A robust variable selection method for grouped data

Kristin Lilly [*] Nedret Billor[†]

Abstract

Group variable selection is a relatively new problem in statistics. When predictor variables can be naturally grouped in the multiple linear regression setting, the objective is to perform variable selection at the group and within-group levels. Several methods have been proposed to perform this type of variable selection, most of which are adapted from existing methods, including the group lasso. However, these methods do not perform optimally in the presence of outliers. As a result, a robust form of the group lasso is presented that is well suited to data with outliers, while still executing group variable selection. Examples with simulated data are shown to assess the performance of this newly proposed method versus existing methods when outliers are present.

Keywords: **Group lasso, Robust variable selection, Multiple regression**

1 Introduction

Suppose the multiple linear regression model is given as:

$$\boldsymbol{y} = \boldsymbol{X}\beta + \varepsilon \tag{1}$$

where $\boldsymbol{y}$ is an n x 1 vector of responses, $\boldsymbol{X}$ is an n x p matrix of predictors, β is a p x 1 vector of regression coefficients, and ε is an n x 1 vector of random errors.

In the multiple linear regression setting, selecting a meaningful subset of predictor variables, known as variable selection, is an important problem, especially with a large number of predictors. An interesting new problem in statistics is group variable selection, where the predictor variables can be naturally grouped, and important groups of variables are to be selected. This type of data is common in many scientific applications. Examples include fMRI data with grouped gene expressions or demographic data that can be grouped by socioeconomic or physical factors. In such cases, it is common to have outliers in the data and multicollinearity between the predictor variables.

[*] Department of Mathematics & Statistics, Auburn University, 221 Parker Hall, Auburn University, Alabama 36849, E-mail: seamokl@auburn.edu

[†] Department of Mathematics & Statistics, Auburn University, 221 Parker Hall, Auburn University, Alabama 36849, E-mail: billone@auburn.edu

Thus, it is necessary to develop a method to do well in the presence of outliers and with high correlation between predictors.

With ideal data, least squares estimators (LSE) for the regression coefficients β_j are usually found to numerically describe the model. The assumptions for the LSE include an approximate linear relationship between the response and predictor variables, and uncorrelated, normally distributed error terms with mean 0 and constant variance σ^2. As a result, the LSE are sensitive to outliers, leading to estimators with high bias in the presence of observations that deviate from a majority of the data points. Traditional variable selection methods, such as forward selection, backwards elimination, and stepwise regression, are based on the LSE; consequently, these methods are sensitive to outliers and also lead to unstable models, which would cause poor prediction results. Thus, a robust method must be used in order to build more accurate linear models to use for prediction or estimation purposes.

Some modern approaches such as LASSO (least absolute shrinkage and selection operator) [1] have been proposed to obtain stable models with good prediction. In this study, we will give an evaluation of a basic “group” variable selection method based on LASSO estimator. Furthermore, a robust group variable selection method, derived from the LAD (least absolute deviation)-LASSO method, is proposed, and an example with real data is presented to demonstrate the performance of the group LASSO versus the group LAD-LASSO methods in the presence of outliers.

2 Group LASSO

For the group LASSO method, assume the predictor variables can be naturally grouped into k groups for $k = 1, \ldots, K$, where each group consists of p_k predictor variables such that $\sum_{k=1}^{K} p_k = p$. The predictor variables should be standardized so that each x_{ij} has mean 0 and variance 1 for $j = 1, \ldots, p$. The criterion to be minimized is:

$$\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p}(\sum_{k=1}^{K} x_{ijk}\beta_{jk}))^2 + \lambda \sum_{j=1}^{p}\sum_{k=1}^{K} |\beta_{jk}| \tag{2}$$

where $\lambda \geq 0$ is a tuning parameter [2]. That is, for each group of predictors, minimize the sum of the squared distances, while simultaneously shrinking unimportant groups with the LASSO penalty (L_1 norm). The tuning parameter λ controls the rate of shrinkage and can be chosen using cross-validation.

The LASSO method of simultaneous estimation and selection is ideal for multicollinearity, but not for data with outliers. In particular, because it uses the LSE, the group LASSO performs poorly in terms of robustness.

