之分量變化模式亦很穩健。對於在公司績效和 CEO 薪酬之間使用替代的應變數或解釋變數其實證關係結果亦很穩健。

關鍵詞：CEO 報酬、公司績效、分量迴歸

台灣財產保險產業其再保險需求之決定性因素：分量迴歸分析的實證

古裕彥
實踐大學風險管理與保險學系，
國立台灣大學會計學系

林姿妤*
實踐大學風險管理與保險學系

摘要

不同於之前使用普通最小平方法（OLS）的文獻，本研究採用分量迴歸分析方法（QR）來探討 2008 年至 2020 年台灣財產保險產業其再保險需求的決定性因素。QR 方法提供更多訊息有關保險公司其再保險需求是如何被影響，尤其是對於在較低和較高分量之下的再保險需求公司。我們發現公司規模對於在較低和較高再保險需求的保險公司產生相反的影響效果。實證結果也發現，大部份公司特性的變數會影響保險公司的再保險需求，而這與之前使用 OLS 方法的研究所發現是一致的。

關鍵詞：再保險需求、分量迴歸、公司特性

Unification of semicompeting risks analysis through causal mediation modeling

Jih-Chang Yu(余日彰)*, Yen-Tsung Huang
Institute of Statistical Science
Academia Sinica

Abstract

Semicompeting risks represent a common problem where an intermediate event and a terminal event are both of interest, but only the former may be truncated by the latter. Copula, frailty and multistate models serve as well-established analytics for semicompeting risks. Here, we cast the semicompeting risks in a causal mediation framework, with the intermediate event as a mediator and the terminal event as an outcome. We define the indirect and direct effects as the effects of an exposure on the terminal event mediated and not mediated through the intermediate event, respectively. We derive respective expressions of estimands for causal mediation under the copula, frailty and multistate models. Next, we propose estimators based on nonparametric maximum likelihood or U-statistics and establish their asymptotic results. Numerical studies demonstrate that the efficiency of copula models leads to potential bias due to model misspecification. Moreover, the robustness of frailty models is accompanied by a loss in efficiency, and multistate models balance the efficiency and robustness. We observe a similar feature when applying the proposed methods to a hepatitis study, indicating that hepatitis B affects mortality by increasing liver cancer incidence. Thus, causal mediation modeling provides a unified framework that accommod ates various semicompeting risks models.

Keywords: causal mediation model, copula model, frailty model, multistate model, semicompeting risks.

Transcriptome-based score predicts hepatitis B virus DNA decrease by Tenofovir

Jia-Ying Su (蘇家瑩) *
Bioinformatics Program, TIGP
Academia Sinica
Institute of Biomedical Informatics
National Yang Ming Chiao Tung University

Yao-Chun Hsu (許耀峻)
Centre for Liver Diseases
E-Da Hospital

Yen-Tsung Huang (黃彥棕)
Institute of Statistical Science
Academia Sinica

Abstract

More than 240 million people worldwide are chronically infected by hepatitis B virus (HBV). Tenofovir (TDF) is recommended as a first-line nucleotide analogue drug for HBV treatment. In this study, we aimed to identify a panel of transcriptome biomarkers to select patients who would benefit more by treating with TDF. We analyzed data from a multicenter, double-blind, placebo-controlled, randomized trial that enrolled patients with chronic hepatitis B in Taiwan. We built multiple linear regressions to select genes with expression levels modifying the effect of TDF on a decrease of HBV DNA. One-year HBV DNA decreased percentage with logit transformation was the dependent variable and log2 transformed baseline gene expression levels (before taking TDF), natural log transformed baseline HBV DNA levels, and their cross-product interaction term were the independent variables. To investigate the potential effect modification by the baseline gene expression, we conduct a transcriptome-wide screening testing the significance of the interaction terms, followed by false discovery rate correction (q-value). We found that *DCPS*, *TOX* and *MATN2* significantly modify the treatment effect of TDF on reducing HBV DNA levels in patients. We further constructed a prediction scores composed of significant genes (q-value < 0.2) weighted by their regression coefficients. Patients

with lower scores benefit more from TDF. In summary, our finding suggests that the baseline expression levels of these three genes may play a critical role in TDF therapy and may be able to provide clinically useful information for HBV patients.

Keywords: Personalized medicine, Effect modification, Transcriptome-wide analysis, Viral hepatitis, HBV.

