Weighted LAD-LASSO for Robust Variable Selection for Grouped Data

Kristin Lilly [1] Nedret Billor [2]

[1]Columbus State University
Columbus, Georgia, USA
lilly_kristin@columbusstate.edu

[2]Auburn University
Auburn, Alabama, USA
billone@auburn.edu

October 3, 2019

Abstract

Group variable selection is an interesting new problem, whereby the predictors in a multiple linear regression setting are non-arbitrarily grouped and selecting a subset of important groups of variable is of interest. In this poster, two new methods are proposed, the group WLAD-LASSO and the adaptive group WLAD-LASSO for selecting significant groups of variables when outliers are present in both the response and the predictor matrix. The theoretical properties of each of the methods is discussed. Preliminary results from a simulation study will be presented.

Motivation

- Description: Data from microarray experiments of mammalian eye tissue samples (Scheetz, 2006)
- Response: The expression level of gene TRIM32, which causes Bardet-Biedl syndrome, of 120 rats
- Predictors: 100 predictors, which are the expression levels of 20 genes, which were expanded using 5 basis B-splines
- Goal: Identify genes that are instrumental in predicting the expression level of TRIM32

Goals

- **Setup**: One response as a function of several predictors
- **New assumption**: Predictors are organized into non-arbitrary groups
- **Problem**: Want to build a model in order to predict the response the best using only a subset of groups; i.e., perform group variable selection
- **Added Consideration**: Outliers in the data
- **Goal**: Create a group variable selection method that performs well in the presence of outliers!

Method 1: Group WLAD-LASSO

- The criterion for minimization for the group WLAD-LASSO method is the following:

$$Q(\boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{1}{2} w_i |y_i - \sum_{k=1}^{K} \boldsymbol{x}_{ik} \boldsymbol{\beta}_k| + n\lambda \sum_{i=1}^{K} ||\boldsymbol{\beta}_k||_1 \qquad (1)$$

  where w_i is a positive weight assigned to each observation for $i = 1, \ldots, n$ and $k = 1, \ldots, K$ is the number of groups.
- Outliers are designed to be downweighted proportionally to a calculated robust distance, such that points farther away from the center of the corresponding distribution of a predictor variables are downweighted more.

Method 2: Adaptive Group WLAD-LASSO

- With an adaptive tuning parameter, we get the following objective function to minimize:

$$Q(\beta) = \sum_{i=1}^{n} \frac{1}{2} w_i |y_i - \sum_{k=1}^{K} \boldsymbol{x}_{ik} \beta_k| + n \sum_{i=1}^{K} \lambda_k ||\beta_k||_1 \qquad (2)$$

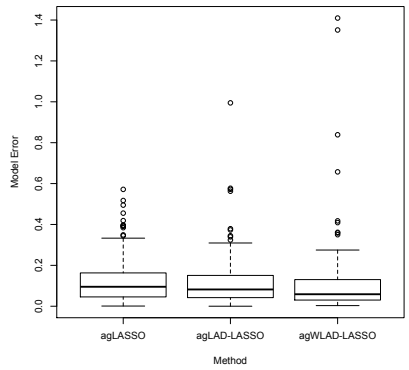
- This results in regression estimators that will be robust to outliers in the response and in the x-direction, while enjoying the shrinkage and nice theoretical properties of the adaptive LASSO to perform group selection.

Results - Table

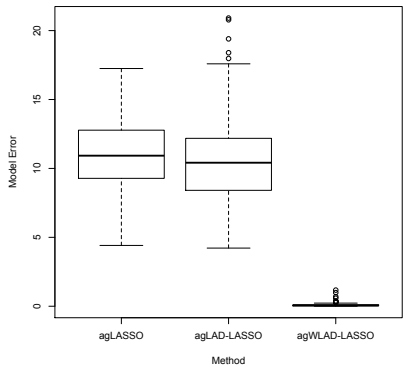
Table: Simulation results for t_3 error for X- and Y-outliers

σ	n	Contamination (%)	Method	Mean Model Error	Median Model Error
1	100	0	ag LASSO	0.11	0.09
			ag LAD-LASSO	0.10	0.08
			ag WLAD-LASSO	0.07	0.04
		0.1	ag LASSO	11.53	11.21
			ag LAD-LASSO	10.51	10.06
			ag WLAD-LASSO	0.12	0.06
		0.2	ag LASSO	29.22	28.50
			ag LAD-LASSO	29.12	29.23
			ag WLAD-LASSO	0.18	0.13
		0.3	ag LASSO	52.84	52.40
			ag LAD-LASSO	51.82	50.53
			ag WLAD-LASSO	0.30	0.15

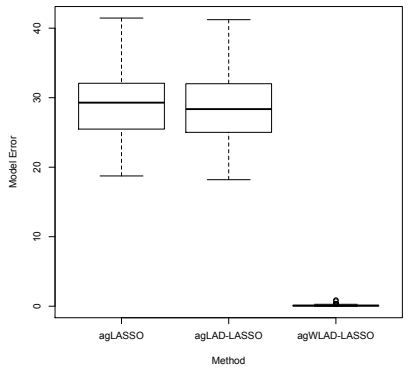
Results - Plots w/ Contamination % for X- and Y- Outliers



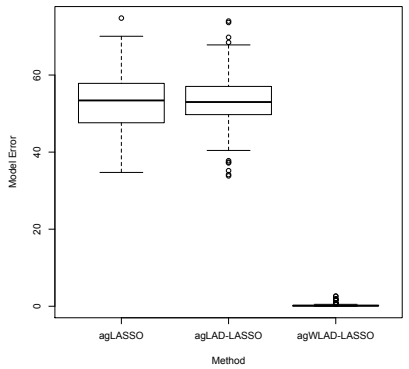
(a) 0%



(b) 10%



(c) 20%



(d) 30%

Conclusions & Future Work

- The Group LAD-LASSO performs well only with outliers in the y-direction.
- The Adaptive Group WLAD-LASSO performs well with outliers in any direction and has nice theoretical properties including: estimation consistency, selection consistency (sparsity), and the oracle property.
- Next steps include a high-dimensional simulation study and a real data application.

References

- Arslan, O. (2012), "Weighted LAD-LASSO method for robust parameter estimation and variable selection in regression," Computational Statistics & Data Analysis, 56, 1952-1965.
- Scheetz, T., Kim, K., Swiderski, R., Philp, A., Braun, T., Knudtson, K., Dorrance, A., DiBona, G., Huang, J., Casavant, T. et al. (2006), "Regulation of gene expression in the mammalian eye and its relevance to eye disease," Proceedings of the National Academy of Sciences, 103 (39), 14429-14434
- Yuan, M. and Lin. Y. (2006), "Model selection and estimation in regression with grouped variables," Journal of the Royal Statistical Society, Series B, 68, 49-67.