Robust Group LASSO Methods

Kristin Lilly * Nedret Billor†

Abstract

Group variable selection is a relatively new problem in statistics. When the predictors can be naturally grouped in regression analysis, it is important to select important groups of variables that are influencing the response. One method of performing group variable selection is a method based on the least absolute shrinkage and selection operator (LASSO), which is called the group LASSO. This method works well in most cases, but has issues when there are outliers in the response. This paper proposes two methods which are based on the least absolute deviation (LAD), the group LAD-LASSO and the adaptive group LAD-LASSO, to perform group variable selection, in the presence of outliers in the response. Both methods perform well when there are outliers in the y-direction; however, only the adaptive version has nice theoretical properties, including the oracle property. Further, selection of the shrinkage parameter and those properties are discussed. Simulation studies and an application to a real data set are also presented for both methods.

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

1 Introduction

Regression analysis, in general, encounters two major problems. The first involves the estimation of the regression coefficients. In real-life situations involving regression analysis, datasets are not always optimal. There may be heavy-tailed errors or outliers in the response or in the predictor variables, and as a result, methods used for estimation in regression need to be adapted for these situations, which are called robust regression methods. For this paper, the focus is on one of these two scenarios: outliers in the response. In this case, the ordinary least squares (OLS) estimators can perform poorly, such that the estimators are unstable and inconsistent. A suitable robust method for this case has been proposed, the least absolute deviation (LAD) estimators, which work well when there are outliers in the response.

In addition to estimating the regression coefficients, variable selection is another important problem in regression analysis. Typically, regression analysis begins with modeling one

*Department of Mathematics, Columbus State University, 211 University Hall, Columbus, Georgia 31907, E-mail: lilly_kristin@columbusstate.edu

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

response with several predictor variables. It is beneficial to select a subset of explanatory variables to predict the response. Including too many predictors in a regression model will result in a model that is inefficient. Predictions from the given model will also be inaccurate. Including too few predictors in a model will lead to a biased model and predictions. As a result, several variable selection methods have been proposed to aid in selecting the true underlying model in a regression problem. Stepwise regression is one of those such methods, which includes forward selection, backwards elimination, and bidirection elimination. Other criteria have been proposed for variable selection including the Akaike Information Criterion (AIC), the Bayesian Information Criterion (BIC), and Mallows's C_p.

One can simultaneously estimate significant regression coefficients and shrink insignificant predictors to zero using the least absolute shrinkage operator (LASSO) method (Tibshirani, 1996). The LASSO is convenient in this respect; however, the LASSO does not perform well in the presence of outliers. Some robust versions of the LASSO have been briefly explored and mentioned. Along the same lines, the LAD-LASSO (Wang et al., 2007) method has been derived to take advantage of the simultaneous estimation and selection of regression coefficients while also doing well with outliers in the response. The LAD-LASSO method combines the LAD regression penalty with the LASSO restriction on the regression coefficients.

Another recent problem of interest involves grouped predictors. That is, the predictors can be naturally grouped in such a way that is inherent to the structure of the data. For example, markers on genes can be grouped such that there are 5 markers per gene. Grouped predictors are also a part of function MRI (fMRI) data, as well as survey data, where variables can be grouped by demographic factors. A few group variable selection methods have been proposed, including the group LASSO (Yuan and Lin, 2006). The group LASSO will estimate the regression coefficients of important groups of variables and shrink the insignificant groups all to zero simultaneously. With normally distributed errors, the group LASSO does well in distinguishing the true model; however, when there are outliers in any direction, it often leads to incorrect models and bad prediction results. As a result, there is a need for methods to execute group variable selection robustly when there are outliers in the data.

The rest of the paper is organized as follows. Section 2 proposes the group LAD-LASSO method and discusses its computation and tuning parameter selection. Section 3 proposes the adaptive group LAD-LASSO and discusses its computation and tuning parameter selection; its properties are also mentioned. Section 4 presents a simulation study demonstrating the robustness of the methods in the presence of outliers. Section 5 applies the methods to a real data set. Section 6 concludes the paper.

2 Group LAD-LASSO

Consider the multiple linear regression model

$$y_i = \beta_0 + \boldsymbol{x}_i^T \boldsymbol{\beta} + \varepsilon_i \tag{1}$$

for $i = 1, \ldots, n$, where y_i are the responses, $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^T$ is the $1 \times p$ vector of predictors, $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^T$ are the regression coefficients, and ε_i are the iid random

errors. Without loss of generality, assume that $\beta_0 = 0$. This can be done practically by centering both the predictors and the response. As a result, consider the linear regression model:

$$y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + \varepsilon_i \tag{2}$$

for $i = 1, \ldots, n$. The ordinary least squares (OLS) criterion for estimating the regression parameter vector $\boldsymbol{\beta}$ requires the minimization of the following objective function:

$$Q(\boldsymbol{\beta}) = \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i^T \boldsymbol{\beta})^2. \tag{3}$$

The solution results in the estimator $\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T Y$, where X is the $n \times p$ design matrix of explanatory variables such that $\boldsymbol{x}_i^T$ is the ith row with rank p and $Y = (y_1, \ldots, y_n)^T$ is the $n \times 1$ vector of responses. The OLS method relies on the assumptions that the errors ε_i are random normal errors with mean 0 and constant variance σ^2. Thus, when the data include outliers, it is known that the OLS estimators perform poorly.

Least absolute deviation (LAD) regression is a method well suited as an alternative to the OLS method when there are outliers in the response. The only assumption the LAD method requires is that the random errors ε have median 0. Consider the LAD regression minimization criterion:

$$Q(\boldsymbol{\beta}) = \sum_{i=1}^{n} |y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}|. \tag{4}$$

In order to also perform variable selection while estimating the regression coefficients, the least absolute shrinkage and selection operator (LASSO) (Tibshirani, 1996) has been proposed. The LASSO minimizes the least squares equation, while including a penalty on the sum of the absolute value of the regression coefficients. The LASSO criterion to be minimized is the following:

$$Q(\boldsymbol{\beta}) = \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i^T \boldsymbol{\beta})^2 + n\lambda \sum_{j=1}^{p} |\beta_j| \tag{5}$$

for $i = 1, \ldots, n$ and $j = 1, \ldots, p$ and where $\lambda > 0$ is a tuning parameter, which controls the amount of shrinkage applied to the regression coefficients. When the assumptions of the OLS method are fulfilled, the LASSO performs optimally.

A method for estimation and variable selection is the LAD-LASSO (Wang et al., 2007), which combines the robustness of the LAD method for estimation and the shrinkage of the LASSO penalty for variable selection. The criterion to be minimized for the LAD-LASSO is the following:

$$Q(\boldsymbol{\beta}) = \sum_{i=1}^{n} |y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}| + n\lambda \sum_{j=1}^{p} |\beta_j| \tag{6}$$

for $i = 1, \ldots, n$ and $j = 1, \ldots, p$ and where $\lambda > 0$ is a tuning parameter. The LAD-LASSO will help to minimize the effect of the outliers in the response, but not the outliers in the explanatory variables. In this paper, we want to adapt this same idea to a multiple regression model with grouped predictors. This requires some extra notation. Consider the linear regression model:

$$y_i = \sum_{j=1}^{p} x_{ij}\beta_j + \varepsilon_i \tag{7}$$

for $i = 1, \ldots, n$ and $j = 1, \ldots, p$, where y_i are the responses, x_{ij} are the values for the individual predictor variables, β_j's are the regression coefficients, and ε_i are the iid random errors with mean 0 and constant variance. Further assume that the predictor variables are grouped such that there are K groups for $k = 1, \ldots, K$ and each group k has p_k predictors where $\sum_{i=1}^{k} p_k = p$.

Therefore, in general, the linear regression model for grouped predictors may be written as the following:

$$y_i = \sum_{k=1}^{K} \boldsymbol{x}_{ik}\boldsymbol{\beta}_k + \varepsilon_i \tag{8}$$

where $\mathbf{x}_{ik}$ is an $1 \times p_k$ vector of predictors in group k and $\boldsymbol{\beta}_k$ is a $p_k \times 1$ vector of regression coefficients for group k for $i = 1, \ldots, n$.

The regression coefficients in the typical linear regression setting are usually estimated with the OLS criterion, minimizing the sum of squares of the residuals. This can also be done for the case of grouped variables, where the sum of the squares of the residuals are still minimized, but with respect to the groupings. An alternative is the group LASSO (Yuan and Lin, 2006), which adds the L_2 penalty to the minimization criterion of the OLS method:

$$Q(\boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{1}{2}(y_i - \sum_{k=1}^{K} \boldsymbol{x}_{ik}\boldsymbol{\beta}_k)^2 + n\lambda \sum_{k=1}^{K} ||\boldsymbol{\beta}_k||_2. \tag{9}$$

The group LASSO will identify important groups and estimate their regression coefficients while shrinking unimportant groups to 0. It is known that the LASSO estimates can be sensitive to outliers, because of the dependency of (9) on the OLS criterion. The algorithm for the group LASSO is based off of the "shooting algorithm" proposed by Yuan and Lin (2006), and it is described as a "group descent" algorithm by Breheny (2015). Its implementation is defined to be the same as that for coordinate descent algorithms (Friedman et al., 2007; Wu and Lange, 2008), but requires a modification to minimize the criterion in (9) with respect to the groups. This algorithm requires a few basic arithmetic operations, which makes for a computationally efficient algorithm (Breheny, 2015).