Semiparametric analysis of covariate-dependent cross ratio for ordered bivariate gap times

連振宇*
國立台灣大學統計碩士學位學程

張淑惠
國立台灣大學流行病學與預防醫學研究所

Abstract

In longitudinal follow-up studies, individual may experience ordered bivariate events before the end of the study. For instance, discharge and rehospitalization are ordered bivariate events for patients having a chronic disease. Association analysis between the first and second gap times between ordered bivariate events in terms of time-invariant and time-varying cross ratios may provide predictive information on the course of chronic disease. Such association may be affected by patient's characteristics, such as gender, genome type and other related factors. Therefore, we introduce parametric log-linear models for cross ratio between ordered bivariate gap times. These parametric cross ratio models can be generated by stratified proportional hazards models via conditional hazards for the second gap times given the first gap time and covariates without specifying the marginal distributions of the first and second gap times. Then, semiparametric estimates of the covariate-dependent cross ratios can be obtained by maximizing the weighted partial likelihood in which the inverse probability of censoring weights is used to tackle the induced dependent censoring of the second gap time. We have shown the consistency and asymptotic normality of the estimated covariate-dependent cross ratios. The finite-sample performance of the proposed methods is examined by simulation studies. The proposed method is also applied to the colon data for the estimation of the covariate-dependent cross ratios for the first gap time from entry study to recurrent colon cancer and the second gap time from recurrent colon cancer to death.

Keywords: Clayton dependence structure, Cross ratio, Gap times, Induced informative censoring, Inverse probability censoring weighted, Pseudo weighted partial likelihood.

Natural history of chronic hepatitis B and C: A causal mediation study

Yi-Ting Huang (黃意婷)*, Yen-Tsung Huang (黃彥棕)
Institute of Statistical Science
Academia Sinica
Institute of Epidemiology and Preventive Medicine, College of Public Health,
National Taiwan University

Abstract

Introduction

Hepatitis C virus (HCV) infection is construed as a systemic disease with extrahepatic manifestations, whereas Hepatitis B virus (HBV) infection is not. Although liver cancer is a common cause of death in carriers of HBV and HCV, they may not share the same natural history of disease. In this study, we aim to investigate the natural course of HBV or HCV infection using causal mediation modeling.

Method

A community-based cohort study (Risk Evaluation of Viral Load Elevation and Associated Liver, REVEAL) was conducted in Taiwan, in which 23,820 individuals recruited during 1991 and 1992. Researchers collected participants' serum samples to profile chronic HBV and HCV infections respectively by hepatitis B surface antigen and antibodies against hepatitis C virus. This cohort data was linked to the National Health Insurance Research Database to determine incidences of potential mediating diseases and mortality. We implemented a non-parametric causal mediation estimator to estimate the causal mediation effect of hepatitis on mortality in relation to the potential mediating disease incidences.

Result

Overall, we found that the effect of HBV infection on mortality was mostly through liver cancer incidence while the mortality of HCV carriers can be mediated through liver cancer or other diseases. Particularly, type II diabetes mellitus, blood/immune diseases, diseases of gallbladder, cerebrovascular disease, and other diseases of intestines and peritoneum also mediated the effect of HCV on mortality with age less than 55 years. In summary, our mediation analyses demonstrate that different natural

histories of HBV and HCV infections and identify separate sets of mortality-mediating diseases for HBV and HCV.

Keywords: causal inference, mediation model, hepatitis C, extrahepatic manifestation.

An m-spread model using branching procsses

Jyy-I (Joy) Hong(洪芷漪)
Department of Mathematical Sciences
National Chengchi University

Abstract

When an infectious disease spreads among a population and it may cause different reactions in individuals. To understand and study the spread patterns is always one of the key things in decision making during an epidemic. In this talk, we will introduce a model to describe a spread pattern which depends on the spreading history within certain time period and construct an induced branching process to study the long term behavior of the spread pattern and the spread rate. This is a joint work with Jung-Chao Ban and Yu-Liang Wu.

Recurrence vs transience in RWCRE

Yuki Chino(千野由喜)
Department of Applied Mathematics
National Yang Ming Chiao Tung University

Abstract

One-dimensional Random Walk in Cooling Random Environment (RWCRE) is obtained as a patchwork of one-dimensional Random Walk in Random Environment (RWRE) by resampling the environment along a sequence of deterministic times. The RWCRE model can be seen as a model that interpolates between the classical static model and the model with i.i.d. resamplings every unit of time. This model shows a crossover between RWRE and a homogeneous model according to how to resample, called the cooling map. This talk is based on a joint work with L. Avena (Leiden University), C. da Costa (Durham University) and F. den Hollander (Leiden University).

