Joint analysis of panel count and interval-censored data using distribution-free frailty analysis

Chi-Chung Wen
Department of Mathematics
Tamkang University
E-mail: ccwen@mail.tku.edu.tw

We propose a joint analysis simultaneously analyzing recurrent and non-recurrent events subject to general types of interval censoring. The proposed analysis allows for general semiparametric models, including the classes of Box-Cox transformation and inverse Box-Cox transformation models for the recurrent and nonrecurrent events, respectively. A frailty variable is used to account for the potential dependence between the recurrent and non-recurrent event processes. We apply the pseudo likelihood for interval-censored recurrent event data, usually termed as panel count data, and the sufficient likelihood for interval-censored non-recurrent event data. Conditioning on the sufficient statistic for the frailty, and using the working assumption of independence over examination times, the sufficient likelihood does not rely on distributional assumptions on the frailty, and can deal with general interval censorship. We illustrate the proposed methodology by a joint analysis of the numbers of occurrences of basal cell carcinoma over time and time to the first recurrence of squamous cell carcinoma based on a skin cancer dataset, as well as a joint analysis of the numbers of adverse events and time to premature withdrawal from study medication based on a scleroderma lung disease dataset. This is a joint work with Yi-Hau Chen.

Keywords: Correlated data; Joint model; Recurrent event; Semiparametric model; Survival analysis.

Analytical expression for the integrated squared density partial derivative of a multivariate normal mixture distribution

Min-Hsiao Tsai
Department of Statistics
National Taipei University
E-mail: mhtsai@mail.ntpu.edu.tw

This investigation describes the derivation of the analytical expression for the integrated squared density partial derivative (ISDPD) in a multivariate normal mixture model. The analytical expression of the ISDPD is derived for arbitrary dimensions with partial derivative orders up to four. Although the value of the ISDPD can be obtained by using the common numerical integration via mathematical software (such as Maple, Mathematica, Matlab, etc), it suffers from the limitation of computation time when the dimension or the number of mixture components of the considered multivariate normal mixture model are large. Moreover, numerical comparison indicates the benefits of speed offered by our proposed analytical expression are far superior to the numerical integration performed by Maple. With this analytical expression, the ISDPD can apace be calculated with no assistance of numerical integration.

Keywords: integrated squared density partial derivative, multivariate normal mixture model, numerical integration, computation time.

台灣參與國際數學奧林匹亞競賽之統計分析

高竹嵐
統計學研究所
國立交通大學
chulankao@gmail.com

台灣於 2019 年參與第六十屆國際數學奧林匹亞競賽之團體排名達到歷史新低，但這是否代表台灣隊之競賽能力退步？此退步與否又應如何評估？本文在考慮各屆參賽國與選手差異，以及各年度題目難度與得分結構之差異下，結合貝氏 ordinal regression 與 EM 演算法，對台灣隊之能力進行分析。研究結果顯示台灣隊之能力並無顯著下降。本文所提出之新研究方式，期望能對未來台灣參與類似競賽與測驗上，從統計上提供新的分析工具。

Keywords: 國際數學奧林匹亞，Ordinal Regression，EM 演算法。

Comparison of the marginal hazard model and the sub-distribution hazard model under an assumed copula

Takeshi Emura
Graduate Institute of Statistics
National Central University, Taiwan
Email: takeshiemura@gmail.com,

For the analysis of competing risks data, three different types of hazard functions have been considered in the literature, namely the cause-specific hazard, the sub-distribution hazard, and the marginal hazard function.

Let X be a nonnegative random variable for time to "Event 1" and Y be the one for time to "Event 2". Under competing risks, we observe the first occurring event time $T = \min(X, Y)$, and the event indicator $\delta = \mathbf{I}(T = X)$, where $\mathbf{I}(\cdot)$ is the indicator function. The *marginal hazard function* for Event 1 is defined as

$$\lambda_1(t) = \Pr(t < X \le t + dt \mid X > t)/dt .$$

The marginal distribution is not identifiable from the distribution of (T, δ) unless some assumptions are made on the joint distribution of (X, Y) [1]. Other types of hazard functions of interest are therefore often considered for competing risks analysis. What can be identified from (T, δ) without knowing or assuming the distribution of (X, Y) is the *cause-specific* (CS) *hazard function* [2]. For Event 1 it is defined as

$$\lambda_1^{CS}(t) = \Pr(t < T \le t + dt, \delta = 1 \mid T > t)/dt .$$

The other identifiable quantity is the *sub-distribution hazard function* [3]. For Event 1, it is defined as

$$\lambda_1^{Sub}(t) = \Pr(t < T \le t + dt, \delta = 1 \mid \{T > t\} \bigcup \{T \le t, \delta = 0\})/dt .$$

