Individual Data Protected Integrative Regression Analysis of High-dimensional Heterogeneous Data

Yin Xia

Fudan University
E-mail: xiayin@fudan.edu.cn

Abstract: Evidence based decision making often relies on meta-analyzing multiple studies, which enables more precise estimation and investigation of generalizability. Integrative analysis of multiple heterogeneous studies is, however, highly challenging in the high dimensional setting. The challenge is even more pronounced when the individual level data cannot be shared across studies due to privacy concerns. Under ultra high dimensional sparse regression models and the constraint of not sharing individual data across studies, we propose in this paper a novel integrative estimation procedure by Aggregating and Debiasing Local Estimators (ADeLE). The ADeLE procedure protects individual data through summary-statistics-based integrating procedure, accommodates between study heterogeneity in both the covariate distribution and model parameters, and attains consistent variable selection. Furthermore, the prediction and estimation errors incurred by aggregating derived data is negligible compared to the statistical minimax rate. In addition, the ADeLE estimator is shown to be asymptotically equivalent in prediction and estimation to the ideal estimator obtained by sharing all data. The finite-sample performance of the ADeLE procedure is studied via extensive simulations. We further illustrate the utility of the ADeLE procedure to derive phenotyping algorithms for coronary artery disease using electronic health records data from multiple disease cohorts.