Robust Variable Selection Methods for Grouped Data

by

Kristin Lee Seamon Lilly

A dissertation submitted to the Graduate Faculty of
Auburn University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Auburn, Alabama
August 1, 2015

Keywords: Group LASSO, Robust variable selection, Multiple regression

Approved by

Nedret Billor, Professor of Mathematics and Statistics
Ash Abebe, Associate Professor of Mathematics and Statistics
Peng Zeng, Associate Professor of Mathematics and Statistics
George Flowers, Dean of the Graduate School

Abstract

When predictor variables possess an underlying grouping structure in multiple regression, selecting important groups of variables is an essential component of building a meaningful regression model. Some methods exist to perform group selection, but do not perform well when the data include outliers. Four methods for robust variable selection of grouped data, based on the group LASSO, are presented: two regular methods and two adaptive methods. For each of the two methods in the regular and adaptive groups, one method works well for data with outliers in the y-direction, and the other method works well for data with outliers in both the x- and y- directions. The effectiveness of each of these methods is illustrated with an extensive simulation study and a real data example.

Keywords: Group LASSO, Robust variable selection, Multiple regression

Acknowledgments

First, I would like to thank my adviser, Dr. Nedret Billor. Without her guidance and support, I would never have completed this dissertation. She always gave me encouragement when I needed it and was always there for a quick chat, whether it be about statistics research or an ongoing tennis tournament.

I would also like to thank the committee members, Dr. Ash Abebe and Dr. Peng Zeng, for their helpful comments on this dissertation and their thorough teaching throughout my time at Auburn.

Next, I would like to thank my family and friends for being a constant source of inspiration and strength, especially when I felt like I couldn't keep going.

Lastly, I would like to thank my husband, Brian, for absolutely everything.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

In regression analysis, variable selection is an important problem. Initially, there may be a large number of explanatory variables in the model. Including more predictors than necessary in the model can result in poor prediction accuracy, and having fewer predictors than needed can increase biases in parameter estimation and prediction results. In addition to these considerations, outliers in the data can also be problematic when performing estimation and variable selection. Therefore, robust regression methods should be utilized in such cases.

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

In this dissertation, we propose four robust methods to simultaneously perform parameter estimation and group variable selection. The first two are regular-type group variable selection methods, where there is a tuning parameter applied to all the parameters. The first performs well in the presence of outliers in the y-direction, and the second excels in the presence of outliers in both the y-direction and the x-direction. The second two are adaptive-type group variable selection methods, where a tuning parameter is applied to each individual group. Similarly to the first two methods, there is an adaptive-type method that is resistant to outliers in the y-direction, and another adaptive-type method that is resistant to outliers in both the x- and y-direction. The second two adaptive-type group variable

selection methods exhibit some nice statistical properties, which will be proven. To show the effectiveness of all of these new methods, we present simulation studies and a real data example.

1.1 Problem Description

The multiple regression model involves modeling one response variable as a function of more than one predictor variable. This model can be written mathematically as:

$$y_i = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + \ldots + x_{ip}\beta_p + \varepsilon_i \tag{1.1}$$

Here, y_i is the response variable, $x_{i1}, \ldots, x_{ip}$, are the predictor variables, β_j's are the regression coefficients, and ε_i's are the error terms for $i = 1, \ldots, n$ and $j = 1, \ldots, p$. Alternatively, these equations can be rewritten in matrix form:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{1.2}$$

where $\boldsymbol{y}$ is an n x 1 vector of responses, $\boldsymbol{X}$ is the n x p matrix of predictors, $\boldsymbol{\beta}$ is the p x 1 vector of regression coefficients, and $\boldsymbol{\varepsilon}$ is the n x 1 vector of random errors. Assume from now on that the response has been centered, and the predictors are standardized such that there need not be an intercept.

The assumptions for the standard multiple regression model include: that there is an approximate linear relationship between the response and predictor variables, and the errors are independent (uncorrelated) and are normally distributed with mean 0 and constant variance σ^2. When these assumptions are fulfilled, the data are ideal. In this case, the ordinary least square (OLS) estimators for the regression coefficients $\hat{\boldsymbol{\beta}}$ can be found, which are the best linear unbiased estimators, a result following from the Gauss-Markov theorem. However, when the errors do not follow the normal distribution or come from a mixture distribution, the least squares estimates can exhibit high bias in the presence of observations

that deviate from a majority of the data points. As a result, choosing a"good" estimation technique can be difficult, depending on the type of outlying observations. In particular, we would like to apply a robust estimation and variable selection methods to grouped data for a couple of different situations of outliers.

1.1.1 Why Variable Selection is Needed

Variable selection is useful for two purposes: interpretation and prediction. Having fewer predictors in the model results in a model that is easier to understand. Patterns and relationships between the predictors and response are easier to explain. With regard to prediction performance, there is a tradeoff. Including more predictors increases the prediction performance, since there is more known information being taken into account when making a prediction. This leads to having more accurate predictions. Having less predictors in the model decreases the variance of the regression model, leading to more precise prediction. However, the results from the prediction can be biased. A good regression model found using variable selection tries to find a balance between interpretation and prediction by including not too few and not too many predictors.

1.1.2 Why Robustness is Needed

Outliers can also cause problems when performing variable selection. Traditional variable selection methods, such as forward selection, backwards elimination, and stepwise regression, are based on the OLS estimators; consequently, these methods are sensitive to outliers and also lead to unstable models, which would cause poor prediction results. Shrinkage methods also exist, which perform variable selection by shrinking unnecessary predictors to zero, effectively eliminating them from the regression model. Nevertheless, shrinkage methods can be badly affected by outliers as well, if they are based on the least squares penalty function. Thus, a robust method must be used in order to build more accurate linear models to use for prediction or estimation purposes.

1.2 Overview of Dissertation

Chapter 2 is a review of the literature, including a discussion of existing variable selection methods and a review of group variable selection methods, which includes the group LASSO. In Chapter 3, we propose two new methods for robust variable selection with grouped data with one tuning parameter and discuss their properties. In Chapter 4, we propose two adaptive robust group variable selection methods and prove some statistical properties. Chapter 5 includes simulation studies and an application on a real data set. Chapter 6 is a summary of the dissertation.

Chapter 2

Literature Review

In this chapter, we review existing variable selection methods. These methods include forward selection, backwards elimination, and stepwise regression, as well as shrinkage methods based on the least absolute shrinkage and selection operator (LASSO). We also review some robust regression techniques involved in estimation and variable selection.

2.1 Classical Variable Selection Methods

In this section, we discuss classical variable selection techniques. These methods include forward selection, backwards elimination, and stepwise regression.

The least squares estimators (LSE) are named as such because they minimize the sum of the squares of the differences between the actual observations and the predicted values. The least squares estimators $\hat{\boldsymbol{\beta}}$ minimize

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{n} \varepsilon_i^2 = \boldsymbol{\varepsilon}^T \boldsymbol{\varepsilon} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}). \tag{2.1}$$

To minimize the equation, take the derivative, set equal to zero, and solve. This results in the following expression:

$$\frac{\partial \mathbf{S}}{\partial \boldsymbol{\beta}}|_{\hat{\boldsymbol{\beta}}} = -2\mathbf{X}^T\mathbf{y} + 2\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{0}. \tag{2.2}$$

Therefore, the least-squares estimator of $\boldsymbol{\beta}$ is:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}. \tag{2.3}$$

For each coefficient $\hat{\beta}_j$ in $\hat{\boldsymbol{\beta}}$, the standard error, $\text{se}(\hat{\beta}_j)$ is computed, and a test statistic $\mathbf{t}_j = \frac{\hat{\beta}_j}{\text{se}(\hat{\beta}_j)}$ can be calculated. This test statistic can be used to determine if the corresponding coefficient is statistically significant, and with this information, various variable selection methods can be used to build a regression model.

