Regularization and Estimation in Regression with Cluster Variables

Qingzhao Yu and Bin Li

March 4, 2013

We propose the Cluster Lasso, a new regularization method for linear regressions. The Cluster Lasso can do variable selections while keeping the correlation structures among variables. In addition, Cluster Lasso encourages selection of clusters of variables, in which variables having the same mechanism in predicting the response variable will be selected in the regression model together. Real microarray data example and simulation studies show that Cluster Lasso outperforms lasso in terms of the prediction performance, particularly when there is collinearity among variables and/or when the number of predictors is larger than the number of observations. The Cluster Lasso paths can be obtained using any established algorithms for lasso solution. An algorithm is proposed to detect variable correlation structures and to compute Cluster Lasso paths efficiently.

Keywords: Clustered Variables, Lasso, Principal Component Analysis.