Heat kernel bounds for nonlocal operators with singular kernels

Moritz Kassmann
Bielefeld University

Kyung-Youn Kim(金璟允)*
National Chengchi University

Takashi Kumagai
Kyoto University

Abstract

We consider *d*-dimensional Markov processes which are written as *d* independent copies of 1-dimensional jump processes so that the jumping measures are singular with respect to the *d*-dimensional Lebesgue measures. We obtain the sharp two-sided bounds of the fundamental solution for the integro-differential operators corresponding to the Markov processes.

Keywords: Markov jump process, heat kernel, integro-differential operator.

Automatic extraction of electronic-component three-view drawings based on CNN and LLDA

Tzung-Pei Hong(洪宗貝)*, Chih-Sheng Hsu(許智勝)
Department of Computer Science and Information Engineering
National University of Kaohsiung, Kaohsiung

Yan-Zhih Wang(王彥智), Yi-Ting Chen(陳怡婷)
Footprintku Inc.

Shih-Feng Huang(黃士峰), Hsiu-Wei Chiu(邱修偉)
Department of Applied Mathematics
National University of Kaohsiung

Abstract

Electronic component specifications, which are usually presented as a PDF document, contain important information about electronic-component design. Traditionally, they need to be checked manually to extract three-view drawings of electronic component specifications from the PDF documents, so it is costly and time-consuming. To address this problem, in this paper, we propose a framework to automatically search for three-view drawings from electronic component specifications. First, the proposed approach parses the PDF documents to get object layouts and then utilizes them to capture the pages containing drawings. The remaining pages without drawings are removed to reduce the number of pages. Next, we use the convolutional neural network and the LLDA analysis to determine whether a page contains three-view drawings. The framework can help us obtain three-view drawings. The experimental results show that our accuracy rate is above 90%, which means that the proposed framework can automatically extract the required information well.

Keywords: Text Mining, Deep Learning, Electronic Components Classification, LLDA.

Feature selection for high-dimensional regression models with heteroscedastic errors

邱海唐*
國立中正大學數學系

彭柏翔、黃學涵、銀慶剛
國立清華大學統計學研究所

Abstract

We consider the problem of selecting high-dimensional regression models with heteroscedastic errors. A high-dimensional dispersion function is employed to account for the heteroscedasticity. By making use of the multi-step orthogonal greedy algorithm and the high-dimensional information criterion, we propose a new model selection procedure that consistently chooses the relevant features in both the regression and the dispersion functions. The finite sample performance of the proposed procedure is illustrated via simulations and real data analysis.

Keywords: Heteroscedasticity, High-dimensional information criterion, Orthogonal greedy algorithm.

Feature selection for high dimensional clustering problems

Cheng-Han Chua(蔡承翰)* and Meihui Guo(郭美惠)
National Sun Yat-sen University

Shih-Feng Huang(黃士峰)
National University of Kaohsiung

Abstract

This study proposes a Kriging-Correlation (KC) score to identify the first few vital clustering features from a large dimensional dataset. The KC score ranks the importance of each feature based on a 3-stage procedure. The first stage is to learn an appropriate distance metric from the data, and accordingly project the original data to a 2-dimensional space by the *t*-distributed stochastic neighbor embedding (t-SNE) method. We treat the coordinates of each sample in the 2-dimensional space as the vital clustering variables in the reduced space. The second stage is to reconstruct each original feature backward from the vital clustering variables in the reduced space by the AutoFRK method. The last stage is to rank the importance of the original features by comparing the dependence between each original feature and its reconstructed estimator, where the number of important features is determined by MANOVA. We applied the KC score to four single-cell datasets and compare its performances to the Laplacian score, which is a well-known unsupervised feature ranking method. The numerical results show that the KC score is capable of selecting a smaller set of features to achieve a better or comparable classification accuracy than the Laplacian score.

Keywords: Clustering, Feature Selection, t-SNE, AutoFRK, Laplacian Score.

Limiting spectral distribution of stochastic block model

Giap Van Su(蘇晉凡)*, Meihui Guo, Hao-Wei Huang
Department of Applied Mathematics
National Sun Yat-Sen University

Abstract

The stochastic block model, abbreviated to SBM, is a generative model for random graphs, which generates graph communities modeled by being connected with particular edge densities. For example, edges may be more common within communities than between communities. The SBM is important in statistics, such as clustering problems, machine learning, and complex network science, where it serves as a robust tool for recovering community structure in graph data. In this talk, we will introduce the SBM with two communities and present its empirical spectral distribution when the size of communities tends to infinity. It turns out that the limiting distribution is closely related to semi-circular law in the Wigner random matrices. The explicit limiting density function, along with a semi-circular approximation, and some applications will be given.