2.1.1 Forward Selection

Forward selection starts with no predictors in the model and sequentially adds significant variables to build an appropriate model. Forward selection is performed with the following steps:

1. Begin with no predictors in the model (other than the intercept).
2. Set an upper limit on the significance level α for entry into the regression model.
3. Calculate the test statistic and p-value for adding each individual predictor variable to the model.
4. Add the most significant predictor with a significance level less than or equal to the previously set α to the model.
5. Calculate the test statistic and p-value for adding the individual candidate predictors to the model, given that the model already includes the intercept and the variable added in the previous step.
6. The most significant predictor variable with a p-value less than or equal to α is then added to the model.
7. Repeat until the next predictor variables that can be added to the model all have a significance level greater than α.

2.1.2 Backwards Elimination

Backwards elimination works in the opposite direction of forward selection. It begins with all predictors in the model and eliminates those that are not significant. The algorithm for backwards elimination is below:

1. Start with all predictors in the model (including the intercept).
2. Set an lower limit on the significance level α for deletion from the regression model.
3. Remove the predictor with the largest p-value greater than α.
4. Next, build the model with the remaining predictors.
5. Remove the predictor with the largest p-value greater than α from the remaining predictor variables, given that the predictor from the previous step is already removed from the model.
6. Repeat until the potential predictor variables to be removed from the model all have a significance level less than α.

2.1.3 Stepwise Regression

Stepwise regression is a combination of forward selection and backwards elimination. It begins with no predictors in the model, except for the intercept. Then, the predictors are potentially added and then reevaluated for potential elimination, depending on how the addition of other predictors may have changed the existing predictors' test statistic value and p-value. This algorithm can be described by the following:

1. Start with no predictors in the model (except for the intercept).
2. Set a significance level α_{IN} for entry into the model and a significance level α_{OUT} for removal from the model.

3. Perform one step of forward selection.

4. Run model to determine significance of variables still in the model.

5. Perform one step of backwards elimination.

6. Run model to determine significance of variables still in the model.

7. Repeat the sequence of forward selection and backwards elimination until the predictors that can be added have a significance level greater than α_{IN} and the predictors that can be removed have a significance level less than α_{OUT}.

2.1.4 Criticisms

The methods of forward selection, backwards elimination, and stepwise regression have been criticized as valid variable selection methods when the assumptions of the least squares estimators have been violated. Some problems include underestimating the standard errors of the regression coefficients, which can result in inflating test statistics, causing p-values to be too low [13]. Also, as a result, parameter estimates can be overestimated. Another shortcoming is the discrete nature of the aforementioned procedures. A variable is either included or excluded at one step. There is no continual process of adding or removing variables. This is an advantage of the following procedures.

2.2 LASSO Estimation Based Techniques

In this section, we describe shrinkage variable selection methods. In particular, this topic is about the least absolute shrinkage and selection operator (LASSO), and its derived robust counterparts.

2.2.1 LASSO

The LASSO [21] was proposed as a compromise between subset selection and ridge regression. Subset selection, such as stepwise regression, is a discrete procedure, which

can result in highly variable models with small changes in the data. Ridge regression is a continuous procedure that is also an estimation method; it shrinks coefficients and is more stable than subset selection. The LASSO will shrink coefficients and set some exactly to 0. Hence, LASSO is both a shrinkage and variable selection method.

The LASSO estimates are originally obtained by minimizing:

$$\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p}\beta_j x_{ij})^2 \text{ subject to } \sum_{j=1}^{p}|\beta_j| \leq t \tag{2.4}$$

where $t \geq 0$ is a tuning parameter. An equivalent way of writing (2.4) is using the Lagrangian form of:

$$\frac{1}{2}\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda\sum_{j=1}^{p}|\beta_j| \tag{2.5}$$

where $\lambda \geq 0$ is the shrinkage parameter, which controls the degree of shrinkage on the estimates. The shrinkage parameter is designed such that the larger it is, the more shrinkage that is applied to the regression coefficients; thus, the larger λ is, the more regression coefficients that will be zero. It is typically chosen using k-fold general cross-validation in order to minimize an estimate of the model error or prediction error, depending on the user's choice. The LASSO solutions do not have a closed form. The constraint of the LASSO makes the solutions of (2.4) nonlinear in the y_i's. The solution to this equation is classified as a quadratic programming problem with an added constraint.

The LASSO was also motivated by a shortcoming in the nonnegative garotte, whose solutions depend explicitly on the least squares estimates. When the predictors are highly correlated, the LSE behave poorly, which in turn affects the garotte solutions, which will also behave badly. The LASSO avoids this problem by not relying explicitly on least squares estimates. However, the LASSO still suffers in the case of predictors with severe multicollinearity; the LASSO is ideal for cases with little to no correlation between predictors [6]. It has been shown that the oracle property does not hold for the LASSO [8]. The LASSO

does automatic variable selection because of the singularity of the L_1 penalty at the origin; however, when the regression coefficients are large, the estimates can be biased when using the LASSO procedure.

2.2.2 Adaptive LASSO

An extension of the LASSO is the adaptive LASSO [27]. Instead of one tuning parameter, there is a tuning parameter for each coefficient. This idea arose from the notion that the LASSO requires that each coefficient is equally penalized in the L_1 penalty, and that may not necessarily be the best way to treat the predictors when they don't all contribute to the regression model. Hence, the adaptive lasso was derived, where each regression coefficient is penalized differently with its own tuning parameter. The adaptive LASSO is designed to minimize the following equation:

$$\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} w_j|\beta_j| \tag{2.6}$$

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, (2.6) can be written as:

$$\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \sum_{j=1}^{p} \lambda_j|\beta_j| \tag{2.7}$$

The solution to (2.6) is 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. One useful algorithm involves a modification of the least angle regression algorithm (LARS algorithm) used by Efron et al [7]. 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. It has been shown by Zou [27] that the oracle properties, including consistency and sparsity, do hold for the adaptive LASSO method. The adaptive LASSO is able to find sparse solutions more efficiently than the LASSO.