In the case of outliers in the response, the LAD estimators can relieve some of this sensitivity, in addition to using the LASSO penalty for shrinkage and selection. Hence, the combination of the LAD-LASSO method with grouped predictors to obtain:

$$Q(\boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{1}{2}|y_i - \sum_{k=1}^{K} \boldsymbol{x}_{ik}\boldsymbol{\beta}_k| + n\lambda \sum_{k=1}^{K} ||\boldsymbol{\beta}_k||_2 \tag{10}$$

which is the minimization criterion for the group LAD-LASSO to simultaneously estimate significant groups and shrink nonsignificant groups to 0. Note that if $p_k = 1$ for all k, then (10) reduces to the LAD-LASSO equation (6). This method is implement in R with a small modification to the loss function in the $grpreg$ package in R.

The tuning parameter, λ, should be chosen so that it is large enough that there is a desired shrinking effect for insignificant groups, but it should also be chosen such that it is small enough that all the group are not shrunk to 0. In general, Cross-validation and generalized cross-validation methods can be used to find the optimal value of the tuning parameter λ (Tibshirani, 1996; Fan and Li, 2001). In this case, we use k-fold cross-validation after modifying the objective function to be that of the group LAD-LASSO to find the best value of λ, such that the cross-validation error is minimized.

Unfortunately, because of using one tuning parameter, λ, to control all shrinkage, the properties of consistency, sparsity, and the oracle property do not hold for the group LAD-LASSO (Wang et al., 2007).

3 Adaptive Group LAD-LASSO

The adaptive LASSO is an extension of the LASSO which penalizes each coefficient differently with its own tuning parameter instead of penalizing each coefficient equally with the L_1 penalty that may not necessarily be the best way to treat the predictors when they don't all contribute to the regression model. (Zou, 2006).

The adaptive LASSO is designed to minimize the following equation:

$$\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} w_j|\beta_j| \tag{11}$$

where the weights are defined to be w_j to be $w_j = \frac{1}{|\hat{\beta}_j|^\nu}$, where $\hat{\beta}_j$ is the LSE for the jth parameter and $\nu > 0$. Equivalently, (11) can be written as:

$$\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \sum_{j=1}^{p} \lambda_j|\beta_j|. \tag{12}$$

The solution to (11) is based on a convex optimization problem. Algorithms used to solve for the LASSO solutions can be used to compute the adaptive LASSO solutions with a very simple modification. The tuning parameter λ_j for each regression coefficient is found using cross-validation along with the LARS algorithm, similar to how it is found for the LASSO. The oracle properties, including consistency and sparsity, hold for the adaptive LASSO method (Zou, 2006).

Wang and Leng (2008) proposed the adaptive group LASSO which assigns a different tuning parameter for each group, allowing the shrinkage to vary from group to group and

also showed that this method has model selection consistency and is efficient. The adaptive group LASSO criterion to minimize is the following:

$$\frac{1}{2}\sum_{i=1}^{n}(y_i - \sum_{k=1}^{K}\boldsymbol{x}_{ik}\boldsymbol{\beta}_k)^2 + n\sum_{k=1}^{K}\lambda_k\|\boldsymbol{\beta}_k\|_2 \tag{13}$$

where $\lambda_k \geq 0$ is an adaptive tuning parameter, y_i is the ith response, $\boldsymbol{x}_{ik}$ is a $1 \times p_k$ vector of predictors in the kth group for the ith observation, and $\boldsymbol{\beta}_k$ is a $p_k \times 1$ vector of regression coefficients for group k. The flexible tuning parameter applies varying amounts of shrinkage to the different groups of predictors. As a result, it can be understood intuitively that applying a high amount of shrinkage to insignificant groups, which would go to 0, and applying a low amount of shrinkage to significant groups, which would be nonzero, would result in an efficient estimator. Even if there is no prior information on which groups are significant and which are not, the shrinkage parameter can be chosen in such as way to get as efficient an estimator as possible.

To choose an appropriate tuning parameter λ_k, usually, cross-validation (CV) or generalized cross-validation (GCV) is used. However, these methods can be too computationally intensive for the adaptive group LASSO, because of the possible high number of tuning parameters that need to be estimated. An ideal candidate for the tuning parameter, according to Wang and Leng (2008) is:

$$\lambda_k = \frac{\lambda}{||\hat{\boldsymbol{\beta}}_k||_2^{\gamma}} \tag{14}$$

where $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p)^T$ is the LSE and $\gamma > 0$ is a prespecified positive number. For their simulation study and real data example, the authors chose $\gamma = 1$. With this choice of shrinkage parameter for each group, the problem of finding an optimal shrinkage parameter reduces to a univariate problem to solve for λ, which can be found similarly as in the case of the LASSO based on various criteria, including C_p, GCV, AIC, and BIC.

Due to the nature of the adaptive tuning parameter, it can be shown that the adaptive group LASSO estimators possess the oracle property (Wang and Leng, 2008).

However it is well known that this method is not robust in the presence of outliers, thus may yield incorrect estimators and as a result incorrect models. Therefore a robust version of this method is needed. The solution to this problem is to combine the adaptive tuning parameter from the adaptive group LASSO with the objective function of the group LAD-LASSO. With this mixture, we consider the following objective function to minimize:

$$Q(\boldsymbol{\beta}) = \frac{1}{2}\sum_{i=1}^{n}|y_i - \sum_{k=1}^{K}\boldsymbol{x}_{ik}\boldsymbol{\beta}_k| + n\sum_{i=1}^{K}\lambda_k||\boldsymbol{\beta}_k||_2. \tag{15}$$

Define $\boldsymbol{x}_{ik}$ to be a $1 \times p_k$ vector of predictors, where p_k is the number of predictors in group k, while $\boldsymbol{\beta}_k$ is a $p_k \times 1$ vector of regression coefficients. The penalty is the typical L_2 norm. The tuning parameter is defined such that $\lambda_k \geq 0$. Effectively, this results in regression estimators that will be robust to outliers in the response, while enjoying the shrinkage and nice theoretical properties of the adaptive LASSO to perform group selection.

The computation is the same as for the group LAD-LASSO with a small adjustment for the adaptive tuning parameter. This is done in R with a small modification to our code using the $grpreg$ package for our simulations (Breheny, 2015).

3.1 Tuning Parameter Selection

In general, the tuning parameter can usually be found using cross-validation (CV) or general cross-validation (GCV). However, this can be computationally intensive for the adaptive group variable selection problems, because there may be a large number of tuning parameters to compute if the number of groups k is large. For the tuning parameter λ_k in the adaptive group LAD-LASSO, we follow the example of Wang and Leng (2008) and choose:

$$\lambda_k = \frac{\lambda}{||\tilde{\boldsymbol{\beta}}_k||_2^{\gamma}} \tag{16}$$

such that $\tilde{\boldsymbol{\beta}} = (\tilde{\beta}_1^T, \ldots, \tilde{\beta}_p^T)^T$ is the LAD estimator and $\gamma > 0$ is a positive number chosen beforehand. For our simulation and real data application, we use $\gamma = 1$, as used by Wang and Leng (2008). As a result, instead of calculating a λ_k for each group, this reduces to a one-dimension problem where we only need to choose an appropriate λ. There are some well known selection criteria for λ, suggested by Wang and Leng (2008), such as CV, GCV, AIC, BIC where all require the df, the degrees of freedom and this is defined as in Yuan and Lin (2006), given by:

$$df = \sum_{k=1}^{K} I\{||\hat{\boldsymbol{\beta}}_k||_2 > 0\} + \sum_{k=1}^{K} \frac{||\hat{\boldsymbol{\beta}}_{\boldsymbol{k}}||_2}{||\tilde{\boldsymbol{\beta}}_k||_2}(p_k - 1). \tag{17}$$

Adapted for the adaptive group LAD-LASSO, $\tilde{\boldsymbol{\beta}}$ are the unpenalized LAD estimators, and $\hat{\sigma}^2$ is the variance estimator associated with $\tilde{\boldsymbol{\beta}}$. For our simulations , we use the default setting of choosing λ with the smallest value of the BIC criterion, which is given by

$$BIC = \log(\frac{1}{n}||\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}||_2^2) + \log(n) * df/n. \tag{18}$$

Using the BIC criterion is an example of a consistent model selection criteria. The consistent model selection criteria have the property of being able to identify the true model consistently, if a finite-dimensional true model exists. This is comparable to the efficient model selection criteria, which have the property of being able to select the best model by an appropriately defined asymptotic optimality criterion, which are useful when the true underlying model is too complicated to be well approximated by a finite-dimensional model (Wang et al, 2007).

3.2 Theoretical Properties

Assume that we decompose the regression coefficient $\boldsymbol{\beta} = (\boldsymbol{\beta}_a^T, \boldsymbol{\beta}_b^T)$, where $\boldsymbol{\beta}_a = (\beta_1, \ldots, \beta_{p_0})^T$ are the significant coefficients and $\boldsymbol{\beta}_b = (\beta_{p_0+1}, \ldots, \beta_p)^T$ are the insignificant coefficients

and denote the corresponding adaptive group LAD-LASSO estimators as $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}_a^T}, \hat{\boldsymbol{\beta}_b^T})$, and let the adaptive group LAD-LASSO objective function be denoted by $Q(\boldsymbol{\beta}) = Q(\boldsymbol{\beta}_a, \boldsymbol{\beta}_b)$.