Keywords: Clustering, Semi-circular law, Spectral distribution, Stochastic block model.

The effect of parameter estimation on X-bar control charts for the lognormal distribution

Wei-Heng Huang(黃偉恆)
Department of Statistics
Feng Chia University

Abstract

Statistical process control (SPC) is a method of monitoring, controlling and improving a process through statistical analysis. In real applications, many industrial process data may follow a positively-skewed distribution such as the lognormal distribution. In this article, under lognormal process, we study the performance of three X-bar control charts, Shewhart control chart, weighted variance control chart and weighted standard deviation control chart, based on the in-control expected average run length (AARL). The standard deviation of ARL (SDARL) metric is used to evaluate the performance for various amounts of phase I data. A formula to approximate the ARL by the normal approximation method is established in this study. The in-control and out-of-control performance of the three X-bar control charts are compared based on the normal approximation and simulation.

Keywords：Average run length, Average of ARL, Lognormal distribution, Standard deviation of ARL, X-bar-chart.

Statistically optimal design of accelerated degradation tests

Ming-Yung Lee(李名鏞)
Department of Data Science and Big Data Analytics
Providence University

Abstract

For highly reliable products, accelerated degradation tests (ADTs) are often used to obtain degradation and/or failure data to estimate the lifetime distribution for developing, for example, a warranty policy or a maintenance schedule for the products. The step-stress ADT (SSADT) and parallel constant-stress ADT (PCSADT) are commonly used designs of ADT. In recent years, researchers also consider the experimental cost and budget, either as the objective function or as a constraint, when developing an optimal test plan. As a result, the lifetime estimate may not be sufficiently accurate or precise. However, for certain consumer products, the cost of conducting a traditional ADT, such as cost of electricity, is relatively low and hence can be ignored. In this paper, we propose a procedure that uses the estimation accuracy of Mean-Time-to-Failure (MTTF) or lifetime percentile (t_p) as the main objective to derive the samples size or the experimental termination time without experimental budget, when designing an optimal ADT plan. We first show that, for any design using an arbitrary accelerated stress function, there exists a corresponding SSADT design such that they have the same objective function value. Secondly, we show that the SSADT and PCSADT are equivalent under some conditions. Thirdly, we derive that the optimal ADT design is either a simple PCSADT or a simple SSADT. Furthermore, we provide a plan procedure for optimal ADT design and determine the optimal samples size and the experiment terminated time. Finally, we use the light emitting diode (LED) testing as an example to illustrate our results.

Keywords: Reliability, Accelerated Degradation Test, Statistically Equivalent Designs, Fisher Information Matrix, MLE, Margin of Error, Estimation Accuracy.

Bayesian analysis of accumulated damage models in lumber reliability

Chun-Hao Yang(楊鈞澔)*
Institute of Applied Mathematical Science
National Taiwan University

James V. Zidek
Department of Statistics
University of British Columbia

Samuel W. K. Wong
Department of Statistics and Actuarial Science
University of Waterloo

Abstract

Wood products that are subjected to sustained stress over a period of long duration may weaken, and this effect must be considered in models for the long-term reliability of lumber. The damage accumulation approach has been widely used for this purpose to set engineering standards. In this article, we revisit an accumulated damage model and propose a Bayesian framework for analysis. For parameter estimation and uncertainty quantification, we adopt approximation Bayesian computation (ABC) techniques to handle the complexities of the model. We demonstrate the effectiveness of our approach using both simulated and real data, and apply our fitted model to analyze long-term lumber reliability under a stochastic live loading scenario. Code is available at https://github.com/wongswk/abc-adm.

Keywords: Approximate Bayesian computation, Duration of load, Failure time distribution.

Design patterns for statistical computing

Wen Hsiang Wei(魏文翔)
Department of Statistics
Tung Hai University

Abstract

In software engineering, reuse-based development is a mainstream approach. Design patterns, originally used in civil engineering and architecture, is a commonly used technique for software reuse, for examples, the packages in Java. GoF (the Gang of Four), the authors of the classical book "Design Patterns: Elements of Reusable Objected-Oriented Software", introduced twenty three useful patterns for software design. Commonly used design patters can be used as developing R codes for statistical methods. Two examples including the patterns corresponding to the codes for a variety of regression models and different point estimators respectively are given for illustrations. The advantage or disadvantage of using design patterns for statistical computing is discussed.