2.2.3 LAD-LASSO

When there are outliers in the response, the LASSO estimates, both the regular LASSO and the adaptive LASSO, can result in an unstable model. In order to effectively perform estimation and variable selection in this case, the LAD-LASSO has been proposed [23]. The LAD-LASSO is based on the least absolute deviation (LAD) estimator, which has been shown to perform well in situations where the data has outlying observations in the y-direction. Instead of minimizing the squared differences of the LSE, whose objective function is shown below,:

$$\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p}\beta_j x_{ij})^2 \tag{2.8}$$

the LAD estimator minimizes the absolute differences:

$$\sum_{i=1}^{n}|y_i - \sum_{j=1}^{p}\beta_j x_{ij}|. \tag{2.9}$$

It is known that the LAD estimators have $\sqrt{n}$-consistency and asymptotic normality under certain conditions [2][17], which is why it would be useful to combine the LAD estimation criterion with other methods. Thus, the LAD-LASSO estimators are designed to minimize:

$$\sum_{i=1}^{n}|y_i - \sum_{j=1}^{p}\beta_j x_{ij}| + \lambda\sum_{j=1}^{p}|\beta_j| \tag{2.10}$$

where $\lambda \geq 0$ is again a shrinkage parameter. When using the same shrinkage parameter for all regression coefficients, the estimators that result can have some issues. For example, in this case, the estimators can be subject to bias [8]. Therefore, we consider the following LAD-LASSO criterion, which combines Zou's adaptive LASSO, to perform consistent variable selection, with LAD regression, to perform robust estimation in the presence of heavy-tailed errors:

$$\sum_{i=1}^{n} |y_i - \sum_{j=1}^{p} \beta_j x_{ij}| + n \sum_{j=1}^{p} \lambda_j |\beta_j| \tag{2.11}$$

The above adaptive LAD-LASSO includes a shrinkage parameter for each regression coefficient. In doing so, several statistical properties can be proven to exist for the resulting estimators. The estimated shrinkage parameter $\hat{\lambda}_j$ is found to be $\hat{\lambda}_j = \frac{\log(n)}{n|\tilde{\beta}_j|}$, where $\tilde{\beta}$ are the unpenalized LAD estimators. The estimated shrinkage parameter results from an idea from Tibshirani [21]. The LAD-LASSO estimator can be seen as a Bayesian estimator with each regression coefficient following a double-exponential prior with location parameter equal to 0 and scale parameter equal to $n\lambda_j$, which leads to the equation $\lambda_j = \frac{1}{n|\beta_j|}$. In order to guarantee both consistency and sparsity, the choice of $\hat{\lambda}_j$ must be $\hat{\lambda}_j = \frac{\log(n)}{n|\tilde{\beta}_j|}$, according to Wang et al. [23].

Under certain conditions, the adaptive LAD-LASSO method fulfills the oracle property, including estimation consistency and sparsity. The consistency of the estimators is a particular nice property, because it implies that the adaptive LAD-LASSO can identify the true model consistently [23]. The computation is easily found using an augmented dataset as described by Wang et al., which is implemented later for a proposed method in Chapter 3 and 4. The method involves defining $\{(y_i^*, \boldsymbol{x}_i^*)\}$ with $i = 1, \ldots, n+p$, where $\{(y_i^*, \boldsymbol{x}_i^*)\} = (y_i, \boldsymbol{x}_i)$ for $1 \leq i \leq n$ and $(y_{n+j}^*, \boldsymbol{x}_{n+j}^*) = (0, n\lambda_j \boldsymbol{e}_j)$ for $1 \leq j \leq p$ such that $\boldsymbol{e}_j$ is a p-dimensional vector with the jth component equal to 1 and all others are equal to 0.

2.2.4 WLAD-LASSO

It is known that if there are outliers in the predictors, but not in the response, the LAD estimators will be outperformed by the LSE. As a result, the LAD-LASSO estimators will also be outperformed by the LASSO estimators in the case of outliers in the predictor space. To account for both outliers in the response and the predictors, the weighted least absolute

deviation (WLAD) method [12] can be combined with the LASSO to create the WLAD-LASSO [1]. The WLAD estimators add an extra weight to the function to be minimized, which will down weigh the outliers in the x-direction:

$$\sum_{i=1}^{n} w_i |y_i - \sum_{j=1}^{p} \beta_j x_{ij}| \tag{2.12}$$

Here, the weights are w_i for $i = 1, \ldots, n$ and are determined by using robust measures. Therefore, the WLAD-LASSO method minimizes the following:

$$\sum_{i=1}^{n} w_i |y_i - \sum_{j=1}^{p} \beta_j x_{ij}| + \lambda \sum_{j=1}^{p} |\beta_j| \tag{2.13}$$

Here, $\lambda \geq 0$ is the shrinkage parameter found by general cross-validation, like before for the LASSO and LAD-LASSO. The weights w_i are found as robust distances, such that more extreme outliers in the x-direction are assigned smaller weights. For this algorithm, the weights are found as follows:

1. For each $\boldsymbol{x}_i$ in $\boldsymbol{X}$ for $i = 1, \ldots, n$, calculate the robust location and scatter estimates, $\tilde{\mu}$ and $\tilde{\Sigma}$.

2. Compute the robust distances: $\text{RD}(\boldsymbol{x}_i) = (\boldsymbol{x}_i - \tilde{\mu})^T \tilde{\Sigma}^{-1} (\boldsymbol{x}_i - \tilde{\mu})$.

3. Calculate the weights $w_i = \min \left\{ 1, \frac{p}{\text{RD}(\boldsymbol{x}_i)} \right\}$ for $i = 1, \ldots, n$.

These weights are designed to decrease as the robust distances increase; hence, the resulting estimators are expected to be robust to outliers in both the x- and y-directions. Theoretically, because of the one tuning parameter λ for all of the regression coefficients, the oracle property does not hold for the WLAD-LASSO in (2.13). As a result, Arslan [1] proposes the adaptive WLAD-LASSO which minimizes the given equation:

$$\sum_{i=1}^{n} w_i |y_i - \sum_{j=1}^{p} \beta_j x_{ij}| + \sum_{j=1}^{p} \lambda_j |\beta_j| \tag{2.14}$$

The adaptive tuning parameters are chosen using general cross-validation, and Arslan [1] suggests $\hat{\lambda}_j = \frac{1}{|\hat{\beta}_j|^\nu}$, where $\nu > 0$. This follows from the same Bayesian logic that is used for the shrinkage parameter for the adaptive LAD-LASSO. The compuation of the adaptive WLAD-LASSO is very similar to that of the adaptive LAD-LASSO. First, calculate $\tilde{y}_i = w_i y_i$ and $\tilde{\boldsymbol{x}}_i = w_i \boldsymbol{x}_i$ for $i = 1, \ldots, n$. Next, define $\{(y_i^*, \boldsymbol{x}_i^*)\}$ for $i = 1, \ldots, n+p$, where $\{(y_i^*, \boldsymbol{x}_i^*)\} = (\tilde{y}_i, \tilde{\boldsymbol{x}}_i)$ for $1 \leq i \leq n$ and $(y_{n+j}^*, \boldsymbol{x}_{n+j}^*) = (0, n\lambda_j \boldsymbol{e}_j)$ for $1 \leq j \leq p$ such that $\boldsymbol{e}_j$ is a p-dimensional vector with the jth component equal to 1 and all others are equal to 0. It has been shown that these estimators possess the properties of $\sqrt{n}$-consistency, sparsity, and asymptotic normality, which together imply the estimators have the oracle property [1].

2.3 Other Regression Methods for Estimation and Selection

2.3.1 Least Angle Regression

Least angle regression (LARS) is a variable selection procedure that can be thought of as accelerated forward selection [7]. All of the regression coefficients are set to zero. Then, the predictor that is most correlated with the response is identified. Geometrically, the algorithm takes the largest step possible in the direction of the most correlated predictor until a second predictor has as much correlation with the current residual that the first predictor does. Then, the algorithm continues in a direction that is equiangular between the two variables until a third predictor enters the model, and so on. This direction is called the "least angle direction." Due to the structure of the algorithm, it is clear that LARS is negatively affected by multicollinearity.