Further, we make the following assumptions:

- The errors ε_i have continuous and positive density at the origin.
- The matrix $\text{cov}(\boldsymbol{x}) = \boldsymbol{\Sigma}$ exists and is positive definite.

and define $a_n = \max\{\lambda_j, j \leq p_0\}$ and $b_n = \min\{\lambda_j, j > p_0\}$. First, we can establish the consistency of the adaptive group LAD-LASSO estimators.

Theorem .1. *(Estimation Consistency) If $\sqrt{n}a_n \to_p 0$, then $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = O_p(\sqrt{n})$.*

Theorem 1 implies that if the shrinkage associated with the relevant nonzero predictors is sufficiently small, then the corresponding adaptive group LAD-LASSO estimator can be $\sqrt{n}$-consistent. The proof can be seen in the Appendix. The next theorem relates to the method's ability to properly estimate insignificant variables as zero.

Theorem .2. *(Selection Consistency) If $\sqrt{n}a_n \to_p 0$ and $\sqrt{n}b_n \to_p \infty$, then $P(\hat{\boldsymbol{\beta}}_b = 0) \to 1$.*

This theorem can also be thought of as proving the sparsity property. In other words, the adaptive group LAD-LASSO can consistently estimate zero coefficients as zero. That is, the method can perform parameter estimation and variable selection simultaneously. The proof of the theorem can be found in the Appendix. With both Theorem 1 & 2, we can establish the Oracle property.

Theorem .3. *(Oracle Property) If $\sqrt{n}a_n \to_p 0$ and $\sqrt{n}b_n \to_p \infty$, then $\sqrt{n}(\hat{\boldsymbol{\beta}}_a - \boldsymbol{\beta}_a) \to_d N(0, \boldsymbol{\Sigma}_a)$.*

Based on Theorem 2, with probability tending to one, all of the zero coefficients will be estimated as such, essentially performing variable selection. Based on Theorem 1, all of the estimates of the nonzero coefficients must be consistent, which implies that the nonzero coefficients must be estimated as such with probability tending to one. Putting these two theorems together leads to the conclusion of Theorem 3, which states that the adaptive group LAD-LASSO has the property to identify the correct model consistently.

The details and proofs of the above theorems are given in the appendix.

4 Simulation Study

Three main simulation studies were performed. The first is a simulation with outliers only in the response to compare the group LASSO to the group LAD-LASSO, while the second also has outliers only in the response to compare the adaptive group LASSO to the adaptive group LAD-LASSO. The package *grpreg* was used for the simulation in the statistical software R. The third simulation study involves high-dimensional data for the adaptive group LAD-LASSO.

4.1 Group LAD-LASSO

For sample sizes n=50,100, and 200, let ϵ be the contamination rate equal to values ϵ=0.1, 0.2, and 0.3 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 $\mathbf{0}$ 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 sequential 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 $\boldsymbol{\mu}_2 \neq \mathbf{0}$ and covariance equal to $\boldsymbol{I}$. Let $\beta_2 \neq \beta_1$. For our simulation, $\boldsymbol{\mu}_2 = \mathbf{5}$ and $\beta_2 = (4.5, 4, 3, 10, 0, 0)$ such that the first group has 4 nonzero values and the second group has 2 zero values. Both vectors were selected beforehand using a random number generator in R. The closer the values of $\boldsymbol{\mu}_2$ and β_2 are to $\boldsymbol{\mu}_1$ and β_1, the smaller the error becomes when there is contamination in the response and vice versa. For each combination of sample size, contamination rate, sigma, and error distribution, the simulation is performed 200 times, and the model error (ME) will be calculated for each of the given method's fit on the data for comparison purposes. The model error is calculated by:

$$ME(\hat{\boldsymbol{\beta}}) = \frac{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T \boldsymbol{X}^T \boldsymbol{X} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})}{n}. \tag{19}$$

The model error can be thought of as another measure of mean square error. Values of model error close to 0 indicate the calculated model is close to the true model, while values far away from 0 indicate the fitted model is not very close to the actual model. The results from the simulation are in Tables 1-3. Figure 1 provides box plots for the distributions of the model errors for the simulation for t_3 errors. Each time the group LAD-LASSO has the smallest median model error.

4.2 Adaptive Group LAD-LASSO

The adaptive group LASSO and adaptive group LAD-LASSO will be evaluated for data with outliers in the y-direction only, performed under the same conditions as the previous simulation for the group LAD-LASSO. In addition to the model error, the tables include a column for the mean % of correct zeros, denoted as Mean % of CZ. For the 200 times the simulation is run, the percentage of correct zeros is calculated (of the zeros found by the model, the percentage of correct zeros is determined as the fraction of coefficients that are actually supposed to be zero and the overall number of zero coefficients), and this column indicates the overall average percentage of correct zeros of those 200 simulations. Tables 4-6 give the resulting model errors for the various setups. Figure 2 gives the box plots for model error for the two adaptive methods. In all cases with contamination, the adaptive group LAD-LASSO gives the smallest model error, which is also close to 0. This is sup-

Table 1: Simulation results for $N(0,1)$ errors for strictly Y-outliers

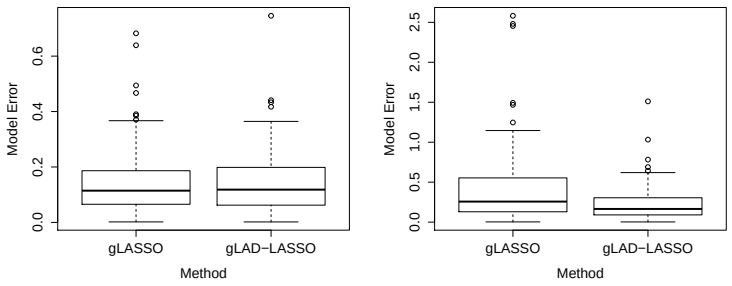
σ	n	ϵ	Method	Mean ME	Median ME
0.5	50	0	g LASSO	0.03	0.02
			g LAD-LASSO	0.03	0.03
		0.1	g LASSO	0.37	0.22
			g LAD-LASSO	0.17	0.13
		0.2	g LASSO	1.27	0.99
			g LAD-LASSO	0.34	0.24
		0.3	g LASSO	2.95	2.36
			g LAD-LASSO	0.70	0.46
	100	0	g LASSO	0.01	0.01
			g LAD-LASSO	0.01	0.01
		0.1	g LASSO	0.26	0.18
			g LAD-LASSO	0.09	0.07
		0.2	g LASSO	1.18	0.95
			g LAD-LASSO	0.21	0.17
		0.3	g LASSO	3.03	2.88
			g LAD-LASSO	0.37	0.33
	200	0	g LASSO	0.01	0.01
			g LAD-LASSO	0.01	0.01
		0.1	g LASSO	0.22	0.18
			g LAD-LASSO	0.06	0.05
		0.2	g LASSO	1.22	1.09
			g LAD-LASSO	0.15	0.13
		0.3	g LASSO	3.08	3.08
			g LAD-LASSO	0.24	0.21
1.0	50	0	g LASSO	0.10	0.10
			g LAD-LASSO	0.11	0.10
		0.1	g LASSO	0.46	0.28
			g LAD-LASSO	0.25	0.20
		0.2	g LASSO	1.44	1.07
			g LAD-LASSO	0.47	0.32
		0.3	g LASSO	3.04	2.33
			g LAD-LASSO	1.02	0.60
	100	0	g LASSO	0.06	0.05
			g LAD-LASSO	0.06	0.05
		0.1	g LASSO	0.37	0.26
			g LAD-LASSO	0.13	0.12
		0.2	g LASSO	1.09	0.91
			g LAD-LASSO	0.25	0.19
		0.3	g LASSO	2.84	2.69
			g LAD-LASSO	0.51	0.38
	200	0	g LASSO	0.03	0.02
			g LAD-LASSO	0.03	0.02
		0.1	g LASSO	0.26	0.22
			g LAD-LASSO	0.07	0.05
		0.2	g LASSO	1.18	1.07
			g LAD-LASSO	0.15	0.12
		0.3	g LASSO	2.99	2.79
			g LAD-LASSO	0.34	0.28

Table 2: Simulation results for t_3 errors for strictly Y-outliers

σ	n	ϵ	Method	Mean ME	Median ME
0.5	50	0	g LASSO	0.07	0.06
			g LAD-LASSO	0.09	0.06
		0.1	g LASSO	0.43	0.23
			g LAD-LASSO	0.19	0.16
		0.2	g LASSO	1.34	1.13
			g LAD-LASSO	0.37	0.27
		0.3	g LASSO	3.03	2.34
			g LAD-LASSO	0.79	0.50
	100	0	g LASSO	0.04	0.03
			g LAD-LASSO	0.04	0.03
		0.1	g LASSO	0.29	0.20
			g LAD-LASSO	0.09	0.06
		0.2	g LASSO	1.20	0.99
			g LAD-LASSO	0.20	0.17
		0.3	g LASSO	3.08	2.91
			g LAD-LASSO	0.38	0.30
	200	0	g LASSO	0.02	0.01
			g LAD-LASSO	0.02	0.02
		0.1	g LASSO	0.23	0.19
			g LAD-LASSO	0.06	0.05
		0.2	g LASSO	1.11	1.02
			g LAD-LASSO	0.16	0.14
		0.3	g LASSO	2.87	2.75
			g LAD-LASSO	0.27	0.24
1.0	50	0	g LASSO	0.31	0.23
			g LAD-LASSO	0.31	0.22
		0.1	g LASSO	0.58	0.36
			g LAD-LASSO	0.38	0.26
		0.2	g LASSO	1.50	1.15
			g LAD-LASSO	0.57	0.39
		0.3	g LASSO	3.24	2.78
			g LAD-LASSO	1.50	0.78
	100	0	g LASSO	0.14	0.11
			g LAD-LASSO	0.14	0.12
		0.1	g LASSO	0.40	0.26
			g LAD-LASSO	0.22	0.17
		0.2	g LASSO	1.23	1.05
			g LAD-LASSO	0.30	0.22
		0.3	g LASSO	2.98	2.70
			g LAD-LASSO	0.93	0.56
	200	0	g LASSO	0.08	0.06
			g LAD-LASSO	0.07	0.06
		0.1	g LASSO	0.31	0.26
			g LAD-LASSO	0.09	0.07
		0.2	g LASSO	1.15	1.02
			g LAD-LASSO	0.18	0.14
		0.3	g LASSO	2.94	2.73
			g LAD-LASSO	0.44	0.34