Keywords: Design Patterns, Software Engineering, R.

A method of establishing a threshold of surrogate endpoint related to clinical endpoint

Yu-Chieh Cheng(鄭宇傑)*、Hsiao-Hui Tsou(鄒小蕙)
Institute of Population Health Sciences
National Health Research Institute

H. M. James Hung(洪賢明)
Division of Biometrics I, OB/CDER
Food and Drug Administration

Abstract

To develop medical products for some diseases, clinical response may take long time to capture. Use of surrogate endpoints has recently been increasingly of much interest at a minimum. If surrogate endpoints are available, then assess the potential treatment effect on the surrogate and subsequently confirm a clinical benefit. When there is a linear relationship between the surrogate and the clinical endpoint, the surrogate may still need to rule out a threshold that corresponds to no clinical benefit. Determination of such a threshold rests upon knowledge of many parameters in the bivariate statistical distribution of the clinical response and the surrogate. In our work, we present a concept of “working” threshold to incorporate statistical uncertainties in determination of such a threshold.

Keywords: surrogate endpoint, linear, null hypothesis, test statistic, type I error rate.

Euclidean statistical mathematics with application to spatial measuring

高正雄
國立中正大學數學系

Abstract

In outdoor survey of landscape or indoor measuring of volumes, surface areas and lengths of curves, the tools and their uses, together with the calculations with traditional mathematics are usually tedious and expensive in labor and costs. The talk shall show that there are ways to transform complicated mathematical problems in doing the survey work to simplified statistical problems regarding Euclidean volumes, areas, and lengths first. Then, by simply applying optical tools including laser distance guns and planary angle measurement devices from just a convenient location, we will be able to obtain the needed data to get solutions for the sizes of the questioned volumes, surface areas and curve lengths in the real world by related statistical treatments upon the data. Practical examples shall be illustrated to explain the speaker's primary thoughts on the whole idea.

On the asymptotic normality and efficiency of Kronecker envelope principal component analysis

Shih-Hao Huang (黃世豪)*
Department of Mathematics
National Central University

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

Abstract

Dimension reduction methods for matrix or tensor data have been an active research field in recent years. Kronecker envelope principal component analysis (PCA) is a two-step procedure, which consists of projecting data onto a multilinear envelope subspace as the first step, followed by ordinary PCA on the projected core tensor. The multilinear envelope subspace preserves the natural Kronecker product structure of observations when searching for the leading principal subspace. The main advantage of preserving the Kronecker product structure is the parsimonious usage of parameters in specifying the leading principal subspace, which mitigates the adverse influence of high dimensionality. The method of PCA will convert possibly correlated variables to uncorrelated ones and further reduce the dimension of the projected core tensor. In this article we derive the asymptotic normality of Kronecker envelope PCA and compare it with ordinary PCA. Utilizing majorization theory, we show that Kronecker envelope PCA is asymptotically more efficient than ordinary PCA. A motivating real data example of cryogenic electron microscopy image clustering and simulation studies are presented to show the merits of Kronecker envelope PCA.

Keywords: cryogenic electron microscopy, dimension reduction, Kronecker envelope, multilinear principal component analysis, principal component analysis, tensor data.

Consensus sparse principal component analysis

Sin-Cheng Ciou(邱信程)*
Mathematics
National Tsing Hua University

Yuh-Jye Lee
Mathematics
National Yang Ming Chiao Tung University

Abstract

Learning from data has become the mainstream in modern machine learning applications. The more data we have, our machine learning methods show better results if the data quality is good enough. However, in many cases, the data owners may not want to share or not allow to share the data they have because of privacy issues or legal concerns. To solve this problem, the consensus learning, also known as federated learning named by Google in 2016 [1]. This framework has been proposed and applied to linear and nonlinear SVM as well as PCA.

In this work, we will apply this framework to sparse principal analysis (SPCA) [2]. The SPCA direction will use as small as possible of the number of nonzero components and still keep the data variation when we project the data on the SPCA direction. Our proposed method, Consensus SPCA(CSPCA), will be solved by a distributed optimization algorithm, ADMM. It will allow different data owners only sharing with the model without sharing their own data, same with other federated learning models. We will demonstrate our CSPCA with synthetics dataset as well as some public datasets.

Keywords: Consensus learning, Federated learning, Sparse Principal Component Analysis.

[1] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data" in *Artificial Intelligence and Statistics*, 2017: PMLR, pp. 1273-1282.