2.3.2 Nonnegative Garotte

The LASSO was inspired by the nonnegative garotte [4]. The nonnegative garrote minimizes the following criterion:

$$\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} c_j \beta_j x_{ij})^2 \tag{2.15}$$

The c_j's are nonnegative factors subject to the following constraints: $c_j \geq 0$ and $\sum c_j \leq \lambda$. Essentially, the garotte takes the least squares estimates of the regression coefficients and scales them, using a nonnegative constant c_j. Because the sum of these constants is restrained by a tuning parameter, λ, this means that the least squares estimates will actually shrink.

The nonnegative garotte tends to have smaller prediction errors than any of the discrete subset selection methods and gives similar results to ridge regression when there are not many small nonzero coefficients, based on simulation studies by Breiman [4]. The garotte's main shortcoming involves its direct dependency on the LSE. In any situation where the LSE perform poorly, such as the case of multicollinearity, the garotte, as a result, would also perform badly.

2.3.3 Ridge Regression

Ridge regression [14] is an estimation method which minimizes:

$$\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \tag{2.16}$$

The ridge procedure is not a variable selection method, but an estimation method. It still shrinks regression coefficients toward zero, but they never quite become exactly equal to zero. This has the effect of increasing the variance of the estimates; however, it does introduce some bias as well. Ridge estimators perform well in the presence of multicollinearity. When there is multicollinearity, the ridge estimators have a variance that is well constrained (although the estimators suffer a small amount of bias), unlike the LSE, which have a variance that becomes inflated (while the estimators themselves remain unbiased) [14].

2.3.4 Elastic Net

The elastic net is a regularization and variable selection method proposed by Zou and Hastie [28]. It is designed to minimize:

$$\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda_2 \sum_{j=1}^{p}\beta_j^2 + \lambda_1 \sum_{j=1}^{p}|\beta_j| \tag{2.17}$$

The penalty involves a convex combination of the LASSO and ridge penalty. The algorithm for the elastic net involves the naive version, which does a dual-type shrinkage. First, it finds the ridge regression coefficients, and then performs a LASSO shrinkage. To correct for the double shrinkage, there is a correction to be applied to the coefficients from the naive version, which is scaling those coefficients by $(1 + \lambda_2)$.

2.3.5 Bridge Regression

Bridge regression is another variable selection method, which was proposed by Frank and Friedman [9]. This method minimizes the following criterion:

$$\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p}|\beta_j|^\gamma \tag{2.18}$$

When $\gamma = 1$ and $\gamma = 2$, the bridge minimization criterion reduces to the LASSO and ridge regression minimization criterion, respectively. If $0 < \gamma \leq 1$, the bridge estimators produce sparse models and are well suited to a case where variable selection is needed when the predictors exhibit multicollinearity. With the appropriate shrinkage parameter choice, the bridge estimators exhibit the oracle property [19].

2.4 Group Variable Selection Methods

In this section, we review existing group variable selection methods and their properties. These methods include the group LASSO, the group LARS, the group bridge, and the adaptive group LASSO.

2.4.1 Group LASSO

In some real data applications, predictors can be grouped in a natural way such that selecting groups of variables is of interest. Genetic data sometimes has this property. For example, data from genes can be grouped such that a group of genes correspond to the same biological pathway. The group LASSO method [26] is ideal for these kind of situations; it will shrink entire groups of predictors to 0 or estimate the regression coefficients for the entire group. The regression coefficients of groups will either all be 0 or all be nonzero.

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 $\sum_{k=1}^{K} p_k = p$. Within each group k, there are j predictors for $j = 1, \ldots, p_k$. 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:

$$\frac{1}{2}\sum_{i=1}^{n}(y_i - \sum_{k=1}^{K} \boldsymbol{x}_{ik}\boldsymbol{\beta}_k)^2 + n\lambda \sum_{k=1}^{K} ||\boldsymbol{\beta}_k||_2 \tag{2.19}$$

where $\lambda \geq 0$ is a tuning parameter, y_i is the ith response, $\boldsymbol{x}_{ik}$ is a 1 x p_k vector of predictors in the kth group for the ith observation, and $\boldsymbol{\beta}_k$ is a p_k x 1 vector of regression coefficients for group k. As for the criterion above, for each group of predictors, minimize the sum of the squared distances, while simultaneously shrinking unimportant groups with the LASSO penalty (the L_2 norm in this case). The tuning parameter λ controls the rate of shrinkage and can be chosen using cross-validation. In particular, the Yuan and Lin [26] use a shrinkage parameter based on an approximate C_p-type criterion.

The LASSO method of simultaneous estimation and selection is ideal for predictors with little to no multicollinearity, but not for data with outliers. In particular, because it uses the LSE, the group LASSO performs poorly in terms of robustness [16].

The computation of the group LASSO is based on the shooting algorithm [11]. Originally, this method was proposed for the LASSO method, but was adapted for the group LASSO [26].

First, rewrite (2.19) with respect to the groups:

$$\frac{1}{2}||\boldsymbol{Y} - \sum_{k=1}^{K} \boldsymbol{X}_k \boldsymbol{\beta}_k|| + n\lambda \sum_{k=1}^{K} ||\boldsymbol{\beta}_k||_2 \tag{2.20}$$

where $\boldsymbol{Y} \sim$ n x 1 vector of responses, $\boldsymbol{X}_k \sim$ n x p_k matrix of predictors from group k, $\boldsymbol{\beta}_k \sim$ 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 LASSO involves applying the following equation iteratively with the groups for $k = 1, \ldots, K$:

$$\boldsymbol{\beta}_k = \left(1 - \frac{\lambda\sqrt{p_k}}{\|\boldsymbol{S}_k\|}\right)_+ \boldsymbol{S}_k \tag{2.21}$$

where $\boldsymbol{S}_k = \boldsymbol{X}_k^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}_{-k})$ with $\boldsymbol{\beta}_{-k} = (\boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_{k-1}^T, \boldsymbol{0}^T, \boldsymbol{\beta}_{k+1}^T, \ldots, \boldsymbol{\beta}_K^T)$, the $\boldsymbol{\beta}$ vector without coefficient vector $\boldsymbol{\beta}_k$, and $\|\boldsymbol{\eta}\| = (\boldsymbol{\eta}^T\boldsymbol{\eta})^{1/2}$. Choose initial $\boldsymbol{\beta}_k$ for $k = 1, \ldots, K$ to be the LSE. 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 [26].

2.4.2 Group LARS

Yuan and Lin [26] also proposed the group LARS. It is best described by considering its algorithm. When all the groups have the same number of predictors ($p_1 = p_2 = \ldots = p_k$), one may define the angle $\theta(r, X_k)$ between an n-vector r and a group represented by X_k as the angle between the vector r and the space that is spanned by the column vectors of X_k. The given angle does not depend on the set of orthonormal contrasts representing the grouping, and it actually ends up being the same as the angle between r and the projection of

r in the space that is spanned by the columns of X_k. Thus, $\cos^2\theta\{(r, X_k)\}$ is the proportion of the total variation sum of squares in r that is explained by the linear regression on X_k.