3 Group LAD-LASSO

To combat the problem of outliers in the response during group variable selection, we propose a modification to the LAD-LASSO [3] that minimizes the following criterion:

$$\sum_{i=1}^{n} |y_i - \sum_{j=1}^{p}(\sum_{k=1}^{K} x_{ijk}\beta_{jk})| + \lambda \sum_{j=1}^{p}\sum_{k=1}^{K} |\beta_{jk}| \tag{3}$$

where $\lambda \geq 0$ is a tuning parameter that controls the shrinkage of the estimators, just like in the group LASSO. Instead of minimizing the sum of the squared distances, this method minimizes the least absolute deviation between the response and each group of predictors, while the second part of the equation will estimate the regression coefficients for each group or shrink them to 0.

The LAD-LASSO method is optimal for data with deviations in the y-direction and highly correlated data, because of the least absolute deviation minimization (good for y-direction outliers) and the LASSO method (good for multicollinearity).

3.1 Computation of the Group LAD-LASSO

The computation of the group LAD-LASSO is based on the shooting algorithm [4]. Originally, this method was proposed for the LASSO method, but was adapted for the group LASSO [2]. Some slight modifications make it an appropriate computational method for the group LAD-LASSO.

Rewrite (3):

$$|Y - \sum_{j=1}^{p}\sum_{k=1}^{K} X_{jk}\beta_{jk}| + \lambda \sum_{j=1}^{p}\sum_{k=1}^{K} |\beta_{jk}| \tag{4}$$

Next, rewrite (4) with respect to the groups:

$$|Y - \sum_{k=1}^{K} X_k\beta_k| + \lambda \sum_{k=1}^{K} |\beta_k| \tag{5}$$

where $Y \sim$ n x 1 vector of responses, $X_k \sim$ n x p_k matrix of predictors from group k, $\beta_k \sim p_k$ x 1 vector of regression coefficients for group k, and $\lambda \geq 0$ is a tuning parameter.

Then, the algorithm for the group LAD-LASSO involves applying the following equation iteratively with the groups for $k = 1, \ldots, K$:

$$\beta_k = \left(1 - \frac{\lambda\sqrt{p_k}}{\|S_k\|}\right)_+ S_k \tag{6}$$

where $S_k = X_k^T|Y - X\beta_{-k}|$ with $\beta_{-k} = (\beta_1^T, \ldots, \beta_{k-1}^T, \mathbf{0}^T, \beta_{k+1}^T, \ldots, \beta_K^T)$, the β vector without coefficient β_k, and $\|\eta\| = (\eta^T\eta)^{1/2}$. Choose initial β_k for $k = 1, \ldots, K$ to be the LAD estimators. This algorithm is stable and reaches convergence tolerance within a few iterations; on the other hand, the computational burden increases dramatically as the number of predictors increases [2].

4 Group WLAD-LASSO

The group WLAD-LASSO is a simultaneous estimation and group variable selection method robust to outliers in both the x- and y-directions. This method is being developed by combining the group LAD-LASSO method with the WLAD-LASSO [5]. The group WLAD-LASSO criterion to be minimized is:

$$\sum_{i=1}^{n} w_i |y_i - \sum_{j=1}^{p} (\sum_{k=1}^{K} x_{ijk}\beta_{jk})| + \lambda \sum_{j=1}^{p}\sum_{k=1}^{K} |\beta_{jk}| \tag{7}$$

where $\lambda \geq 0$ is a tuning parameter. Define the robust distance $RD(\boldsymbol{x}_i) = (\boldsymbol{x}_i - \hat{\mu})^T \hat{\Sigma}^{-1}(\boldsymbol{x}_i - \hat{\mu})$ for $i = 1, \ldots, n$, where $\hat{\mu}$ and $\hat{\Sigma}$ are robust location and scale measures. Large values of $RD(\boldsymbol{x_i})$ indicate leverage points. Then, calculate the positive weights such that $w_i = min\{1, \frac{p}{RD(\boldsymbol{x}_i)}\}$ for $i = 1, \ldots, n$.

The least absolute deviation restriction from (7) helps to minimize the effects of the y-direction outliers when fitting the linear model, while the weights counteract the x-direction outliers. That is, as $RD(\boldsymbol{x}_i)$ gets larger, w_i gets smaller, giving high leverage points smaller weights in the model. While the group WLAD-LASSO and group LAD-LASSO perform almost equivalently for data with outliers in the response, the group WLAD-LASSO does best out of the two with outliers in both directions, as well as with multicollinearity between the predictors.

