High-dimensional Heteroscedastic Regression with an Application to eQTL Data Analysis

John Daye*, University of Pennsylvania
Jinbo Chen, University of Pennsylvania
Hongzhe Li, University of Pennsylvania

Abstract

We consider the problem of high-dimensional regression under non-constant error variances. Despite being a common phenomenon in biological applications, heteroscedasticity has, so far, been largely ignored in high-dimensional analysis of genomic data sets. We propose a new methodology that allows non-constant error variances for high-dimensional estimation and model selection. Our method incorporates heteroscedasticity by simultaneously modeling both the mean and variance components via a novel regularized approach. Extensive Monte Carlo simulations indicate that our proposed procedure can result in better estimation and variable selection than existing methods when heteroscedasticity arises from the presence of predictors explaining error variances and outliers. Further, we demonstrate the presence of heteroscedasticity in and apply our method to an expression quantitative trait loci (eQTLs) study. The new procedure can automatically account for heteroscedasticity in identifying the eQTLs that are associated with gene expression variations and lead to smaller prediction errors. These results demonstrate the importance of considering heteroscedasticity in eQTL data analysis.

Keywords: Generalized least squares; Heteroscedasticity; Large p small n; Model selection; Sparse regression; Variance estimation.

* Presenting author