Beginning with all regression coefficient vectors equal to 0, the group LARS algorithm finds the group that has the smallest angle with $\boldsymbol{y}$ and proceeds in the direction of the projection of $\boldsymbol{y}$ on the space that is spanned by the factor until some other group has as small an angle with the current residual. Group LARS will proceed in the direction of the projection of the current residual on the space that is spanned by those two groups; it will continue in that direction until a third group has an equally small angle and then slightly change direction, and so on. There is a small adjustment when all the group sizes are not equal.

2.4.3 Group Nonnegative Garotte

Another group variable selection method is the group nonnegative garotte, which is based on the nonnegative garotte [4]. The group nonnegative garotte takes the LSE, arranged into vectors according to the grouping information, and scales the coefficients in each group vector by a constant factor. This minimization criterion can be written as the following:

$$\frac{1}{2}\sum_{i=1}^{n}(y_i - \sum_{k=1}^{K}(\sum_{j=1}^{p_k} c_j x_{ijk}\beta_{jk}))^2 + \lambda\sum_{k=1}^{K} p_k c_k \tag{2.22}$$

The algorithm for the group nonnegative garotte is similar to the algorithm for the group LARS. Because the group nonnegative garotte relies explicitly on the least squares estimates, it is, therefore, not considered to be a robust method [16]. The next method, the adaptive group LASSO, is preferable due to the introduction of an adaptive tuning parameter, allowing for an overall better fit of shrinkage to the groups separately.

2.4.4 Adaptive Group LASSO

Wang and Leng [24] saw a need for combination of the adaptive LASSO method with the group LASSO method. Because both the LASSO and the group LASSO apply the same

amount of shrinkage to all of the regression coefficients, those two methods are not consistent in terms of model selection [8]. Efficiency can also suffer due to the one shrinkage parameter [27]. As a result, an adaptive tuning parameter is introduced, which assigns a different tuning parameter for each group, allowing the shrinkage to vary from group to group. The adaptive group LASSO, like the group LASSO, will shrink insignificant groups to 0 and estimate significant groups to be nonzero. The adaptive group LASSO criterion to minimize is the following:

$$\frac{1}{2}\sum_{i=1}^{n}(y_i - \sum_{k=1}^{K}\boldsymbol{x}_{ik}\boldsymbol{\beta}_k)^2 + n\sum_{k=1}^{K}\lambda_k\|\boldsymbol{\beta}_k\|_2 \tag{2.23}$$

where $\lambda_k \geq 0$ is an adaptive tuning parameter, y_i is the ith response, $\boldsymbol{x}_{ik}$ is a 1 x p_k vector of predictors in the kth group for the ith observation, and $\boldsymbol{\beta}_k$ is a p_k x 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 [24] is:

$$\lambda_k = \frac{\lambda}{||\hat{\boldsymbol{\beta}}_k||_2^{\gamma}} \tag{2.24}$$

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 [24].

Chapter 3

Regular Robust Group Variable Selection

In this chapter, we propose two robust group variable selection methods. The first is the group LAD-LASSO, which is based on the LAD-LASSO. The second is the group WLAD-LASSO, which is based on the WLAD-LASSO.

3.1 Group LAD-LASSO

The group LASSO will identify important groups and estimate their regression coefficients and shrink unimportant groups such that are of their regression coefficients are 0. It is known that the LASSO estimates can be sensitive to outliers, because of the dependency of (2.19) on the OLS criterion. 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:

$$Q(\boldsymbol{\beta}) = \frac{1}{2}\sum_{i=1}^{n}|y_i - \sum_{k=1}^{K}\boldsymbol{x}_{ik}\boldsymbol{\beta}_k| + n\lambda\sum_{i=1}^{K}||\boldsymbol{\beta}_k||_2 \tag{3.1}$$

which is the minimization criterion for the group LAD-LASSO to simultaneously estimate significant groups and shrink nonsignificant groups to 0. The penalty is the typical L_2 norm. Define $\boldsymbol{x}_{ik}$ to be the ith row of predictors in group k; that is, $\boldsymbol{x}_{ik}$ is a 1 x p_k vector, where p_k is the number of predictors in group k. Note that if $p_k = 1$ for all k, then (3.1) reduces to the LAD-LASSO equation in (2.10). The vector $\boldsymbol{\beta}_k$ is a p_k x 1 vector of regression coefficients. The computation for our simulation is done using the package *grpreg* in the statistical program R [3].

3.1.1 Tuning Parameter Selection

The tuning parameter, λ, must be selected carefully. It should be chosen such that it is large enough to have a desired shrinking effect for insignificant groups, but it should also be chosen such that it is small enough that all the groups 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 [8] [21]. In this case, we use k-fold cross-validation after modifying the objective function to be that of the group WLAD-LASSO to find the best value of λ, such that the cross-validation error is minimized.

3.1.2 Theoretical Properties

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 [8].

3.2 Group WLAD-LASSO

As before, it is known that the LAD method is adapted to do well for the regression setting when there are outliers in the response; however, it has been studied and determined that the same method performs poorly in the prescense of outliers in the explanatory variables. In this case, an adjustment is made the minimization criterion of (2.9) to downweight those observations that are outliers in the predictors, and the weighted LAD (WLAD) method is used. The criterion for minimization for the WLAD method is the following:

$$Q(\beta) = \sum_{i=1}^{n} w_i |y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}| \tag{3.2}$$

where w_i are the weights assigned to downweight observations in the predictors for $i = 1, \ldots, n$. These 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.

Similarly to the LAD-LASSO, the WLAD method can be combined with the LASSO in order to minimize the following criterion:

$$Q(\beta) = \sum_{i=1}^{n} w_i |y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}| + n\lambda \sum_{j=1}^{p} |\beta_j| \tag{3.3}$$

where w_i is a positive weight assigned to each observation for $i = 1, \ldots, n$. The WLAD-LASSO [1] is ideally used for data with outliers in both the response and the predictors. Minimizing the LAD lessens the effect of outliers in the response, while the weights will relieve the effect of outliers in the explanatory variables. We would like to also extend this idea to an application with grouped predictors.

While the group LAD-LASSO method works well on data where there are outlying observations in the response, it does not do as well when there are outliers also in the predictors. As a result, we propose a small modification to the WLAD-LASSO criterion to extend it to grouped predictors:

$$Q(\boldsymbol{\beta}) = \frac{1}{2} \sum_{i=1}^{n} w_i |y_i - \sum_{k=1}^{K} \boldsymbol{x}_{ik} \boldsymbol{\beta}_k| + n\lambda \sum_{i=1}^{K} ||\boldsymbol{\beta}_k||_2 \tag{3.4}$$

where w_i is a positive weight assigned to each observation for $i = 1, \ldots, n$. The above is the objective function to be minimized for the group WLAD-LASSO. Denote $\boldsymbol{x}_{ik}$ to be the ith row of predictors corresponding to group k, and let $\boldsymbol{\beta}_k$ be a p_k x 1 vector of regression coefficients. The penalty is the typical L_2 norm.