5 Simulation Study

A small simulation study is performed to compare the group LASSO with the group LAD-LASSO and group WLAD-LASSO. For sample sizes n=50,100, and 200, let ϵ be the contamination rate equal to values ϵ=0.1,0.2,0.3, and 0.4 such that $m = [\epsilon n]$ is the number of contaminated data points. The first $n - m$ data points are generated from the true model $\boldsymbol{y_1} = \boldsymbol{X_1}\beta_1 + \sigma\varepsilon$, where $\boldsymbol{X}$ is multivariate normal with $\boldsymbol{0}$ mean and the pairwise correlation between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ equal to $cor(\boldsymbol{x}_i, \boldsymbol{x}_j) = 0.5^{|i-j|}$. The regression parameter vector is set to be $\beta_1 = (3, 1.5, 2, 0, 0, 0)$, such that there are two groups of three variables. The errors ε are generated from the standard normal distribution, the t-distribution with 3 degrees of freedom, and the t-distribution with 5 degrees of freedom, while σ will be 0.5 and 1. This will allow for heavy-tail error distributions and some outliers in the response direction. The m points from the contaminated data are produced with the following model: $\boldsymbol{y_2} = \boldsymbol{X_2}\beta_2$, where $\boldsymbol{X}_2$ is multivariate normally distributed with $\boldsymbol{\mu}_2 \neq \boldsymbol{0}$ and covariance equal to $\boldsymbol{I}$. Let $\beta_2 \neq \beta_1$. For each combination of sample size, contamination rate, sigma, and error distribution, the simulation is performed 200 times, and the relative prediction error (RPE) will be calculated from the group LASSO, group LAD-LASSO, and group WLAD-LASSO fits for comparison purposes.

The *grpreg* package and *rrcov* package were both utilized to perform the simulation in R. The results are shown in Tables 1-3. In each case, with the exception of $\epsilon = 0$, the group WLAD-LASSO method results in the smallest RPE. Results are similar for $n = 100, 200$.

Table 1: Simulation results for standard normally distributed errors.

n	σ	ϵ	Method	MeanRPE	MedianRPE
50	0.5	0	Group LASSO	0.07	0.06
			Group LAD-LASSO	0.07	0.06
			Group WLAD-LASSO	0.27	0.18
		0.1	Group LASSO	10.89	10.25
			Group LAD-LASSO	6.11	5.96
			Group WLAD-LASSO	0.29	0.20
		0.2	Group LASSO	16.57	16.50
			Group LAD-LASSO	9.20	9.01
			Group WLAD-LASSO	0.15	0.10
		0.3	Group LASSO	21.08	21.08
			Group LAD-LASSO	11.47	11.28
			Group WLAD-LASSO	0.30	0.24
		0.4	Group LASSO	24.40	23.71
			Group LAD-LASSO	12.93	12.69
			Group WLAD-LASSO	0.69	0.64
50	1	0	Group LASSO	0.27	0.24
			Group LAD-LASSO	0.24	0.21
			Group WLAD-LASSO	0.90	0.79
		0.1	Group LASSO	16.68	16.28
			Group LAD-LASSO	10.94	10.75
			Group WLAD-LASSO	0.16	0.14
		0.2	Group LASSO	16.28	15.92
			Group LAD-LASSO	9.39	9.11
			Group WLAD-LASSO	0.26	0.22
		0.3	Group LASSO	20.96	20.52
			Group LAD-LASSO	11.82	11.33
			Group WLAD-LASSO	0.43	0.39
		0.4	Group LASSO	23.30	23.10
			Group LAD-LASSO	13.63	13.57
			Group WLAD-LASSO	0.78	0.75

Table 2: Simulation results for t_3 errors.

n	σ	ϵ	Method	MeanRPE	MedianRPE
50	0.5	0	Group LASSO	0.18	0.13
			Group LAD-LASSO	0.19	0.15
			Group WLAD-LASSO	0.82	0.43
		0.1	Group LASSO	14.83	14.55
			Group LAD-LASSO	10.44	9.09
			Group WLAD-LASSO	0.56	0.57
		0.2	Group LASSO	15.92	15.91
			Group LAD-LASSO	9.44	9.13
			Group WLAD-LASSO	0.23	0.16
		0.3	Group LASSO	20.88	20.89
			Group LAD-LASSO	11.65	11.49
			Group WLAD-LASSO	0.40	0.36
		0.4	Group LASSO	23.53	23.37
			Group LAD-LASSO	13.62	13.39
			Group WLAD-LASSO	0.73	0.68
50	1	0	Group LASSO	0.27	0.25
			Group LAD-LASSO	0.65	0.46
			Group WLAD-LASSO	0.72	0.66
		0.1	Group LASSO	17.28	16.86
			Group LAD-LASSO	11.33	11.23
			Group WLAD-LASSO	0.37	0.28
		0.2	Group LASSO	16.48	16.20
			Group LAD-LASSO	11.59	9.20
			Group WLAD-LASSO	0.48	0.38
		0.3	Group LASSO	21.17	21.16
			Group LAD-LASSO	11.82	11.33
			Group WLAD-LASSO	0.73	0.60
		0.4	Group LASSO	23.67	23.43
			Group LAD-LASSO	13.62	13.39
			Group WLAD-LASSO	0.71	0.62