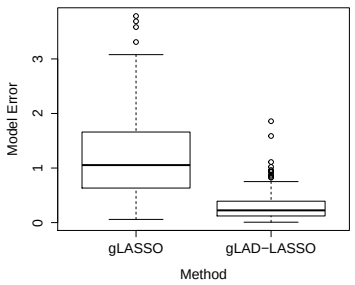
Table 3: Simulation results for t_5 errors for strictly Y-outliers

σ	n	ϵ	Method	Mean ME	Median ME
0.5	50	0	g LASSO	0.05	0.04
			g LAD-LASSO	0.05	0.04
		0.1	g LASSO	0.39	0.21
			g LAD-LASSO	0.17	0.12
		0.2	g LASSO	1.21	0.91
			g LAD-LASSO	0.41	0.27
		0.3	g LASSO	3.04	2.42
			g LAD-LASSO	0.66	0.47
	100	0	g LASSO	0.02	0.02
			g LAD-LASSO	0.02	0.02
		0.1	g LASSO	0.27	0.17
			g LAD-LASSO	0.09	0.08
		0.2	g LASSO	1.11	1.01
			g LAD-LASSO	0.22	0.15
		0.3	g LASSO	2.76	2.46
			g LAD-LASSO	0.38	0.28
	200	0	g LASSO	0.01	0.01
			g LAD-LASSO	0.01	0.01
		0.1	g LASSO	0.24	0.20
			g LAD-LASSO	0.06	0.05
		0.2	g LASSO	1.21	1.16
			g LAD-LASSO	0.15	0.13
		0.3	g LASSO	3.00	2.89
			g LAD-LASSO	0.25	0.23
1.0	50	0	g LASSO	0.17	0.14
			g LAD-LASSO	0.17	0.14
		0.1	g LASSO	0.56	0.36
			g LAD-LASSO	0.30	0.22
		0.2	g LASSO	1.41	0.99
			g LAD-LASSO	0.57	0.37
		0.3	g LASSO	3.40	2.74
			g LAD-LASSO	1.25	0.63
	100	0	g LASSO	0.10	0.09
			g LAD-LASSO	0.09	0.08
		0.1	g LASSO	0.37	0.27
			g LAD-LASSO	0.14	0.12
		0.2	g LASSO	1.30	1.10
			g LAD-LASSO	0.27	0.22
		0.3	g LASSO	2.90	2.60
			g LAD-LASSO	0.61	0.41
	200	0	g LASSO	0.04	0.04
			g LAD-LASSO	0.05	0.04
		0.1	g LASSO	0.29	0.24
			g LAD-LASSO	0.08	0.07
		0.2	g LASSO	1.13	0.98
			g LAD-LASSO	0.17	0.14
		0.3	g LASSO	2.92	2.84
			g LAD-LASSO	0.37	0.32

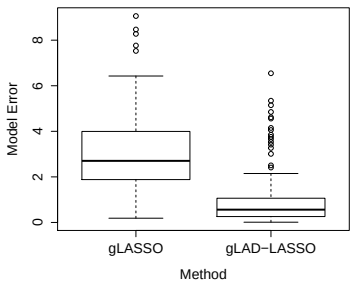


(a) 0%

(b) 10%

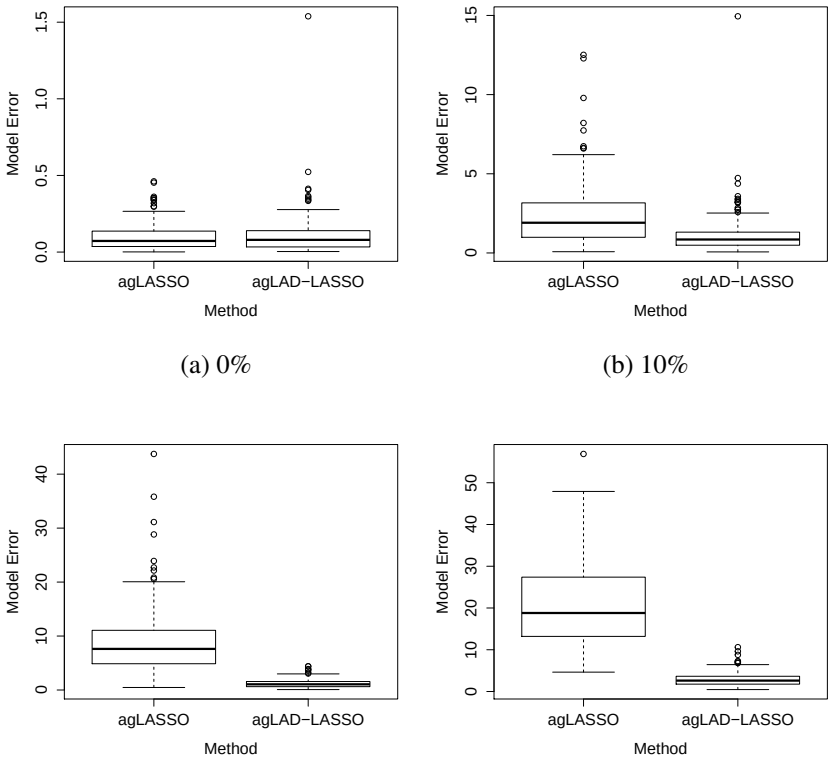


(c) 20%



(d) 30%

Figure 1: Boxplots for Model Error for the strictly y-outlier simulation for comparing the group LASSO (gLASSO) to the group LAD-LASSO (gLAD-LASSO) for various contamination levels for $\varepsilon \sim t_3$ over 200 simulations for $\sigma = 1$ and $n = 100$.



(a) 0%

(b) 10%

(c) 20%

(d) 30%

Figure 2: Boxplots for Model Error for the strictly y-outlier simulation for comparing the adaptive group LASSO (agLASSO) to the adaptive group LAD-LASSO (agLAD-LASSO) for various contamination levels for $\varepsilon \sim t_3$ over 200 simulations for $\sigma = 1$ and $n = 100$.

ported visually by the box plots, indicating that the adaptive group LAD-LASSO works well for data with contaminations in the response variable.

4.3 Adaptive Group LAD-LASSO for High-Dimensional Data

In order to demonstrate the effectiveness of the group variable selection methods with high-dimensional data, we perform a simulation study where $p > n$ (Huang et al, 2008). The conditions for calculating $\boldsymbol{y}_1$, $\boldsymbol{X}_1$, $\boldsymbol{y}_2$, and $\boldsymbol{X}_2$ are the same as for the previous two simulation studies. For sample size n, let ϵ be the contamination rate equal to values ϵ=0.1,

Table 4: Simulation results for $N(0, 1)$ errors for strictly Y-outliers

σ	n	ϵ	Method	Mean % of CZ	Mean ME	Median ME
0.5	50	0	ag LASSO	99.7	0.03	0.02
			ag LAD-LASSO	100	0.02	0.01
		0.1	ag LASSO	26.4	0.57	0.39
			ag LAD-LASSO	99.8	0.36	0.24
		0.2	ag LASSO	26.0	1.75	1.37
			ag LAD-LASSO	98.0	0.20	0.08
		0.3	ag LASSO	29.8	4.02	3.51
			ag LAD-LASSO	95.6	0.72	0.38
	100	0	ag LASSO	100	0.01	0.01
			ag LAD-LASSO	100	0.01	0.01
		0.1	ag LASSO	30.1	0.47	0.34
			ag LAD-LASSO	100	0.31	0.24
		0.2	ag LASSO	29.9	1.54	1.38
			ag LAD-LASSO	100	0.08	0.15
		0.3	ag LASSO	33.4	3.46	3.24
			ag LAD-LASSO	96.7	0.67	0.28
	200	0	ag LASSO	100	0.01	0.01
			ag LAD-LASSO	100	0.00	0.00
		0.1	ag LASSO	24.5	0.35	0.31
			ag LAD-LASSO	100	0.23	0.20
		0.2	ag LASSO	25.6	1.31	1.22
			ag LAD-LASSO	100	0.36	0.17
		0.3	ag LASSO	43.1	3.17	3.03
			ag LAD-LASSO	99.4	0.30	0.28
1.0	50	0	ag LASSO	97.8	0.08	0.07
			ag LAD-LASSO	92.3	0.08	0.06
		0.1	ag LASSO	24.6	0.62	0.44
			ag LAD-LASSO	91.5	0.50	0.34
		0.2	ag LASSO	25.7	1.79	1.47
			ag LAD-LASSO	93.0	0.42	0.15
		0.3	ag LASSO	23.2	4.16	3.39
			ag LAD-LASSO	94.7	0.20	0.12
	100	0	ag LASSO	99.0	0.04	0.03
			ag LAD-LASSO	100	0.04	0.03
		0.1	ag LASSO	24.2	0.40	0.31
			ag LAD-LASSO	95.7	0.31	0.25
		0.2	ag LASSO	38.3	1.51	1.36
			ag LAD-LASSO	96.2	0.09	0.01
		0.3	ag LASSO	43.3	3.53	3.21
			ag LAD-LASSO	97.0	0.16	0.16
	200	0	ag LASSO	98.8	0.02	0.02
			ag LAD-LASSO	99.3	0.02	0.01
		0.1	ag LASSO	39.5	0.32	0.26
			ag LAD-LASSO	100	0.24	0.20
		0.2	ag LASSO	37.6	1.34	1.24
			ag LAD-LASSO	96.6	0.24	0.21
		0.3	ag LASSO	24.8	3.27	3.14
			ag LAD-LASSO	97.4	0.37	0.26