3.2.1 Weights

In order to compute the estimators for the group WLAD-LASSO in (3.4), the weights must be calculated first. These weights will be based on a robust distance [15]. Given a set of points $x_1, \ldots, x_n$, the weights can be found with the following steps:

1. For each $\boldsymbol{x}_i$ in $\boldsymbol{X}$ for $i = 1, \ldots, n$, calculate the robust location and scatter estimates, $\tilde{\mu}$ and $\tilde{\Sigma}$.

2. Compute the robust distances: $\text{RD}(\boldsymbol{x}_i) = (\boldsymbol{x}_i - \tilde{\mu})^T \tilde{\Sigma}^{-1} (\boldsymbol{x}_i - \tilde{\mu})$.

3. Calculate the weights $w_i = \min\left\{1, \frac{p}{\text{RD}(\boldsymbol{x}_i)}\right\}$ for $i = 1, \ldots, n$.

One such set robust location and scatter estimates could be the MVE and MCD. For our simulations and real data example, we use the minimum covariance determinant (MCD) estimator of location and scatter. The method finds the $h(> \frac{n}{2})$ observations out of the n total observations whose classical covariance matrix has the lowest possible determinant. The raw MCD estimates of location and scatter are the average of the h points and their covariance matrix, respectively. The raw estimates are reweighted to increase the finite-sample efficiency, and these reweighted MCD estimates of location and scatter are used to help create the robust distances. This algorithm is utilized with the *rrcov* package in R [22].

Large values of $RD(\boldsymbol{x_i})$ indicate leverage points. For high leverage points, which are points that are considered outliers in the explanatory variables, these weights will be small, while for other points considered regular, the weights will be close to 1. The statistical software program R is used for simulations and analysis using the package *grpreg* [3].

3.2.2 Tuning Parameter Selection

The tuning parameter λ for group WLAD-LASSO is selected in the same way as for the group LAD-LASSO method, but now the consideration of the weight for each observation is also taken into account when calculating λ. Thus, 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.

3.2.3 Theoretical Properties

Similarly as for the group LAD-LASSO, because of the nature of the tuning parameter λ, the properties of consistency, sparsity, and the oracle property cannot be proven.

Chapter 4

Adaptive Robust Group Variable Selection Methods

In this section, we discuss the adaptive robust group variable selection methods, which include the adaptive group LAD-LASSO and the adaptive group WLAD-LASSO.

4.1 Adaptive Group LAD-LASSO

For the first proposal of an adaptive robust group variable selection method, we combine the adaptive tuning parameter from the adaptive group LASSO with the objective function of the group LAD-LASSO. With this mixture, we get the following objective function to minimize:

$$Q(\boldsymbol{\beta}) = \frac{1}{2}\sum_{i=1}^{n}|y_i - \sum_{k=1}^{K}\boldsymbol{x}_{ik}\boldsymbol{\beta}_k| + n\sum_{i=1}^{K}\lambda_k||\boldsymbol{\beta}_k||_2 \tag{4.1}$$

Define $\boldsymbol{x}_{ik}$ to be a 1 x p_k vector of predictors, where p_k is the number of predictors in group k, while $\boldsymbol{\beta}_k$ is a p_k x 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. This is done in R with a small modification to our code using the *grpreg* package for our simulations [3].

4.1.1 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 [24] and choose:

$$\lambda_k = \frac{\lambda}{||\tilde{\boldsymbol{\beta}}_k||_2^{\gamma}} \tag{4.2}$$

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 [24]. As a result, instead of calculating a λ_k for each group, this reduces to a one-dimension problem where we need only need to choose an appropriate λ. Some selection criteria for λ, suggested by Wang and Leng [24], are as follows:

$$C_p = \frac{||\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}||_2^2}{\hat{\sigma}^2} - n + 2 * df \tag{4.3}$$

$$GCV = \frac{||\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}||_2^2}{(1 - n^{-1} * df)^2} \tag{4.4}$$

$$AIC = \log(\frac{1}{n}||\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}||_2^2) + 2 * df/n \tag{4.5}$$

$$BIC = \log(\frac{1}{n}||\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}||_2^2) + \log(n) * df/n \tag{4.6}$$

The df are the degrees of freedom as defined in Yuan and Lin [26], given by:

$$df = \sum_{k=1}^{K} I\{||\hat{\boldsymbol{\beta}}_k||_2 > 0\} + \sum_{k=1}^{K} \frac{||\hat{\boldsymbol{\beta}}_{\boldsymbol{k}}||_2}{||\tilde{\boldsymbol{\beta}}_k||_2}(p_k - 1) \tag{4.7}$$

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 with software, we

use the default setting of choosing λ with the smallest value of the BIC criterion, which is the equation in (4.7).

4.1.2 Theoretical Properties

In order to establish a few theoretical properties, we need to make some important assumptions and define some notations. First, 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. 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 $Q(\boldsymbol{\beta}) = Q(\boldsymbol{\beta}_a, \boldsymbol{\beta}_b)$.

In addition to the above, we must make the following assumptions:

- The errors ε_i have continuous and positive density at the origin.
- The matrix $cov(\boldsymbol{x}_1) = \boldsymbol{\Sigma}$ exists and is positive definite.

We must also 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 4.1. *(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 4.2. *(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 4.3. *(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 shown in appendix A.

4.2 Adaptive Group WLAD-LASSO

Similarly, we can extend the adaptive tuning parameter to the group WLAD-LASSO to create the adaptive group WLAD-LASSO. By combining the adaptive LASSO with the group WLAD-LASSO, we will get a method with nice theoretical properties that is able to perform well in the presence of outliers in both the response and predictor space. The adaptive group WLAD-LASSO requires the minimization of the following criterion:

$$Q(\boldsymbol{\beta}) = \frac{1}{2}\sum_{i=1}^{n} w_i |y_i - \sum_{k=1}^{K} \boldsymbol{x}_{ik}\boldsymbol{\beta}_k| + n\sum_{i=1}^{K} \lambda_k ||\boldsymbol{\beta}_k||_2 \tag{4.8}$$

Define $\boldsymbol{x}_{ik}$ to be a 1 x p_k vector of predictors, where p_k is the number of predictors in group k, while $\boldsymbol{\beta}_k$ is a p_k x 1 vector of regression coefficients. The penalty is the typical L_2 norm. The tuning parameter is defined such that $\lambda_k \geq 0$. The objective function is a combination of the adaptive group LAD-LASSO with a weight w_i to downweight high leverage points.

4.2.1 Weights

The weights w_i are calculated the same as for the regular group WLAD-LASSO. This procedure can be found in section 3.2.1. Like before, we use the minimum covariance determinant (MCD) estimator of location and scatter.

Large values of $RD(\boldsymbol{x_i})$ will indicate high leverage points, which will be assigned a small weight close to 0. For points that are not considered outliers, their weights will be assigned such that they are close to 1. The weights and computation are found in R using the *grpreg* package [3] with a small modification.

4.2.2 Tuning Parameter Selection

For the adaptive group WLAD-LASSO, we stick with the earlier choice for the adaptive group LAD-LASSO and choose:

$$\lambda_k = \frac{\lambda}{||\tilde{\boldsymbol{\beta}}_k||_2^{\gamma}} \tag{4.9}$$

such that $\tilde{\boldsymbol{\beta}} = (\tilde{\beta}_1^T, \ldots, \tilde{\beta}_p^T)^T$ is the WLAD 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 [24], and we also use the default setting in the *grpreg* package in R, which chooses the λ with the smallest value for the BIC criterion (4.7).