[2] H. Zou, T. Hastie, and R. Tibshirani, "Sparse principal component analysis", *Journal of computational and graphical statistics*, vol. 15, no. 2, pp. 265-286, 2006.

具有固定共變數下製程優化之實驗設計策略

黃玉潔*、羅夢娜、曾聖澧
國立中山大學應用數學系

摘要

實驗設計在工業生產製程優化上一直擔任重要角色，傳統作法是先提出適當實驗設計，進行實驗並找出影響因子，進而改善製程。但實務上常只能取得產線上有限實驗設計因子組合下之相關數據，近乎觀察性資料。除了部分可控制之設計因子外，還有眾多不可控共變數，亦會影響實驗結果。同時在生產過程中，時有每個批次生產資料裏，僅能觀察到一組樣本的輸入變因的情形。又由於產品之檢測屬高成本破壞性試驗，在有相當多無法控制共變數情況下，進行傳統實驗設計並不可行。本研究旨在探討如何依據產線上所觀測到有限個因子試驗組合下，以及各類可控和不可控共變數，與產品品質參數等，來建構適當模型，提供最佳產品良率之實驗設計策略。期能依所建之模型，可提供至少一組具有高良率的因子設計組合，或直接標記此組物料品質可能異常，建議進行檢查，以減少物料浪費與生產耗損。

關鍵詞：不可控共變數、因子設計、高成本試驗、觀察性實驗、製程優化

A new statistical control for multiple stream processes

Meng-Chun Wu(吳孟峻)*, Su-Fen Yang, Yuo-Hsien Shiau,
Chia-Hu Huang, Tzee-Ming Huang
National Chengchi University, Taiwan

Abstract

A multiple stream process is a process that generates several streams of output. From the statistical process control standpoint, the quality variable and its specifications are the same in all streams. A filling machine with eight identical filling heads producing the cosmetic bottle covers every 4 hours. The cosmetic bottle cover weights are measured. The cover weights from each filling machine should have the same distributions and specifications. Hence, the production process is a kind of multiple stream process. In order to monitor whether the processes are in-control or out-of-control, we propose the EWMA mean and covariance control charts. We found that the proposed EWMA mean and covariance control charts perform well.

The discrepancy between GMM and FCN for segmenting medical images: Demonstration on breast sonography

Tai-Been Chen(陳泰賓)
Department of Medical Imaging and Radiological Sciences
I-Shou University
Institute of Statistics
National Yang Ming Chiao Yung University

Abstract

The segmentation of medical images is crucial for annotation, representation, and diagnostic lesions. Especially, the blurring and spike noise were intrinsic effects in the B-mode sonography. It is difficult to segment lesions in sonography. The conventional methods are applied K-means or Gaussian mixture method (GMM) to segment interesting regions on medical images. However, the novel deep learning method, fully convolutional neural network (FCN), was approved efficiently and accurately to depict the boundary of lesions. In this study, the breast sonography was used to investigate the segmented performance and discrepancy between GMM and FCN. These breast ultrasound B-mode images were obtained from the Open Access Series of Breast Ultrasonic Data (OASBUD) which were provided by the Polish Academy of Sciences. There were 52 malignant and 48 benign cases in dataset. Each case was two breast images. The global accuracy, mean accuracy, mean IoU (intersection over union), weighted IoU, and mean BF (boundary F1) Score were used to evaluated the performance of presented FCN model. The global accuracy, mean accuracy, mean IoU, weighted IoU, and mean BF score were 0.989, 0.970, 0.905, 0.979, and 0.725 provided by FCN. The experimental results were demonstrated the presented method feasibly and acceptably. Although, the information of pre-labeled boundary of images was needed to train FCN, the accurate boundary generated by FCN was higher than those by GMM. This method could also be further expanded on real clinical application at the hospital in the future work.

Keywords: GMM, FCN, Breast sonography.

Mixtures of factor analyzers with covariates for modeling multiply censored dependent variables

林宗儀
國立中興大學統計所

Abstract

Censored data arise frequently in diverse applications in which observations to be measured may be subject to some upper and lower detection limits due to the restriction of experimental apparatus such that they are not exactly quantifiable. Mixtures of factor analyzers with censored data (MFAC) have been recently proposed for model-based density estimation and clustering of high-dimensional data in the presence of censored observations. In this paper, we consider an extended version of MFAC by considering regression equations to describe the relationship between covariates and multiply censored dependent variables. Two analytically feasible EM-type algorithms are developed for computing maximum likelihood estimates of model parameters with closed-form expressions. Moreover, we provide an information-based method to compute asymptotic standard errors of mixing proportions and regression coefficients. The utility and performance of our proposed methodology are illustrated through a simulation study and two real data examples.