Table 5: Simulation results for t_3 errors for strictly Y-outliers

σ	n	ϵ	Method	Mean % of CZ	Mean ME	Median ME
0.5	50	0	ag LASSO	99.1	0.09	0.05
			ag LAD-LASSO	99.8	0.18	0.04
		0.1	ag LASSO	45.7	2.87	1.96
			ag LAD-LASSO	98.7	0.47	0.32
		0.2	ag LASSO	38.4	9.91	7.84
			ag LAD-LASSO	96.8	0.30	0.14
		0.3	ag LASSO	29.5	24.11	21.95
			ag LAD-LASSO	94.8	0.23	0.14
	100	0	ag LASSO	93.6	0.03	0.03
			ag LAD-LASSO	95.8	0.02	0.02
		0.1	ag LASSO	44.7	2.23	1.75
			ag LAD-LASSO	95.7	0.30	0.27
		0.2	ag LASSO	32.0	8.55	7.33
			ag LAD-LASSO	99.5	0.20	0.15
		0.3	ag LASSO	23.3	21.39	19.90
			ag LAD-LASSO	97.1	0.36	0.23
	200	0	ag LASSO	99.3	0.02	0.02
			ag LAD-LASSO	99.7	0.01	0.01
		0.1	ag LASSO	39.6	1.99	1.71
			ag LAD-LASSO	91.4	0.23	0.19
		0.2	ag LASSO	31.5	7.78	7.49
			ag LAD-LASSO	90.7	0.48	0.44
		0.3	ag LASSO	21.8	20.29	19.11
			ag LAD-LASSO	94.5	0.44	0.32
1.0	50	0	ag LASSO	95.9	0.28	0.17
			ag LAD-LASSO	96.5	0.23	0.15
		0.1	ag LASSO	48.1	3.43	2.39
			ag LAD-LASSO	95.9	0.64	0.45
		0.2	ag LASSO	44.7	11.20	9.51
			ag LAD-LASSO	93.5	0.55	0.49
		0.3	ag LASSO	23.6	24.10	22.36
			ag LAD-LASSO	92.4	0.44	0.39
	100	0	ag LASSO	97.0	0.10	0.07
			ag LAD-LASSO	99.0	0.11	0.08
		0.1	ag LASSO	47.3	2.35	1.91
			ag LAD-LASSO	94.8	0.37	0.28
		0.2	ag LASSO	41.5	8.83	7.62
			ag LAD-LASSO	92.7	0.33	0.29
		0.3	ag LASSO	21.4	20.69	18.80
			ag LAD-LASSO	92.1	0.27	0.28
	200	0	ag LASSO	99.3	0.05	0.04
			ag LAD-LASSO	99.8	0.05	0.03
		0.1	ag LASSO	45.3	1.86	1.55
			ag LAD-LASSO	95.4	0.28	0.21
		0.2	ag LASSO	28.1	7.83	7.35
			ag LAD-LASSO	92.4	0.54	0.48
		0.3	ag LASSO	20.2	20.12	19.56
			ag LAD-LASSO	91.3	0.39	0.28

Table 6: Simulation results for t_5 errors for strictly Y-outliers

σ	n	ϵ	Method	Mean % of CZ	Mean ME	Median ME
0.5	50	0	ag LASSO	97.4	0.05	0.04
			ag LAD-LASSO	97.5	0.03	0.02
		0.1	ag LASSO	48.7	3.80	2.44
			ag LAD-LASSO	93.7	0.35	0.24
		0.2	ag LASSO	43.0	10.69	9.03
			ag LAD-LASSO	93.0	0.53	0.45
		0.3	ag LASSO	32.9	23.98	22.08
			ag LAD-LASSO	90.0	0.43	0.40
	100	0	ag LASSO	98.4	0.02	0.02
			ag LAD-LASSO	98.7	0.01	0.01
		0.1	ag LASSO	48.0	2.44	1.81
			ag LAD-LASSO	94.1	0.28	0.21
		0.2	ag LASSO	36.5	8.48	7.51
			ag LAD-LASSO	93.1	0.58	0.48
		0.3	ag LASSO	27.6	21.21	20.02
			ag LAD-LASSO	91.6	0.32	0.28
	200	0	ag LASSO	99.9	0.01	0.01
			ag LAD-LASSO	100	0.01	0.00
		0.1	ag LASSO	47.3	1.98	1.67
			ag LAD-LASSO	95.6	0.24	0.22
		0.2	ag LASSO	36.1	7.53	6.95
			ag LAD-LASSO	93.3	0.64	0.40
		0.3	ag LASSO	23.0	19.87	18.74
			ag LAD-LASSO	92.8	0.32	0.25
1.0	50	0	ag LASSO	97.2	0.14	0.12
			ag LAD-LASSO	97.8	0.15	0.10
		0.1	ag LASSO	47.8	3.85	2.40
			ag LAD-LASSO	94.2	0.54	0.34
		0.2	ag LASSO	43.0	11.69	9.21
			ag LAD-LASSO	92.8	0.30	0.25
		0.3	ag LASSO	30.5	23.65	22.12
			ag LAD-LASSO	90.5	0.35	0.20
	100	0	ag LASSO	98.6	0.07	0.05
			ag LAD-LASSO	98.7	0.06	0.05
		0.1	ag LASSO	46.6	2.29	1.82
			ag LAD-LASSO	94.4	0.33	0.27
		0.2	ag LASSO	40.1	8.40	7.25
			ag LAD-LASSO	93.1	0.33	0.29
		0.3	ag LASSO	26.4	20.95	19.68
			ag LAD-LASSO	90.8	0.39	0.28
	200	0	ag LASSO	99.1	0.03	0.02
			ag LAD-LASSO	99.2	0.02	0.02
		0.1	ag LASSO	44.6	1.90	1.55
			ag LAD-LASSO	97.0	0.26	0.21
		0.2	ag LASSO	32.0	8.02	7.71
			ag LAD-LASSO	93.8	0.30	0.16
		0.3	ag LASSO	21.1	19.48	18.49
			ag LAD-LASSO	92.5	0.25	0.17

0.2, and 0.3 such that $m = [\epsilon n]$ is the number of contaminated data points as before. 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 $\mathbf{0}$ 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 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 $\boldsymbol{\mu}_2 \neq \mathbf{0}$ and covariance equal to $\boldsymbol{I}$. Let $\beta_2 \neq \beta_1$. Both vectors were selected beforehand using a random number generator in R.

This simulation specifically sets a high-dimensional setting for the data where $p > n$. To achieve this, $n = 100$ and $p = 400$. The first 15 predictors are nonzero, while the last 385 predictors are zero, which also allows for a sparse predictor matrix. The predictor matrix is partitioned into 80 groups of 5 predictors each such that $p_k = 5$. The first five predictors of $\boldsymbol{\beta}_1$ are equal to 2.5, the next five predictors are equal to 1.5, the third group of five predictors is equal to 0.5, while the rest of the groups of five predictors are all equal to 0. The contamination is included such that $\mu_2 = \mathbf{5}$, and the first five components of β_2 are equal to 5, the second five components are equal to 4, the third group of five components is equal to 3, the fourth group of five components is equal to 2, the fifth group of five components is equal to 1, and the remaining 375 components, which comprise the remaining 75 groups of five components each, are all equal to 0. The errors are t-distributed with 5 degrees of freedom, while $\sigma = 1$. The simulation is replicated 200 times in R. The model error (ME) will be calculated for each of the given method's fit on the data for comparison purposes as in (19).

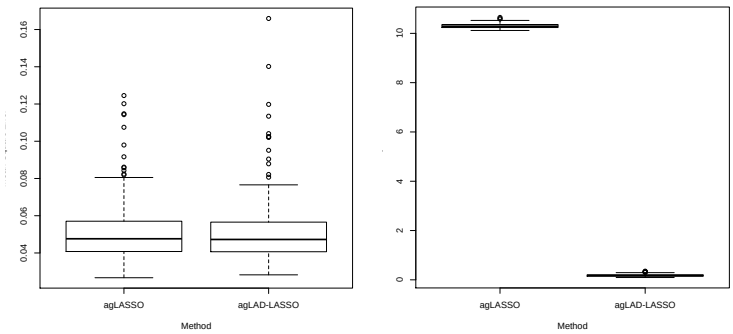
The results of the high-dimensional simulation are shown in table 7, and the box plots of the model error for the various contamination levels are shown in figure 3. The adaptive group LASSO and the adaptive group LAD-LASSO perform similarly when there is no contamination with model errors both close to zero. However, once there is contamination, the advantage in the robust adaptive group LAD-LASSO becomes apparent. The mean model error for the adaptive group LASSO begins to increase to 10.30, 44.95, and 71.14, as the contamination level increases to 10%, 20%, and 30%, respectively, while the mean model error for the more robust adaptive group LAD-LASSO stays close to 0 by being equal to 0.18, 0.14, 0.11, as the contamination level increases to 10%, 20%, and 30%, respectively. The results also show that the adaptive group LAD-LASSO correctly estimates the zero groups as zero more often than the adaptive group LASSO when there is contamination, supporting the sparsity property of the adaptive group LAD-LASSO. Results are similar for non-adaptive comparisons between the group LASSO and the group LAD-LASSO.

