Improving Inference and Variable Selection for Two-Phase Studies with High-Dimensional Covariates

Haoyang Wang[1], Qingning Zhou[2], and **Kin Yau Wong**[1]

[1]*The Hong Kong Polytechnic University*

[2]*The University of North Carolina at Charlotte*

ABSTRACT

The two-phase study design is widely used to improve estimation efficiency and reduce cost. In many two-phase studies, the outcome and inexpensive covariates are obtained on all subjects in Phase I, while expensive covariates are measured only on a subset of subjects in Phase II. As a result, regression analysis of two-phase studies faces a missing data problem. In this presentation, we consider two-phase studies with high-dimensional covariates, where one faces a more challenging high-dimensional missing data problem. For this problem, complete-case analyses are generally inefficient, while imputation or likelihood-based methods usually require a model for the missing covariates, which is almost impossible to correctly specify in high-dimensional settings. To overcome these limitations, we propose a two-step estimation method that refines a complete-data estimator by incorporating the incomplete data, and auxiliary information if available, to improve estimation efficiency. This method avoids the need to correctly specify a model for the missing covariates, and the resulting estimator is guaranteed to be at least as efficient as the complete-data estimator. We also establish theoretical guarantees for the proposed method, including estimation consistency, inference validity, and variable selection consistency. We evaluate the performance of the proposed method via simulation studies and provide an application to a major cancer study for illustration.

Keywords: Debiased estimation; High-dimensional regression; Missing data; Robust estimation.

Efficient Estimation for Functional Accelerated Failure Time Model

Changyu Liu, Wen Su, **Kin Yat Liu**, Guosheng Yin, Xingqiu Zhao

Department of Statistics and Data Science, The Chinese University of Hong Kong

ABSTRACT

We propose a functional accelerated failure time model to characterize effects of both functional and scalar covariates on the time to event of interest, and provide regularity conditions to guarantee model identifiability. For efficient estimation of model parameters, we develop a sieve maximum likelihood approach where parametric and nonparametric coefficients are bundled with an unknown baseline hazard function in the likelihood function. Not only do the bundled parameters cause immense numerical difficulties, but they also result in new challenges in theoretical development. By developing a general theoretical framework, we overcome the challenges arising from the bundled parameters and derive the convergence rate of the proposed estimator. Furthermore, we prove that the finite-dimensional estimator is root-n consistent, asymptotically normal and achieves the semiparametric information bound. And we demonstrate the nonparameteric optimality of functional estimator and construct the asymptotic simultaneous confidence band. The proposed inference procedures are evaluated by extensive simulation studies and illustrated with an application to the National Health and Nutrition Examination Survey data.

Keywords: Functional Data Analysis; Survival Analysis

Semiparametric Causal Inference for Right-Censored Outcomes with Many Weak Invalid Instruments

Qiushi Bu[1,2], Wen Su[1], Xingqiu Zhao[3], and Zhonghua Liu[4]

[1]*Department of Biostatistics, City University of Hong Kong, Hong Kong*

[2]*Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, China*

[3]*Department of Applied Mathematics, The Hong Kong Polytechnic University, Hong Kong*

[4]*Department of Biostatistics, Columbia University, New York, NY, USA*

ABSTRACT

We propose a semiparametric framework for causal inference with right-censored survival outcomes and many weak invalid instruments, motivated by Mendelian randomization in biobank studies where classical methods may fail. We adopt an accelerated failure time model and construct a moment condition based on augmented inverse probability of censoring weighting, incorporating both uncensored and censored observations. Under a heteroscedasticity-based condition on the treatment model, we establish point identification of the causal effect despite censoring and invalid instruments. We propose GEL-NOW (Generalized Empirical Likelihood with Non-Orthogonal and Weak moments) for valid inference under these conditions. A divergent number of Neyman orthogonal nuisance functions is estimated using deep neural networks. A key challenge is that the conditional censoring distribution is a non-Neyman orthogonal nuisance, contributing to the first-order asymptotics of the estimator for the target causal effect parameter. We derive the asymptotic distribution and explicitly incorporate this additional uncertainty into the asymptotic variance formula. We also introduce a censoring-adjusted over-identification test that accounts for this variance component. Simulation studies and UK Biobank applications demonstrate the method's robustness and practical utility.

Keywords: Censored outcomes; Deep neural networks; Generalized empirical likelihood; Mendelian randomization; Over-identification test; Semiparametric theory; Weak and invalid instruments

Efficient Estimation for Deep Accelerated Failure Time Model with Application to Credit Risk Analysis

Kun Ren[1], Li Liu[2,*], Wen Su[1], Xingqiu Zhao[3,*]

[1]*Department of Biostatistics, City University of Hong Kong*

[2]*School of Mathematics and Statistics, Wuhan University*

[3]*Department of Applied Mathematics, The Hong Kong Polytechnic University*

ABSTRACT

Time-to-event data, frequently encountered in credit risk analysis and engineering reliability studies, present significant analytical challenges due to censoring and data heterogeneity. We develop a deep semiparametric accelerated failure time model designed for right-censored data, which accommodates complex data structures while maintaining interpretability through parametric components. Our approach integrates the flexibility of deep neural networks with traditional smoothing techniques, preserving the benefits of parametric interpretability. Furthermore, we establish both nonasymptotic and asymptotic properties for the proposed method, including prediction error bounds, asymptotic normality, and semiparametric efficiency of the estimator. Simulation studies demonstrate the superior performance of our method compared to conventional smoothing approaches. Finally, we apply our method to analyze the German credit data.

Keywords: Default risk; Deep neural network; Semiparametric efficiency; Survival data.