Keywords: AECM algorithm, Censored data, Detection limit, Factor analysis, ML estimation, Truncated multivariate normal distribution.

基於不正確登錄之雙零件供應商型一區間設限資料之可靠度推論

蔡宗儒
淡江大學統計學系

Hon Keung Tony Ng
Department of Statistical Science
Southern Methodist University

Hoang Pham
Department of Industrial Systems Engineering
Rutgers University

Yuhlong Lio
Department of Mathematical Sciences
University of South Dakoda

江俊佑
西南財經大學統計學院

摘要

本研究討論當一產品的零件由雙供應商提供，使用型一區間設限方案來登錄資料，但資料登錄不正確下，如何進行產品的可靠度推論，研究中將討論如何使用剖面最大概似估計法及貝氏估計法來得到推論的結果，並以一組 VGA 轉接器的實際案例說明本方法的應用。

關鍵詞：型一區間設限、剖面最大概似估計、貝氏估計、馬可夫鏈蒙地卡羅法

探討不同深度學習架構應用於胸部 X 光肺結節分類之性能評估

王豪善*、謝文權
義守大學生物科技學系

黃詠暉、陳泰賓
義守大學醫學影像暨放射科學系

蔡汎修
義守大學學士後中醫學系

摘要

目的: 胸部 X 光(Chest X-Ray)造影為肺癌篩檢最常見的工具之一；透過影像肺結節型態進行診斷。然而胸部 X 光為三維投影成二維影像，因此病灶與正常組織因重疊或肋骨干擾而影響診斷。故本研究將探討不同深度與機器學習架構應用於胸部 X 光肺結節分類之性能評估。

材料與方法：本研究為回顧性研究。胸部 X 光影像資料採用 JSRT(Japanese Society of Radiological Technology)之結節與非結節影像各 50 張；接著分割資料 70%與 30%分別作為訓練組與驗證組；同時評估六種深度及三種機器學習模型組合;六種深度學習模型分別為 VGG16、VGG19、Resnet50、Resnet101、Darknet19 及 Darknet53；三種機器學習模型為 SVM (Support Vector Machine)、KNN (K-Nearest Neighbors)及 LR (Logistic Regression)分類模型，共評估 18 種分類模型；以靈敏度、特異性、陽性預測值、陰性預測值及 Kappa 值進行模型效能評估與比較。

結果：透過驗證組資料評估顯示 Darknet53 具有最好的靈敏度、特異性、陽性預測值、陰性預測值、準確性、Kappa 值；分別為 0.93、0.80、0.82、0.92、0.86、0.73。

結論:透過深度及機器學習模型對胸部X光進行肺結節分類可達到良好的分類效能；未來除增加臨床案例，將持續探討深度學習在不同層(Layers)產生之影像特徵矩陣並配合機器學習模型參數調整，檢視分類效能以輔助臨床辨識之可能性。

關鍵詞：胸部 X 光、深度學習模型、機器學習模型

全卷積類神經網路演算法應用於 T1 與 T2 大腦磁振融合影像進行腫瘤分割

花郁婷*、謝文權
義守大學生物科技學系

陳泰賓
義守大學醫學影像暨放射科學系

摘要

目的：磁振造影(Magnetic Resonance Imaging, MRI)對腦部檢查能夠提供高解析度軟組織辨別度；然而， T1 和 T2 權重影像用於鑑別惡性腦腫瘤時，常因腫瘤邊緣會與正常組織強度近似，而易生分辨困擾。因此透過深度學習中的全卷積類神經網路(Fully Convolutional Neural Network, FCN)進行腫瘤分割，達成電腦自動輔助圈選腫瘤位置及邊界之目的。

材料與方法：本研究為回顧性研究，收集 44 組經 T1 與 T2 造影之腦腫瘤影像。透過 T1、T2、以及融合 T1 與 T2 成為 RGB 三通道之彩色影像，再透過 FCN 模型進行切割腫瘤模型之訓練；其效能評估採用整體準確率(Global Accuracy)、平均準確率(Mean Accuracy)、平均交疊率(Mean Intersection over Union, IoU)、加權交疊率(Weighted IoU)及平均邊界 F-1 分數(Mean Boundary F-1 Score, BF)。

