Model selection for correlated data with diverging number of parameters

Hyunkeun Cho and Annie Qu
Department of Statistics, University of Illinois at Urbana-Champaign

Abstract

High-dimensional longitudinal data arises frequently in biomedical and genomic research. It is important to select relevant covariates when the dimension of the parameters diverges as the sample size increases. We propose the penalized quadratic inference function to perform model selection and estimation simultaneously in the framework of a diverging number of regression parameters. The penalized quadratic inference function can easily take correlation information from clustered data into account, yet it does not require specifying the likelihood function. This is advantageous compared to existing model selection methods for discrete data with large cluster size. In addition, the proposed approach enjoys the oracle property, which is able to identify non-zero components consistently with probability tending to 1, and any finite linear combination of the estimated non-zero components also follows an asymptotic normal distribution. We propose an efficient algorithm by selecting an effective tuning parameter to solve the penalized quadratic inference function. Monte Carlo simulation studies illustrate that the proposed method selects the correct model with a high frequency and estimates covariate effects accurately even when the dimension of parameters is high. We illustrate the proposed approach through analyzing periodontal disease data.

Key words and phrases: diverging number of parameters, longitudinal data, model selection, oracle property, quadratic inference function, SCAD.