Table 3: Simulation results for t_5 errors.

n	σ	ϵ	Method	MeanRPE	MedianRPE
50	0.5	0	Group LASSO	0.09	0.07
			Group LAD-LASSO	0.11	0.10
			Group WLAD-LASSO	0.42	0.28
		0.1	Group LASSO	4.74	4.46
			Group LAD-LASSO	2.06	2.01
			Group WLAD-LASSO	0.44	0.28
		0.2	Group LASSO	16.58	16.16
			Group LAD-LASSO	9.51	9.31
			Group WLAD-LASSO	0.86	0.46
		0.3	Group LASSO	23.38	23.17
			Group LAD-LASSO	13.48	13.36
			Group WLAD-LASSO	0.35	0.28
		0.4	Group LASSO	23.38	23.71
			Group LAD-LASSO	12.93	12.69
			Group WLAD-LASSO	0.71	0.65
50	1	0	Group LASSO	0.41	0.33
			Group LAD-LASSO	0.49	0.42
			Group WLAD-LASSO	0.82	0.81
		0.1	Group LASSO	16.30	15.91
			Group LAD-LASSO	10.71	10.36
			Group WLAD-LASSO	0.24	0.21
		0.2	Group LASSO	16.44	16.40
			Group LAD-LASSO	9.15	9.02
			Group WLAD-LASSO	0.32	0.28
		0.3	Group LASSO	21.15	21.12
			Group LAD-LASSO	11.81	11.42
			Group WLAD-LASSO	0.62	0.52
		0.4	Group LASSO	23.47	23.31
			Group LAD-LASSO	13.30	12.96
			Group WLAD-LASSO	0.86	0.80

6 Conclusion

Three group variable selection methods were discussed, the group LASSO, the group LAD-LASSO, and the group WLAD-LASSO with the intention of selecting important groups of predictor variables and estimating the regression coefficients of the groups. While the group LASSO is useful for highly correlated data, it does not work well for data with outliers, in general. The proposed group LAD-LASSO does work well for outliers in the y-direction, but has some problems with outliers in the x-direction. The group WLAD-LASSO method is designed to work well for outliers in both directions. The example highlighted the differences in the two methods with available code for a real data set with outliers.

A more in-depth examination of the group LASSO, group LAD-LASSO, and group WLAD-LASSO with a simulation study, as well as comparing the two methods analytically, is recommended. Then, applying other robust methods to the group variable selection problem to see how they compare to the methods discussed here is a natural next step. Ideally, a method robust to all types of outliers while estimating and selecting groups of variables in the group variable selection setting can be derived from existing robust methods not sensitive to outliers.

References

[1] R. Tibshirani,“Regression Shrinkage and Selection via the Lasso”, *Journal of the Royal Statistical Society. Series B (Methodological)*, Volume 58, Issue 1 (1996), 267-288.

[2] M. Yuan and Y. Lin, “Model selection and estimation in regression with grouped variables”, *Journal of the Royal Statistical Society. Series B (Methodological)* Volume 68, Issue 1 (2006), 49-67.

[3] H. Wang, G. Li, G. Jiang, “Robust Regression Shrinkage and Consistent Variable Selection Through the LAD-Lasso”, *Journal of Business & Economic Statistics*, Volume 25, Issue 3 (2007) , 347-355.

[4] W. J. Fu, “Penalized Regressions: The Bridge Versus the Lasso”, *Journal of Computational and Graphical Statistics*, Volume 7, Issue 3 (1999) , 397-416.

[5] O. Arslan, “Weighted LAD-LASSO method for robust parameter estimation and variable selection in regression”, *Computational Statistics and Data Analysis* Volume 56 (2012), 1952-1965.