5 Real Data Application

In order to show the effectiveness of the two methods, a real data example is presented. The data are from microarray experiments of mammalian eye tissue samples and contain gene expression information from 120 subjects (Scheetz et al., 2006). The response is the expression level of gene TRIM32, which causes Bardet-Biedl syndrome. There are 100

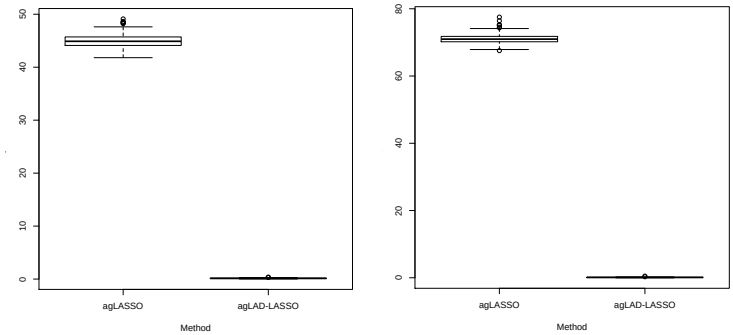
Table 7: Simulation results for t_5 errors for strictly Y-outliers with High-Dimensional Data

σ	n	ϵ	Method	Mean % of CZ	Mean ME	Median ME
1.0	100	0	ag LASSO	98.6	0.05	0.05
			ag LAD-LASSO	99.0	0.05	0.05
		0.1	ag LASSO	77.2	10.30	10.29
			ag LAD-LASSO	95.6	0.18	0.17
		0.2	ag LASSO	73.7	44.95	44.89
			ag LAD-LASSO	94.0	0.14	0.13
		0.3	ag LASSO	69.5	71.14	70.98
			ag LAD-LASSO	93.4	0.11	0.11



(a) 0%

(b) 10%



(c) 20%

(d) 30%

Figure 3: Boxplots for Model Error for the strictly y-outlier simulation with high-dimensional data for comparing the group LASSO (gLASSO) to the group LAD-LASSO (gLAD-LASSO) for 0% (top left), 10% (top right), 20% (bottom left), and 30% (bottom right) contamination levels for $\varepsilon \sim t_5$ over 200 simulations for $\sigma = 1$, $n = 100$, $p = 400$.

Table 8: MSE for the application on the Bardet data set.

Method	Contamination	Mean MSE	Median MSE
gLASSO	0%	0.023	0.021
gLAD-LASSO		0.029	0.028
agLASSO		0.010	0.010
agLAD-LASSO		0.010	0.010
gLASSO	20% y-outliers	84.949	85.476
gLAD-LASSO		0.173	0.176
agLASSO		56.194	56.263
agLAD-LASSO		0.060	0.040

predictors, which are the expression levels of 20 genes, which were expanded using 5 basis B-splines (Yang and Zou, 2012). That is, each 5 consecutive columns corresponds to a grouped gene.

Preliminary analyses of the data indicate there is multicollinearity between the predictors. For example, marker 4 has a correlation equal to 0.77 with marker 19, and marker 5 has a correlation equal to 0.99 with marker 30. However, since this happens with only a few pairs of variables, the multicollinearity is not severe enough to warrant a change from the LASSO-based methods (Dormann et al, 2013). A scatter plot matrix indicates that there is at least one outlier in the response, and some outliers in the predictor space, including about nine observations for marker 4. As a result, 24 observations in the response are randomly chosen and shifted to become outliers. (24 observations are 20% of the overall 120 observations, indicating there will be 20% contamination of outliers).

All four methods are performed on the data set to see which groups of genes are important in predicting the expression level of gene TRIM32. The methods are examined by using the following measure. We find the mean square error for each method over 100 runs of fitting the model with k-fold cross-validation and report the average of the 100 mean square errors.

$$\text{MSE} = \frac{1}{n-p}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{20}$$

The results are in Table 7. Box plots of the mean square error are provided

When there are no outliers, all of the methods perform equivalently. When accounting for outliers in the response, it is clear the the group LAD-LASSO and the adaptive group LAD-LASSO outperform their non-robust counterparts, the group LASSO and the adaptive group LASSO. This can also be seen from the box plots.

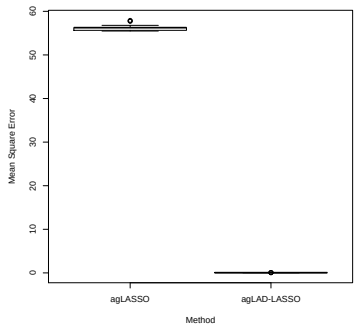


Figure 4: Box plot for the Mean Square Error for the adaptive group LASSO (agLASSO) and the adaptive group LAD-LASSO (agLAD-LASSO) for the given condition over 100 fittings on the Bardet data set.

6 Conclusions

In this paper, two new methods for robust variable selection with grouped predictors were proposed. The group LAD-LASSO and the adaptive group LAD-LASSO is appropriate when the data includes outliers in the response. Both methods prove to be more robust in simulations and in a real data example than the group LASSO and the adaptive group LASSO. The adaptive group LASSO is shown to have nice theoretical properties, including the oracle property, due to its adaptive tuning parameter. In all cases presented when there are outliers in the response and group selection is a priority, the group LAD-LASSO and adaptive group LAD-LASSO are well-suited to the task.

Additionally, this work leads to ideas to pursue for further research in the area of group variable selection methods that perform group selection and variable selection within the groups simultaneously, which the group LASSO is not designed to do. One such method that could do this is the group bridge. For instance, in the real data set example, both of the adaptive methods selected all three groups, so a group selection method that also does within-group variable selection would be useful in that case in order to see if a subset of the individual variables within the group are actually significant.

References

[1] Bloomfield, P., and Steiger, W. L. 1983. Least Absolute Deviation: Theory, Applications, and Algorithms Birkhauser, Boston

[2] Breheny, P. and Huang, J. 2015. Group descent algorithms for a non convex penalized linear and logistic regression models with grouped predictors. Statistics and Computing 25:173-187

[3] Davis, R.A., Knight, K., and Liu, J. 1992. M-Estimation for Autoregressions with Infinite Variance. Stochastic Process and Their Applications 40:145-180

[4] Dormann, C.F., Elith, J., Bacher, S., Buchmann, C. et al. 2013. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. Ecography 36:027-046

[5] Fan, J. and Li, R. 2001. Variable selection via non concave penalized likelihood and its oracle properties. Journal of the American Statistical Association 96:1348-1360

[6] Friedman, J., Hastie, T., Höfling, H., and Tibshirani, R. 2007. Pathwise coordinate optimization. Annals of Applied Statistics 1:302-332

[7] Giloni, A., Simonoff, J., and Sengupta, B. 2005. Robust weighted LAD regression. Computational Statistics & Data Analysis 50:3124-3140

[8] Huang, J., Ma, S., Zhang, C.H. 2008. Adaptive LASSO for Sparse High-Dimensional Regression Models. Statistica Sinica 18:1603-1618

[9] Hubert, M. and Rousseeuw, P. 1997. Robust regression with both continuous and binary regression. Journal of Statistical Planning and Inference 57:153-163

[10] Knight, K. 1998. Limiting Distributions for L_1 Regression Estimators Under General Conditions. The Annals of Statistics 26:755-770

[11] Koenker, R., and Zhao, Q. 1996. Conditional Quantile Estimation and Inference for Arch Models. Econometric Theory 12:793-813

[12] 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

[13] Tibshirani, R. J. 1996. Regression shrinkage and selection via the LASSO. Journal of the Royal Statistical Society Series B 58:267-288

[14] Wang, H., Li, G., and Jiang, G. 2007. Robust regression shrinkage and consistent variable selection via the LAD-LASSO. Journal of Business and Economics Statistics 25:347-355

[15] Wang, H., and Leng, C. 2008. A note on the adaptive group lasso. Computational Statistics & Data Analysis 52:5277-5286

[16] Wu, T. and Lange, K. 2008. Coordinate descent algorithms for LASSO penalized regression. Annals of Applied Statistics 2:224-244

[17] Yang, Y. and Zou, H. 2015. A Fast Unified Algorithm for Computing Group-LASSO Penalized Learning Problems. Statistics and Computing 25:1126-1141

[18] 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

[19] Zou, H. 2006. The adaptive lasso and its oracle properties. Journal of the American Statistical Association 101:1418-1429

7 Appendix

7.1 Proof of Theorem 1

Before the proof, first assume all of the conditions presented before the theorem in section 3.3. Therefore, we assume the groups are ordered such that all significant nonzero groups are first in the grouping order, and all insignificant zero groups are ordered to be last. For example, if there are four groups, and two are significant, groups 1 and 2 would be the significant groups, and groups 3 and 4 would be the nonsignificant groups. Furthermore, assume k_0 is the largest value of k such that the group k_0 is significant and nonzero.