While the relationship between the cause-specific hazard and the sub-distribution hazard has been extensively studied [4], the relationship to the marginal hazard function has not yet been analyzed due to the difficulties related to non-identifiability. In this paper, we adopt an assumed copula model [5] to deal with the model identifiability issue, making it possible to establish a relationship between the sub-distribution hazard and the marginal hazard function.

To model the dependence between two competing event times, we adopt a *survival copula model* [6]

$$\Pr(X > x, Y > y) = C_\theta\{ S_1(x), S_2(y) \}$$

where $C_\theta : [0,1]^2 \mapsto [0,1]$ is a copula function with a parameter θ [7]. The copula function can be any bivariate distribution function having the uniform marginal distribution on (0,1).

Under the survival copula model, we derive a mathematical relationship between $\lambda_1(\cdot)$ and $\lambda_1^{Sub}(\cdot)$. Furthermore, we establish a necessary and sufficient condition for $\lambda_1(\cdot)$ and $\lambda_1^{Sub}(\cdot)$ to be equivalent. We then compare the two methods of fitting the Cox model to competing risks data [3, 8]. We also extend our comparative analysis to *clustered* competing risks data under a *joint frailty-copula model* [9, 10]. For illustration, we analyze two survival datasets from lung cancer and bladder cancer patients.

This full paper [11] is currently under review and joint work with Shih JH, Il Do Ha, and Ralf Wilke.

Key words: Clustered survival data, competing risk, Cox model, frailty model, survival analysis

References:

[1] Tsiatis A (1975). A nonidentifiability aspect of the problem of competing risks. *Proc. Natn. Acad. Sci*. USA, 72: 20-22.

[2] Kalbfleisch JD, Prentice RL (2002). *The Statistical Analysis of Failure Time Data, 2nd Edition*, John Wiley and Sons, New York

[3] Fine JP, Gray RJ (1999). A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association* 94: 548-560.

[4] Bakoyannis G, Touloumi G (2012). Practical methods for competing risks data: a review. *Statistical Methods in Medical Research*; 21: 257-272.

[5] Zheng M, Klein JP (1995). Estimates of marginal survival for dependent competing risks based on an assumed copula. *Biometrika*, 82(1), 127-138.

[6] Escarela G, Carriere JF (2003). Fitting competing risks with an assumed copula. *Statistical Methods in Medical Research* 12: 333-349.

[7] Nelsen RB (2006). *An Introduction to Copulas, 2nd Edition*. Springer Series in Statistics, Springer-Verlag, New York.

[8] Chen YH (2010). Semiparametric marginal regression analysis for dependent competing risks under an assumed copula, *Journal of the Royal Statistical Society, Ser. B*; 72: 235-51.

[9] Emura T, Nakatochi M, Murotani K, Rondeau V (2017). A joint frailty-copula model between tumour progression and death for meta-analysis, *Statistical Methods in Medical Research* 26 (6): 2649-2666.

[10] Emura T, Matsui S, Rondeau V (2019), *Survival Analysis with Correlated Endpoints, Joint Frailty-Copula Models*, JSS Research Series in Statistics, Springer, Singapore.

[11] Emura T, Shih JH, Ha ID, Wilk RA (2020), Comparison of the marginal hazard model and the sub-distribution hazard model under an assumed copula, *in revision*.

