Additive partially linear models for massive heterogeneous data

Binhuan Wang

NYU School of Medicine
E-mail: binhuan.wang@nyumc.org

Abstract: We consider an additive partially linear framework for modelling massive heterogeneous data. The major goal is to extract multiple common features simultaneously across all sub-populations while exploring heterogeneity of each sub-population. We propose an aggregation type of estimators for the commonality parameters that possess the asymptotic optimal bounds and the asymptotic distributions as if there were no heterogeneity. This oracle result holds when the number of sub-populations does not grow too fast and the tuning parameters are selected carefully. A plugin estimator for the heterogeneity parameter is further constructed, and shown to possess the asymptotic distribution as if the commonality information were available. Furthermore, we develop a heterogeneity test for the linear components and a homogeneity test for the non-linear components accordingly. The performance of the proposed methods is evaluated via simulation studies and an application to the Medicare Provider Utilization and Payment data.

Keywords and phrases: Divide-and-conquer, homogeneity, heterogeneity, oracle property, regression splines.