Statistics and Its Interface Volume 11 (2018) 201–202

Developments in the analysis of longitudinal data, dimension reduction, and beyond

Heping Zhang

Analysis of longitudinal data is a very important topic. Kürüm et al. [1] proposed a copula-based joint modeling framework for mixed longitudinal responses. Their approach permits all model parameters to vary with time, which is useful in revealing dynamic response–predictor relationships and response–response associations as demonstrated through their analysis of Women's Interagency HIV Study. Understanding HIV viral load dynamics also requires the modeling of complex longitudinal data for which nonlinear mixed-effects (NLME) models are commonly used. In those models, random errors are generally assumed normally distributed, but normality may not be validate in practice. Han et al. [2] adopt a Bayesian-frequentist hybrid (BFH) approach to NLME models with a skew-normal distribution for the covariate measurement errors. Their approach jointly models the response and covariate processes, and led to insightful understanding of the HIV viral dynamics from the data in an AIDS clinical trial. Related to [1, 2], modeling survival time AIDS patients is of clinical significance. Gómez [3] introduced two families of distributions that are suitable for fitting unimodal as well as bimodal symmetric and asymmetric censored data. Their models extend the skew normal model to bimodal symmetric and asymmetric situations and typically involves less parameters to be estimated than mixtures of normal distributions. The utility of their approach is demonstrated through the analysis of the data from a study on antiretroviral therapy (HAART) to AIDS patients.

Also on analysis of longitudinal data, Wang and Castro [4] were concerned with multiple trajectories following arbitrary growth patterns in the presence of outliers and possible missing responses. Because the likelihood function is intractable, they devised a fully Bayesian estimating procedure to account for the uncertainties of model parameters, random effects, and missing responses via the Markov chain Monte Carlo method.

Missing data in longitudinal data give rise to serious challenges in the statistical modeling, including the identifiability of model parameters. To overcome those challenges, Zhao et al. [5] made use of several novel techniques such as the generalized method of moments, an augmented inverse probability weighting estimator, and the importance sampling.

Again on the longitudinal data, Chen et al. [6] proposed a nonparametric multivariate control scheme for simultaneously monitoring several related characteristics of a process in time. This method can quickly detect small mean and/or variance shifts in various types of longitudinal processes, Gaussian or non-Gaussian.

Dimension reduction is one of the most challenging and lasting problem in statistical modeling. Principal component analysis (PCA) is the most commonly used approach, but it raises computational challenges when it is applied big data. Zhang and Yang [7] introduced a practical improvement over PCA by providing exact solutions when the size of observed data exceeds the memory size of a computing system. High-dimensional stationary time series arise from finance industry. Xia et al. [8] suggested a method to determine the number of factors in factor modeling for such data. Interactions among genes, known as epistasis, are important to our understanding of common complex diseases. Niu et al. [9] considered interaction screening for high dimensional quadratic regression models. They proposed a main-effect-adjusted interaction screening procedure which selects interactions while taking into account main effects. It is a unified framework and can be employed to Pearson correlation coefficient, and as well as nonparametric rank-based measures such as Spearman's and Kendall's correlation coefficients.

How to deal with zero-total-event data, meaning zero events in both treatment and control arms, has long been debated, and received much renewed interest recently. Xie et al. [10] provided a timely and detailed comparison of two approaches: the regular likelihood approach and the classical conditional likelihood approach. They found that, "when we assume the underlying population event rates are not zero, an observed zero-total-event study actually contains information for inference on the parameters such as the common odds ratio in meta-analysis and cannot be left out in our analysis. This is contrary to the belief held by many statisticians that an observed zero-total-event study does not contribute to meta-analysis because it does not contain any information concerning the common odds ratio." Their finding helps clarify a difficult question concerning how to deal with zero-total-event studies in meta-analysis of rare event studies.

Count data are commonly collected in healthcare industry, and sometimes they have excess zeros or excess ones. To deal with this particular issue, Liu et al. [11] proposed a zero-one-inflated Poisson model. Also on count data, Li et al. [12] dealt with the topic on missing data in two-way contingency tables which often occur in multi-center clinical trials by presenting a reasonable joint distribution of the observed counts in an incomplete contingency table and testing the homogeneity between two correlated proportions in multiple incomplete two-way contingency tables.

