Logistics Regression Model for Differentially-Private Matrix Masking Data

Linh H. Nghiem[1], Aidong Adam Ding[2], **Samuel S. Wu[3]**

[1]*School of Mathematics and Statistics, University of Sydney, Sydney, Australia*

[2]*Department of Mathematics, Northeastern University, Boston, USA*

[3]*Health Informatics Institute, University of South Florida, Tampa, USA*

ABSTRACT

A recently proposed scheme utilizing local noise addition and matrix masking enables data collection while protecting individual privacy from all parties, including the central data manager. Statistical analysis of such privacy-preserved data is particularly challenging for nonlinear models like logistic regression. By leveraging a relationship between logistic regression and linear regression estimators, we propose the first valid statistical analysis method for logistic regression under this setting. Theoretical analysis of the proposed estimators confirmed its validity under an asymptotic framework with increasing noise magnitude to account for strict privacy requirements. Simulations and real data analyses demonstrate the superiority of the proposed estimators over naive logistic regression methods on privacy-preserved data sets.

Keywords: Logistic regression; Differential privacy; Matrix masking; Mixture normal

Partially-Global Fréchet Regression

Danielle C. Tucker and **Yichao Wu[1]**

Department of Mathematics, Statistics, and Computer Science, University of Illinois Chicago

ABSTRACT

We propose a partially-global Fréchet regression model by extending the profiling technique for the partially linear regression model (Severini and Wong 1992). This extension allows for the response to come from a generic metric space and can incorporate a combination of Euclidean predictors and a predictor which comes from another generic metric space. By melding together the local and global Fréchet regression models proposed by Petersen and Müller (2019), we gain a model that is more flexible than global Fréchet regression and more accurate than local Fréchet regression when the data generating process relies on a non-Euclidean predictor or is truly "global (linear)" for some scalar predictors. In this paper, we provide theoretical support for partially-global Fréchet regression and demonstrate its competitive finite-sample performance when applied to both simulated data and to real data which is too complex for traditional statistical methods.

Keywords: metric distance, partially linear model; random object

Learning nonparametric graphical model on heterogeneous network-linked data

Junhui Wang

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

ABSTRACT

Graphical models have been popularly used for capturing conditional independence structure in multivariate data, which are often built upon independent and identically distributed observations, limiting their applicability to complex datasets such as network-linked data. In this talk, we introduce a nonparametric graphical model that addresses these limitations by accommodating heterogeneous graph structures without imposing any specific distributional assumptions. The introduced estimation method effectively integrates network embedding with nonparametric graphical model estimation. It further transforms the graph learning task into solving a finitedimensional linear equation system by leveraging the properties of vectorvalued reproducing kernel Hilbert space. We will also discuss theoretical properties of the proposed method in terms of the estimation consistency and exact recovery of the heterogeneous graph structures. Its effectiveness is also demonstrated through a variety of simulated examples and a real application to the statistician coauthorship dataset.

Keywords: Graphical model, score matching, network-linked data

Residual-Based Subdata Selection for Local Linear Regression and Its Extension to Partial Linear Models

Chia-Wei Lin[1], Li-Shan Huang[2,*]

Institute of Statistics and Data Science, National Tsing Hua University

ABSTRACT

The rapid growth of data has introduced considerable computational challenges in statistical analysis. His study addresses this issue in local linear regression through representative subdata selection to reduce the computational burden, then extends the method to partial linear models. For local linear egression, a residual-based subdata selection (RESS) method is introduced. RESS yields a lower asymptotic mean squared error than existing methods in a neighborhood where the absolute asymptotic bias is largest. For partial linear models, an integrated method, termed IBRESS, combines ESS for the nonlinear component with information-based optimal subdata selection (IBOSS) for the linear component. IBRESS leverages the strengths of both methods and satisfies two theoretical properties: (i) similar to IBOSS, the convergence rate of the linear component depends on the full data size; and (ii) the nonlinear component retains the asymptotic properties of RESS. Simulation studies demonstrate that IBRESS reduces computational cost while maintaining estimation accuracy.

Keywords: semiparametric regression, big data, data reduction.