Robustness of Supervised Clustering Methods to Different Types of Inactive Variables

Rebecca Marion[1], Johannes Lederer[2], Bernadette Govaerts[1], Rainer von Sachs[1]

[1]ISBA/LIDAM, Université catholique de Louvain, Louvain-la-Neuve, Belgium
[2]Department of Mathematics, Ruhr-Universität Bochum, Bochum, Germany

E-mail for correspondence: `rebecca.marion@uclouvain.be`

Abstract: Model regularization methods that perform embedded variable selection during the model estimation process have great potential for increasing model interpretability. In cases where the predictor variables form clusters (such as in gene expression data, where genes belong to different regulatory pathways), the selection of important variables using model regularization is more challenging. This is especially true when the variable clusters are not known a priori. "Supervised clustering" methods in the literature, such as Pairwise Absolute Clustering and Sparsity (Sharma (2013)), Simultaneous Supervised Clustering and Feature Selection (Shen (2012)) and Cluster Elastic Net (Witten (2014)) are able to learn variable clusters from the data and select important clusters during model estimation. However, the correlation structure of inactive variables (i.e. variables that do not predict the response) can impact the variable selection, clustering and prediction quality of these methods. Inactive variables come in several types: they can be correlated or uncorrelated with each other, as well as correlated or uncorrelated with active variables. This poster presents a simulation study comparing state-of-the-art supervised clustering methods and demonstrating the impact of the correlation structure of inactive variables on prediction and clustering performance.

Key words: Variable clustering; Variable selection; Regression; Regularization

Sharma D, Bondell HD and Zhang H (2013). Consistent group identification and variable selection in regression with correlated predictors. Journal of Computational and Graphical Statistics: 22(2), 319–340.

Shen X, Huang HC and Pan W (2012). Simultaneous supervised clustering and feature selection over a graph. Biometrika: 99(4), 899– 914.

Witten DM, Shojaie A and Zhang F (2014). The cluster elastic net for high-dimensional regression with unknown variable grouping. Technometrics: 56(1), 112–122.