The presence of ordinal data is reality of our lives. For example, Nandram and Peiris [13] described a taste-testing experiment in which foods were withdrawn from storage at various times and a panel of tasters were asked to rate the foods on a nine-point hedonic scale. Nandram and Peiris [13] provided a Bayesian procedure to assess the difference between fresh foods and foods withdrawn a few months later.

Identifying diffierentially expressed genes is important for cancer diagnosis. Yang et al. [14] presented a stochastic variable selection approach for gene selection with different two level hierarchical prior distributions for regression coefficients.

Lastly, in this issue, we introduced a new forum on challenges arising from the practice [15]. We welcome short essays on emerging and important problems that call for novel statistical ideas and solutions.

REFERENCES

[1] Esra Kürüm, John Hughes, Runze Li, and Saul Shiffman (2018). Time-varying copula models for longitudinal data. *Statistics and Its Interface.* **11** 203–221.

[2] Gang Han, Yangxin Huang, and Ao Yuan (2018). Bayesian-frequentist hybrid approach for skew-normal nonlinear mixed-effects joint models in the presence of covariates measured with errors. *Statistics and Its Interface.* **11** 223–236.

[3] Guillermo Martínez-Flórez, Heleno Bolfarine, and Héctor W. Gómez, (2018). Censored bimodal symmetric-asymmetric families. *Statistics and Its Interface.* **11** 237–249.

[4] Wan-Lun Wang and Luis Mauricio Castro (2018). Bayesian inference on multivariate-t nonlinear mixed-effects models for multiple longitudinal data with missing values. *Statistics and Its Interface.* **11** 251–264.

[5] Puying Zhao, Lei Wang, and Jun Shao (2018). Analysis of longitudinal data under nonignorable nonmonotone nonresponse. *Statistics and Its Interface.* **11** 265–279.

[6] Yuhui Chen, Mingwei Sun, and Timothy Hanson (2018). Nonparametric multivariate Polya Tree EWMA control chart for process changepoint detection. *Statistics and Its Interface.* **11** 281–293.

[7] Tonglin Zhang and Baijian Yang (2018). Dimension reduction for big data. *Statistics and Its Interface.* **11** 295–306.

[8] Qiang Xia, Rubing Liang, Jianhong Wu, and Heung Wong (2018). Determining the number of factors for high-dimensional time series. *Statistics and Its Interface.* **11** 307–316.

[9] Yue Selina Niu, Ning Hao, and Hao Helen Zhang (2018). Interaction screening by partial correlation. *Statistics and Its Interface.* **11** 317–325.

[10] Min-ge Xie, John Kolassa, Dungang Liu, Regina Liu, and Sifan Liu (2018). Does an observed zero-total-event study contain information for inference of odds ratio in meta-analysis? *Statistics and Its Interface.* **11** 327–337.

[11] Wenchen Liu, Yincai Tang, and Ancha Xu (2018). A zero-one-inflated Poisson model and its application. *Statistics and Its Interface.* **11** 339–351.

[12] Huiqiong Li, Niansheng Tang, Guoliang Tian, and Hongyuan Cao (2018). Testing the equality of risk difference among multiple incomplete two-way contingency tables. *Statistics and Its Interface.* **11** 353–368.

[13] Balgobin Nandram and Thelge Buddika Peiris (2018). Bayesian analysis of a ROC curve for categorical data using a skew-binormal model. *Statistics and Its Interface.* **11** 369–384.

[14] Aijun Yang, Heng Lian, Niansheng Tang, and Xinyuan Song (2018). Sparse Bayesian variable selection for classifying high-dimensional data. *Statistics and Its Interface.* **11** 385–395.

[15] Hongying Kuang, Hao Huang, Haizhu Tan, Allen R. Kenselman, and Heping Zhang (2018). Challenges in analyzing time to live birth. *Statistics and Its Interface.* **11** 397–399.

Heping Zhang
Department of Biostatistics
Yale University School of Public Health
New Haven, CT 06520-8034
USA
E-mail address: heping.zhang@yale.edu