結果：根據 FCN 模型效能評估，其整體準確率、平均準確率、平均交疊率、加權交疊率及平均邊界 F-1 分數分別為 0.99、0.96、0.92、0.99、0.95。

結論：透過 T1、T2、以及融合 T1 與 T2 成為 RGB 三通道之彩色影像，能提供 FCN 模型較準確切割腦腫瘤邊界；未來除了繼續增加訓練樣本數，亦將朝向臨床應用以達成輔助腦腫瘤電腦切割之目的。

關鍵詞：磁振造影、深度學習、全卷積網路

基於機器學習 XGBoost 模型之台南永康區氣象預測

何應承*、楊松霈
崑山科技大學綠能科技研究中心

摘要

本研究先透過網路爬蟲氣象系統，串接氣象局應用程式介面(Application Programming Interface, API)將台灣各鄉鎮市區氣象預報未來 2 天的公開資料，以攫取氣象局(Central Weather Bureau, CWB)鄉鎮預報編號 F-D0047-093 的壓縮檔資料，即可收集獲得所需的台南市永康區氣象資料集。本研究使用台南市永康區 2020 年 8 月至 2021 年 10 月初的氣象歷史資料，採用機器學習 XGBoost(Extreme Gradient Boosting)演算法做模型訓練，在建模過程中探討影響氣象溫度之重要變數，發現時間索引相關的 8 個特徵欄位中，以每年第幾天(day of year)為其最重要的影響變數。本研究係將多個弱學習器用迴歸樹(Classification And Regression Tree, CART)模型累加集成，以優化成為強學習器，進而使目標損失函數達到極小，故從預測資料測試集結果，即可得 XGBoost 模型以極限梯度提升，因此預測準確度相當好。實驗結果 XGBoost 模型的損失函數 MSE 為 1.28、RMSE 為 1.13、MAE 為 0.87，以及 MAPE 為 3.09%。

關鍵詞：人工智慧、機器學習、XGBoost、氣象預測、大數據分析

A network-based method for predicting fungal essential genes through identification of core genes

Pei-Yu Lin(林蓓郁)[1]*, Yu-Chun Huang(黃郁珺)[1,2]
[1]Insititute of Plant and Microbial Biology,
Academia Sinica
[2]Bioinformatics, Taiwan International Graduate Program,
Academia Sinica, National Taiwan University

Hsiao-Ching Lin(林曉青)
Institute of Biological Chemistry
Academia Sinica

Pao-Yang Chen(陳柏仰)
Insititute of Plant and Microbial Biology
Academia Sinica
Bioinformatics, Taiwan International Graduate Program
Academia Sinica, National Taiwan University

Abstract

In fungal species, the essential genes are particularly helpful for the identification of antifungal drug targets, and the prediction of biosynthetic gene clusters. With resource expensive for experimentally constructing a catalog of essential genes, computational approaches to precisely identify candidate essential genes would be invaluable. Here, we present a network-based approach to predict fungal essential genes. To this end, a total of 491 fungal genomes (68% of assembled) kindly shared by communities were collected. We implemented Louvain algorithm for effectively clustering 6M fungal proteins from these fungal genomes based on sequence similarity, resulted in 67,826 orthologous gene clusters; each represents a group of similar proteins sharing amongst several fungal species. Each ortholog cluster was exploited as a sub-network where nodes are proteins and edges are the protein similarity. With their network statistics as parameters, a generalized linear model (GLM) and a random forest model were built to accurately rank these subnets by

their likelihood of originating from an essential gene. We found that the top ranked subnets were of two types, one exhibits in many species (global) and the other is only vital for a few closed-relative fungi (local) with a higher network density, suggesting these local essential genes may be unique to specific yet close fungal families for living. For examples two of our predicted local essential genes, CFT1 which takes part in mRNA cleavage and UTP6 encoding for nucleolar proteins, are found to be close species well within genus Nakaseomyces. As a validation, our approach coupling with either GLM or random forest model reached 84% and 91% accuracy in predicting known essential genes from three well-studied species (S. pombe, S. cerevisiae and A. fumigatus). Additionally, GLM-based prediction tends find more previously undiscovered essential genes, for instance, several novel gene subnets associated with specific functions such as tetratricopeptide repeats coding protein or citrate synthase which have been shown to be essential for Eukaryotes. Our prediction strategy is based on a large number of fungal genomes, combining network biology, statistical modelling and machine learning, to provide a ranked list of fungal core genes. The results as a web database would serve as a valuable resource for fungal genomic research.

Keywords: essential genes, fungi, orthologs, GLM, network statistics.