4.2.3 Theoretical Properties

It can be shown that with the appropriate choice in tuning parameter, the adaptive group WLAD-LASSO possesses the properties of consistency, sparsity, and, therefore, the oracle. We must make the same assumptions as before with the adaptive group LAD-LASSO and remind ourselves of the notation required.

First, 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. Denote

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

In addition to the above, we must make the following assumptions, like before:

- The errors ε_i have continuous and positive density at the origin.
- The matrix $cov(\boldsymbol{x}_1) = \boldsymbol{\Sigma}$ exists and is positive definite.
- The weights w_i are defined such that $0 < w_i \leq 1$.

We must also 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 WLAD-LASSO estimators.

Theorem 4.4. *(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 WLAD-LASSO estimator can be $\sqrt{n}$-consistent. The next theorem establishes the sparsity of the adaptive group WLAD-LASSO estimators.

Theorem 4.5. *(Selection Consistency) If* $\sqrt{n}a_n \to_p 0$ *and* $\sqrt{n} \to_p \infty$*, then* $P(\hat{\boldsymbol{\beta}}_b = 0) \to 1$.

The above theorem states that the adaptive group WLAD-LASSO can consistently estimate insignificant coefficients as zero. This allows for simultaneous parameter estimation and variable selection. With the previous two theorems and the assumptions from earlier, the oracle property can be established for the adaptive group WLAD-LASSO estimators.

Theorem 4.6. *(Oracle Property) If* $\sqrt{n}a_n \to_p 0$*, then* $\sqrt{n}(\hat{\boldsymbol{\beta}}_a - \boldsymbol{\beta}_a) \to_d N(0, \boldsymbol{\Sigma}_a)$.

Based on Theorem 5, with probability tending to one, all of the zero coefficients will be estimated as such, essentially performing variable selection. Based on Theorem 4, 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 6, which states that the adaptive group WLAD-LASSO has the property to identify the correct model consistently. The details for the proof of Theorems 4, 5, and 6 can be found in the Appendix.

Chapter 5

Simulation Studies and Real Data Application

In this chapter, several simulation studies and a real data application are presented, which can be divided into two categories. The first simulation study is for the regular robust group variable selection methods, while the other is for the adaptive robust group variable selection methods. For the regular methods, there is a simulation study showing the effectiveness of the group LAD-LASSO method when the data presents with outliers in the y-direction. Then, there is a study comparing the performance of the group WLAD-LASSO to that of the group LASSO and group LAD-LASSO when there are outliers in both the x- and y-direction. For the adaptive methods, a similar study is presented for the adaptive group LAD-LASSO for outliers only in the y-direction and for the adaptive group WLAD-LASSO for outliers in both directions. The previous simulations are all for predictors split into two groups. Another simulation study is presented comparing the adaptive group methods for predictors separated into seven groups for data with x- and y-outliers. The last section presents a real data application.

5.1 Simulation Study: 2 Groups

5.1.1 Simulation Setup

There will be four simulation studies presented in total in this section. They can be distinguished by the type of group variable selection method (regular or adaptive) and by the type of outliers present in the data (strictly y-outliers or both x- and y-outliers). Thus, in order, we will present studies to showcase the effectiveness of the group LAD-LASSO (regular/y-outliers), the group WLAD-LASSO (regular/x- and y-outliers), the adaptive

group LAD-LASSO (adaptive/y-outliers), and the adaptive group WLAD-LASSO (adaptive/x- and y-outliers).

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 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:

$$ME(\hat{\boldsymbol{\beta}}) = \frac{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T \boldsymbol{X}^T \boldsymbol{X} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})}{n} \tag{5.1}$$

Ideally, this model error will be very close to 0, indicating the method is doing a great job of estimating the actual model.

In addition to the model error, the simulations for the adaptive methods include a column for the mean percentage of correct zeros, which is denoted as mean % of CZ in the tables

5.1.2 Regular GVS Methods: Y-Direction Outliers

First, the group LASSO and group LAD-LASSO will be evaluated for data with outliers in the y-direction only. The results for the various errors can be found in Tables 5.1-5.3. For all cases of contamination greater than 0 ($\varepsilon > 0$), the group LAD-LASSO has the smallest model error. The group LAD-LASSO also has the model error that is close to 0, indicating the group LAD-LASSO is the better group variable selection method with outliers in the y-direction. Figure 5.1 shows the box plots of model error for each method at various contamination levels for t_3 errors where $\sigma = 0.1$ and $n = 100$.

5.1.3 Regular GVS Methods: X-Direction and Y-Direction Outliers

All of the regular group selection methods will be compared using simulated data with both x- and y-direction outliers: the group LASSO, the group LAD-LASSO, and the group WLAD-LASSO. The results can be found in Tables 5.4-5.6 for $\sigma = 1$. Results for $\sigma = 0.5$ are similar and can be found in tables B.1-B.3 in Appendix B. The group WLAD-LASSO consistently has the smallest model error that is also closest to 0. Figure 5.2 gives the box plots for model error comparing the three methods in the case of contamination in both the x- and y-directions. In each case, it can be seen that the group WLAD-LASSO gives the smallest model error, especially in the cases of contamination.

Table 5.1: Simulation results for $N(0,1)$ errors for strictly Y-outliers

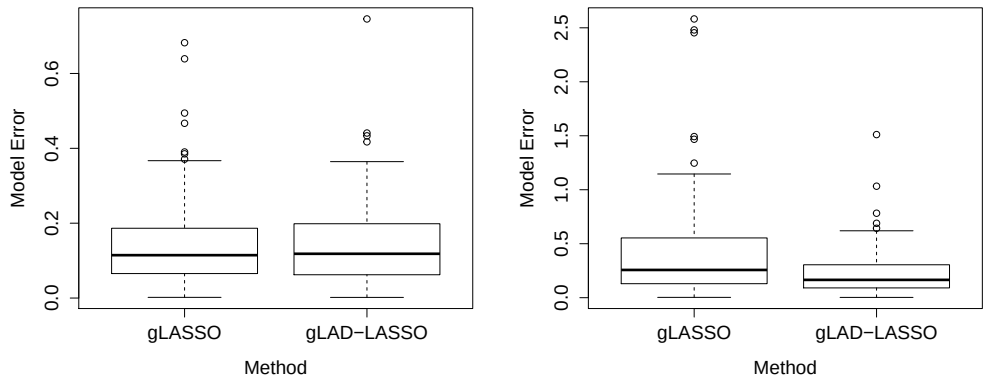
σ	n	ϵ	Method	Mean ME	Median ME
0.5	50	0	g LASSO	0.03	0.02
			g LAD-LASSO	0.03	0.03
		0.1	g LASSO	0.37	0.22
			g LAD-LASSO	0.17	0.13
		0.2	g LASSO	1.27	0.99
			g LAD-LASSO	0.34	0.24
		0.3	g LASSO	2.95	2.36
			g LAD-LASSO	0.70	0.46
	100	0	g LASSO	0.01	0.01
			g LAD-LASSO	0.01	0.01
		0.1	g LASSO	0.26	0.18
			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 5.2: 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