Local variable selection criterion based on a prediction perspective

Chun-Shu Chen
Graduate Institute of Statistics
National Central University
cschen1207@ncu.edu.tw

Variable selection and spatial prediction both are important issues in spatial statistics. If spatially varying means exist among different subareas, globally fitting a spatial regression model for the study area may be not suitable. To alleviate deviations from model assumptions, we propose a local variable selection criterion to locally select variables for each subarea. The proposed local criterion considers the global spatial dependence of observations and the characteristics of each subarea are also identified. It results in a composite spatial predictor which not only provides a more accurate spatial prediction, but also reduces the prediction variance. Statistical inferences of the proposed methodology are justified both theoretically and numerically.

Keywords: Information criterion, prediction variance, resampling, squared prediction error

Data-driven multistratum designs with the generalized Bayesian D-D criterion for highly uncertain models

Chang-Yun Lin

Department of Applied Mathematics and Institute of Statistics,

National Chung Hsing University, Taichung, Taiwan, 40227

Abstract

Multistratum designs have gained much attention recently. Most criteria, such as the D criterion, select multistratum designs based on a given model that is assumed to be true by the experimenters. However, when the true model is highly uncertain, the model used for selecting the optimal design can be seriously misspecified. If this is the case, then the selected multistratum design will be not efficient for fitting the true model. To deal with the problem of high uncertain models, we propose the generalized Bayesian D-D (GBDD) criterion, which selects multistratum designs based on the experimental data. Under the framework of multistratum structures, we develop theorems and formula that are used for conducting Bayesian analysis and extracting information about the true model from the data to reduce model uncertainty. The GBDD criterion is easy and flexible in use. We provide several examples to demonstrate how to construct the GBDD-optimal split-plot, strip-plot, and staggered-level designs. By comparing with the D-optimal designs and one-stage generalized Bayesian

D-optimal designs, we show that the GBDD-optimal designs have higher efficiency on fitting the true models. The extensions of the GBDD criterion for more complicated cases, such as more than two stages of experiments and more than one class of potential terms, are also developed.

KEY WORDS: Bayesian D criterion, D criterion, split-plot design, staggered-level design, strip-plot design, two-stage experiment.

A non-parametric method for inferring parameters in ecological networks

Wei-Chung Liu
Institute of Statistical Science
Academia Sinica
E-mail: wliu56@gate.sinica.edu.tw

The simplest type of ecological networks, or a food web, is a representation of who eats whom in an ecosystem. Such a network is often generated by a single dataset aggregated from one or several surveys. Point estimates of network parameters can be calculated from the data, but how to quantify their corresponding interval estimates still remains elusive. Here, a simple bootstrap-based resampling procedure is proposed for inferring network parameters. First, for a particular network parameter, we obtain its point estimate by calculating the corresponding statistics from the original network. Second, we generate a resampled network by sampling with replacement the same number of species from the original network, and for each resampled species we record how many prey items it consumes in the original network. Third, a resampled species is allowed to consume its original prey species if such a species is also present; if not then it instead consumes the resampled species that is most topologically similar to its original prey species. Several resample networks can be constructed from which the sampling distribution and the interval estimate for this particular statistics can then be determined. We demonstrate our methodology on two different ecological networks and discuss its application in comparing ecosystems of various sizes and complexity.

Keywords: ecological network, food web, network parameter, bootstrap, resampled network, sampling distribution, interval estimate.

On the matrix condition of phylogenetic tree

Tony Jhwueng
Department of Statistics
Feng-Chia University
E-mail: dcjhwueng@fcu.edu.tw

Phylogenetic comparative analyses incorporate phylogenetic tree to study evolutionary relationship among a group of related species. A phylogenetic tree of n taxa can be algebraically transformed into an n by n squared symmetric phylogenetic covariance matrix C where each element c_{ij} in C represents the affinity between extant species i and extant species j. Because C plays an important role in phylogenetic comparative analysis, it deserves a rigorous investigation of the matrix condition of C. The condition number of matrix C denoted by κ is defined by the ratio of the maximum eigenvalue of C to the minimum eigenvalue of C. When tree has ill-conditioned matrix C, results obtained from subsequent analyses such as computing the likelihood that requires inversion of C may not be stable. To remediate this problem, we propose several methods to appropriately adjust the phylogenetic tree and improve the matrix condition of C for the purpose of obtaining reliable results.

Keywords: matrix condition, Brownian motion, phylogenetic comparative analysis