Distributed Sufficient Dimension Reduction for Heterogeneous Massive Data

Liping Zhu

Institute of Statistics and Big Data, Renmin University

Abstract

We propose a distributed sufficient dimension reduction to process massive data characterized by high dimensionality, a huge sample size, and heterogeneity (heterogeneity, and huge sample sizes). To address the high dimensionality, we replace the high-dimensional explanatory variables with a small number of linear projections that are sufficient to explain the variabilities of the response variable. We allow for distinctive function maps for data scattered at different locations, thus addressing the problem of heterogeneity. We assume that the dimension reduction subspaces at different local nodes are identical. This allows us to aggregate the local results obtained from each local node to yield a final esti- mate on a central server. We explicitly examine the sliced inverse regression and cumulative slicing estimation, and investigate the nonasymptotic error bounds of the resulting dimensionality reduction. Our theoretical results are further sup- ported by simulation studies and an application to meta-genome data from the American Gut Project.

Keywords: Cumulative slicing estimation, distributed estimation, heterogeneity, sliced inverse regression, sufficient dimension reduction