It should be noted that the objective function of the adaptive group LAD-LASSO $Q(\boldsymbol{\beta})$ (15) is convex. As long as we can show a local minimizer of $Q(\boldsymbol{\beta})$, which is $\sqrt{n}$-consistent, then by global convexity of $Q(\boldsymbol{\beta})$, the local minimizer must be $\hat{\beta}$, the adaptive group LAD-LASSO estimators. In order to show the existence of a $\sqrt{n}$-consistent local minimizer, we want to show that for any given $\epsilon > 0$, there exists a sufficiently large constant C such that

$$\liminf_{n} P\left\{\inf_{||\boldsymbol{u}||=C} Q(\boldsymbol{\beta} + n^{-1/2}\boldsymbol{u}) > Q(\boldsymbol{\beta})\right\} > 1 - \epsilon, \tag{21}$$

where $\boldsymbol{u} = (u_1, \ldots, u_p)^T$ is a p-dimensional vector such that $||\boldsymbol{u}|| = C$. Let $D_n(\boldsymbol{u}) = Q(\boldsymbol{\beta} + n^{-1/2}\boldsymbol{u}) - Q(\boldsymbol{\beta})$. Then,

$$
\begin{aligned}
D_n(\boldsymbol{u}) = & \sum_{i=1}^{n} \frac{1}{2}|y_i - \sum_{k=1}^{K} \boldsymbol{x}_{ik}(\boldsymbol{\beta}_k + n^{-1/2}\boldsymbol{u}_k)| + n\sum_{k=1}^{K} \lambda_k ||\boldsymbol{\beta}_k + n^{-1/2}\boldsymbol{u}_k||_2 \\
& - \sum_{i=1}^{n} \frac{1}{2}|y_i - \sum_{k=1}^{K} \boldsymbol{x}_{ik}\boldsymbol{\beta}_k| + n\sum_{i=1}^{K} \lambda_k ||\boldsymbol{\beta}_k||_2 \qquad (22) \\
= & \sum_{i=1}^{n} \frac{1}{2}\{|y_i - \sum_{k=1}^{K} \boldsymbol{x}_{ik}(\boldsymbol{\beta}_k + n^{-1/2}\boldsymbol{u}_k)| - |y_i - \sum_{k=1}^{K} \boldsymbol{x}_{ik}\boldsymbol{\beta}_k|\} \\
& + n\sum_{k=1}^{K} \lambda_k ||\boldsymbol{\beta}_k + n^{-1/2}\boldsymbol{u}_k||_2 - n\sum_{i=1}^{K} \lambda_k ||\boldsymbol{\beta}_k||_2 \qquad (23) \\
= & \sum_{i=1}^{n} \frac{1}{2}\{|y_i - \sum_{k=1}^{K} \boldsymbol{x}_{ik}(\boldsymbol{\beta}_k + n^{-1/2}\boldsymbol{u}_k)| - |y_i - \sum_{k=1}^{K} \boldsymbol{x}_{ik}\boldsymbol{\beta}_k|\} \\
& + n\sum_{k=1}^{K} \lambda_k ||\boldsymbol{\beta}_k + n^{-1/2}\boldsymbol{u}_k||_2 - n\sum_{k=1}^{k_0} \lambda_k ||\boldsymbol{\beta}_k||_2 \qquad (24) \\
\geq & \sum_{i=1}^{n} \frac{1}{2}\{|y_i - \sum_{k=1}^{K} \boldsymbol{x}_{ik}(\boldsymbol{\beta}_k + n^{-1/2}\boldsymbol{u}_k)| - |y_i - \sum_{k=1}^{K} \boldsymbol{x}_{ik}\boldsymbol{\beta}_k|\} \\
& + n\sum_{k=1}^{k_0} \lambda_k (||\boldsymbol{\beta}_k + n^{-1/2}\boldsymbol{u}_k||_2 - ||\boldsymbol{\beta}||_2) \qquad (25) \\
\geq & \sum_{i=1}^{n} \frac{1}{2}\{|y_i - \sum_{k=1}^{K} \boldsymbol{x}_{ik}(\boldsymbol{\beta}_k + n^{-1/2}\boldsymbol{u}_k)| - |y_i - \sum_{k=1}^{K} \boldsymbol{x}_{ik}\boldsymbol{\beta}_k|\} \\
& + p_0\sqrt{n}a_n \sum_{k=1}^{k_0} ||\boldsymbol{u}_k||_2 \qquad (26)
\end{aligned}
$$

Equation (25) follows from (24), because $\boldsymbol{\beta}_k = 0$ for any $j > p_0$. Divide equation (26) into two parts, separated by the $+$. Denote the first part as $L_n(\boldsymbol{u})$. Because of the theorem's conditions, we know $\sqrt{n}a_n = o(1)$, which implies the second and last term is of $o(1)$. Next, we must show how $L_n(\boldsymbol{u})$ behaves.

Using an equation from Knight (1998), for $x \neq 0$:

$$|x - y| - |x| = -y[I(x > 0) - I(x < 0)] + 2\int_0^y [I(x \leq s) - I(x \leq 0)]ds$$

Then $L_n(\boldsymbol{u})$ can be rewritten as:

$$\sum_{i=1}^{n}\{|y_i - \sum_{k=1}^{K} \boldsymbol{x}_{ik}\boldsymbol{\beta}_k - \sum_{k=1}^{K} \boldsymbol{x}_{ik}n^{-1/2}\boldsymbol{u}_k| - |y_i - \sum_{k=1}^{K} \boldsymbol{x}_{ik}\boldsymbol{\beta}_k| \qquad (27)$$

which, in turn, can be written as (with help from Knight (1998):

$$-n^{-1/2}\boldsymbol{u}\sum_{i=1}^{n}\boldsymbol{x}_i[I(\epsilon_i>0)-I(\epsilon_i<0)]+2\sum_{i=1}^{n}\int_0^{n^{-1/2}\boldsymbol{u}^T\boldsymbol{x}_i}[I(\epsilon\le s)-I(\epsilon\le 0)]ds \quad (28)$$

By the Central Limit Theorem, the first term of (28) converges in distribution to $\boldsymbol{u}^T\boldsymbol{W}$, where $\boldsymbol{W}$ is a p-dimensional normal random vector with mean 0 and covariance matrix $\boldsymbol{\Sigma}$. Now, as for the second part of (28), denote the c.d.f. of ϵ_i by F and $\int_0^{n^{-1/2}\boldsymbol{u}^T\boldsymbol{x}_i}[I(\epsilon \le s) - I(\epsilon \le 0)]ds$ by $Z_{ni}(\boldsymbol{u})$. Hence,

$$nE[Z_{ni}(\boldsymbol{u})I(n^{-1/2}|\boldsymbol{u}^T\boldsymbol{x}_i|\ge\eta)] \le nE\{(\int_0^{n^{-1/2}|\boldsymbol{u}^T\boldsymbol{x}_i|}2ds)^2 I(n^{-1/2}|\boldsymbol{u}^T\boldsymbol{x}_i|\ge\eta)\} \quad (29)$$

$$= 4E[|\boldsymbol{u}^T\boldsymbol{x}|^2 I(|\boldsymbol{u}^T\boldsymbol{x}|\ge\sqrt{n}\eta)] \quad (30)$$

$$= o(1) \quad (31)$$

However, due to the continuity of f, there exists an $\eta > 0$ and $0 < \kappa < \infty$ such that $\sup_{|x|<\eta} f(x) < f(0) + \kappa$. Let $R = nE[Z_{ni}^2(\boldsymbol{u})I(n^{-1/2}|\boldsymbol{u}^T\boldsymbol{x}_i| < \eta)]$. Then,

$$R \le 2n\eta E\{\int_0^{n^{-1/2}|\boldsymbol{u}^T\boldsymbol{x}_i|}|I(\epsilon_i\le s) - I(\epsilon_i\le 0)ds * I(n^{-1/2}|\boldsymbol{u}^T\boldsymbol{x}_i| < \eta)\} \quad (32)$$

$$\le 2n\eta E\{\int_0^{n^{-1/2}|\boldsymbol{u}^T\boldsymbol{x}_i|}[F(s) - F(0)]ds * I(n^{-1/2}|\boldsymbol{u}^T\boldsymbol{x}_i| < \eta)\} \quad (33)$$

$$\le 2n\eta\{f(0)+\kappa\}E\{\int_0^{n^{-1/2}|\boldsymbol{u}^T\boldsymbol{x}_i|}sds * I(n^{-1/2}|\boldsymbol{u}^T\boldsymbol{x}_i| < \eta)\} \quad (34)$$

$$\le \{f(0)+\kappa\}E|\boldsymbol{u}^T\boldsymbol{x}_i|^2 \quad (35)$$

The terms in (35) converge to 0 as $\eta \to 0$. This implies that R is dominated by the given function. It follows that as $n \to \infty$, $Var(\sum_{i=1}^n Z_{ni}) = \sum_{i=1}^n Var(Z_{ni}) \le nE(Z_{ni}^2(\boldsymbol{u})) \to 0$. Hence, $\sum_{i=1}^n\{Z_{ni}(\boldsymbol{u}) - E[Z_{ni}(\boldsymbol{u})]\} = o(1)$. Furthermore,

$$E(\sum_{i=1}^{n}Z_{ni}(\boldsymbol{u})) = nE[Z_{ni}(\boldsymbol{u})] \quad (36)$$

$$= nE\{\int_0^{n^{-1/2}\boldsymbol{u}^T\boldsymbol{x}_i}[F(s) - F(0)]ds\} \quad (37)$$

$$= E\int_0^{n^{-1/2}}\boldsymbol{u}^T\boldsymbol{x}_i s f(0)ds\} + o(1) \quad (38)$$

$$= 0.5f(0)\boldsymbol{u}^T(\boldsymbol{x}_i\boldsymbol{x}_i^T)\boldsymbol{u} + o(1) \quad (39)$$

because

$$P\{n^{-1/2}\max(|\boldsymbol{u}^T\boldsymbol{x}_1|,\ldots,|\boldsymbol{u}^T\boldsymbol{x}_n| > \eta^*)\} \tag{40}$$

$$\leq nP\{|\boldsymbol{u}^T\boldsymbol{x}_1| > \eta^* n^{1/2}\} \tag{41}$$

$$\leq \frac{1}{(\eta^*)^2}E\{|\boldsymbol{u}^T\boldsymbol{x}_1|^2 I(|\boldsymbol{u}^T\boldsymbol{x}_1| > \eta^* n^{1/2})\} \to 0 \tag{42}$$

