DIMENSION REDUCTION STRATEGIES FOR MODELING BI-CLUSTERED HIGH DIMENSIONAL DATA

A thesis presented by

MICHAEL VAN B. SUPRANES

to the

School of Statistics

In partial fulfillment of the requirements for the degree of

Master of Science in Statistics

School of Statistics

University of the Philippines

Diliman, Quezon City

May 2016

Abstract

A three-stage framework is developed for fitting a mixture of regressions for high dimensional data. The method combines a hierarchical agglomerative grouping algorithm, regression-based clustering, and a sequential, group-wise sparse estimation called Layered Elastic Net Selection (LENS). A simulation study is used to compare the method with LASSO-type and PC-based strategies in terms of predictive accuracy, selection optimality, and clustering accuracy. All simulation scenarios are high dimensional (n<<p) with varying correlation structures, and group-wise predictive contributions. When the group of most important predictors varies among regression components, the combination of OLS and the method using LENS outperforms LASSO-type and PC-based strategies in terms of prediction and clustering accuracy. Based on simulation, the method (termed as MixLENS) results to optimal variable selection, and applying OLS on selected variables results to better prediction and clusters. OLS-MixLENS may result to a more interpretable model that is as predictive as a full model (e.g. Mixture of PCRs). In general, MixLENS is likely to select an optimal small subset of predictors for modeling.

Keywords: high dimensionality, mixture of regressions, dimension reduction, variable selection, layered selection, elastic net selection, FMR LASSO