Thus, because of the Law of Large Numbers, it follows that $\sum_{i=1}^n Z_{ni}(\boldsymbol{u}) \to_p \frac{1}{2}f(0)\boldsymbol{u}^T\boldsymbol{\Sigma}\boldsymbol{u}$, which is a quadratic function in $\boldsymbol{u}$. Therefore, the second part of (31) converges to $f(0)\boldsymbol{u}^T\boldsymbol{\Sigma}\boldsymbol{u}$ in probability. Hence, when C is sufficiently large, the second term of (28) dominates both the first part of (28) and the last term in (26). This implies (21) and completes the proof.

7.2 Proof of Theorem 2

First, assume all conditions from the proof of theorem 1 are true. Using an argument from Bloomfield and Steiger (1983), it follows that $Q(\boldsymbol{\beta})$ is piecewise linear and reaches the minimum at some breaking point. Take the first derivative of $Q(\boldsymbol{\beta})$ at any differentiable point $\tilde{\boldsymbol{\beta}}$ with respect to $\boldsymbol{\beta}_j$, $j = p_0 + 1, \ldots, p$, to obtain:

$$n^{-1/2}\frac{\partial Q(\tilde{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}_j} = -n^{-1/2}\sum_{i=1}^n \text{sgn}(y_i - \boldsymbol{x}_i^T\tilde{\boldsymbol{\beta}})x_{ik} + \sqrt{n}\lambda_k\frac{\hat{\boldsymbol{\beta}}_{\boldsymbol{b}}}{||\hat{\boldsymbol{\beta}}_{\boldsymbol{b}}||_2} \tag{43}$$

where

$$\text{sgn}(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{if } x = 0 \\ -1 & \text{if } x < 0 \end{cases} \tag{44}$$

.

For any $\boldsymbol{\Delta} \in \mathbb{R}^p$, let

$$V(\boldsymbol{\Delta}) = n^{-1/2}\sum_{i=1}^n \boldsymbol{x}_i \text{sgn}(\epsilon_i - n^{-1/2}\boldsymbol{x}_i^T\boldsymbol{\Delta}). \tag{45}$$

By the Central Limit Theorem, it follows that

$$V(\boldsymbol{0}) = n^{-1/2}\sum_{i=1}^n \boldsymbol{x}_i \text{sgn}(\epsilon_i) \to_d N(\boldsymbol{0}, \boldsymbol{\Sigma}), \tag{46}$$

where $\to_d$ means 'convergence in distribution.' Because $n^{-1/2}\max\{|\boldsymbol{u}^T\boldsymbol{x}_i|\} = o(1)$ and because of lemma A.2 from Koenker and Zhao (1996), it follows that

$$\sup_{||\boldsymbol{\Delta}|| \leq M} |V(\boldsymbol{\Delta}) - V(0) + f(0)\boldsymbol{\Sigma}\boldsymbol{\Delta}| = o(1) \tag{47}$$

where M is any fixed number. Then, for any $\tilde{\boldsymbol{\beta}} = (\tilde{\boldsymbol{\beta}}_{\boldsymbol{a}}^T, \tilde{\boldsymbol{\beta}}_{\boldsymbol{b}}^T)^T$ such that $\sqrt{n}(\tilde{\boldsymbol{\beta}}_{\boldsymbol{a}} - \boldsymbol{\beta}_a) = O_p(1)$ and $|\tilde{\boldsymbol{\beta}}_{\boldsymbol{b}} - \boldsymbol{\beta}_b| \leq \epsilon_n = Mn^{-1/2}$,

$$n^{-1/2}\sum_{i=1}^{n}\text{sgn}(y_i - \boldsymbol{x}_i^T\tilde{\boldsymbol{\beta}}) - n^{-1/2}\sum_{i=1}^{n}\boldsymbol{x}_i\text{sgn}(\epsilon) + f(0)\boldsymbol{\Sigma}\boldsymbol{\Delta}^* = o(1) \tag{48}$$

where $\boldsymbol{\Delta}^* = \sqrt{n}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})$. Ultimately, this implies

$$n^{-1/2}\sum_{i=1}^{n}\boldsymbol{x}_i^T\text{sgn}(y_1 - \boldsymbol{x}_i^T\tilde{\boldsymbol{\beta}}) = o(1), \tag{49}$$

which, in turn, implies that the first term of (43) is $o(1)$. As for the second term of (43), note that if $\hat{\boldsymbol{\beta}}_{\boldsymbol{b}} \neq 0$, there exists a c such that $|\hat{\beta}_{bc}| = \max\{|\hat{\boldsymbol{\beta}}_{\boldsymbol{bc'}}| : 1 \leq c' \leq p_k\}$. Without loss of generality, we can assume $c = 1$, then we must have

$$\frac{|\hat{\boldsymbol{\beta}}_{\boldsymbol{b1}}|}{||\hat{\boldsymbol{\beta}}_{\boldsymbol{b}}||_2} \geq \frac{1}{\sqrt{p_k}} > 0. \tag{50}$$

Note that $\sqrt{n}\lambda_k \geq \sqrt{n}b_n \to \infty$. This implies that $\frac{\sqrt{n}\lambda_k\hat{\boldsymbol{\beta}}_{\boldsymbol{bc}}}{||\hat{\boldsymbol{\beta}}_{\boldsymbol{b}}||_2}$ dominates the first term in (43) with probability tending to 1. This means (43) cannot be true as long as the sample size is sufficiently large. Hence, we can conclude that with probability tending to 1, $\|\hat{\boldsymbol{\beta}}_{\boldsymbol{b}}\|$ must be undifferentiable. Therefore, $\hat{\boldsymbol{\beta}}_{\boldsymbol{b}}$ has to be exactly zero.

7.3 Proof of Theorem 3

With theorem 1 and 2, theorem 3 implies that the group LAD-LASSO estimator is robust against heavy-tailed errors, because the $\sqrt{n}$-consistency of $\hat{\boldsymbol{\beta}}_a$ is established without making any moment assumptions on the regression error. Also, it implies that the resulting estimator has the same asymptotic distribution as the group LAD-LASSO estimator obtained under the true model establishing the oracle property of the estimator. Combining theorem 1 and 3, we know that $\hat{\boldsymbol{\beta}}_{\boldsymbol{k}} \neq 0$ for $k_0 < p_0$ and $\hat{\boldsymbol{\beta}}_{\boldsymbol{k}} = 0$ for $k_0 > p_0$.

For any $\boldsymbol{v} = (v_1, \ldots, v_{p_0})^T \in \mathbb{R}^{p_0}$, let $S_n(\boldsymbol{v}) = Q(\boldsymbol{\beta}_a + n^{-1/2}\boldsymbol{v}, 0) - Q(\boldsymbol{\beta}_a, 0)$. Then,

$$S_n(\boldsymbol{v}) = \sum_{i=1}^{n}\{|y_i - \boldsymbol{x}_{ia}\boldsymbol{\beta}_a - n^{-1/2}\boldsymbol{v}^T\boldsymbol{x}_{ia}| - |y_i - \boldsymbol{x}_{ia}^T\boldsymbol{\beta}_a| + n\sum_{j=1}^{p_0}\lambda_j\{|\beta_j + n^{-1/2}\boldsymbol{v}_j| - |\boldsymbol{\beta}_j|\} \tag{51}$$

where $\boldsymbol{x}_{ia} = (x_{i1}, \ldots, x_{ip_0})^T$. Similar to the proof of theorem 1, the first term of (51), such that (51) is separated by the $+$, converges in distribution to $\boldsymbol{v}^T\boldsymbol{W}_a + f(0)\boldsymbol{v}^T\boldsymbol{\Sigma}\boldsymbol{v}$, where $\boldsymbol{W}_a$ is a p_0-dimensional normal random vector with mean $\boldsymbol{0}$ and variance matrix $\boldsymbol{\Sigma}_a$. Also, the absolute value of the second term of (51), which can be denoted by $**$ is constrained by the following

$$|**| \leq \sqrt{n}a_n\sum_{j=1}^{p_0}|\boldsymbol{v}_j| \to 0 \tag{52}$$

Using the results from theorem 2 and remark 1 from Davis (1992), the central limit theorem follows, which completes the proof of theorem 3.