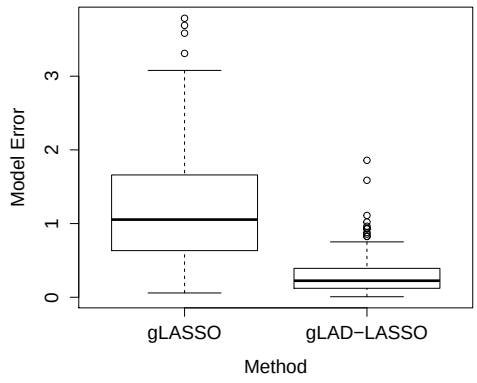
Table 5.3: Simulation results for t_5 errors for strictly Y-outliers

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

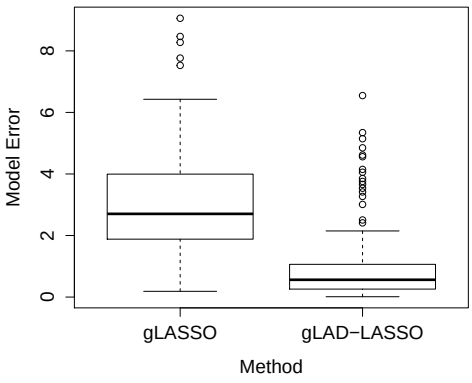


(a) 0%

(b) 10%



(c) 20%



(d) 30%

Figure 5.1: 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 $n = 100$.

Table 5.4: Simulation results for $N(0,1)$ error for X- and Y-outliers

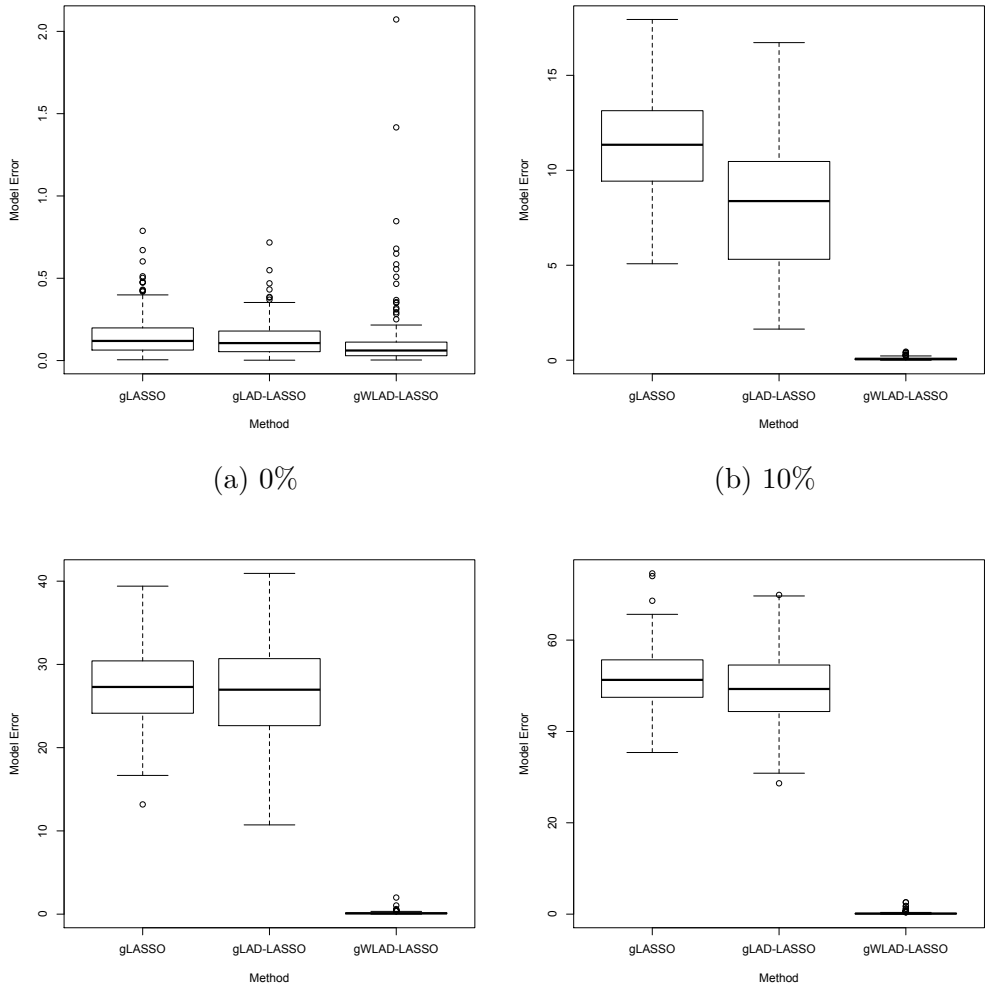
σ	n	ϵ	Method	Mean ME	Median ME
1.0	50	0	g LASSO	0.11	0.10
			g LAD-LASSO	0.13	0.10
			g WLAD-LASSO	0.21	0.17
		0.1	g LASSO	11.09	10.85
			g LAD-LASSO	7.94	7.50
			g WLAD-LASSO	0.30	0.22
		0.2	g LASSO	27.88	27.79
			g LAD-LASSO	27.28	26.61
			g WLAD-LASSO	0.64	0.34
		0.3	g LASSO	50.40	50.84
			g LAD-LASSO	48.92	48.22
			g WLAD-LASSO	0.58	0.12
	100	0	g LASSO	0.05	0.04
			g LAD-LASSO	0.06	0.05
			g WLAD-LASSO	0.10	0.08
		0.1	g LASSO	10.64	10.41
			g LAD-LASSO	7.49	7.07
			g WLAD-LASSO	0.09	0.09
		0.2	g LASSO	28.56	28.63
			g LAD-LASSO	26.45	26.41
			g WLAD-LASSO	0.29	0.21
		0.3	g LASSO	51.58	51.31
			g LAD-LASSO	49.54	49.34
			g WLAD-LASSO	0.22	0.10
	200	0	g LASSO	0.03	0.03
			g LAD-LASSO	0.03	0.03
			g WLAD-LASSO	0.04	0.04
		0.1	g LASSO	10.58	10.45
			g LAD-LASSO	7.20	7.40
			g WLAD-LASSO	0.06	0.05
		0.2	g LASSO	28.66	28.71
			g LAD-LASSO	26.79	27.22
			g WLAD-LASSO	0.23	0.17
		0.3	g LASSO	51.76	51.58
			g LAD-LASSO	51.46	51.40
			g WLAD-LASSO	0.11	0.09

Table 5.5: Simulation results for t_3 error for X- and Y-outliers

σ	n	ϵ	Method	Mean ME	Median ME
1.0	50	0	g LASSO	0.24	0.22
			g LAD-LASSO	0.30	0.23
			g WLAD-LASSO	0.52	0.32
		0.1	g LASSO	10.90	10.28
			g LAD-LASSO	8.63	8.03
			g WLAD-LASSO	0.73	0.39
		0.2	g LASSO	27.26	27.62
			g LAD-LASSO	25.45	25.27
			g WLAD-LASSO	0.10	0.06
		0.3	g LASSO	50.95	50.37
			g LAD-LASSO	48.55	48.45
			g WLAD-LASSO	0.64	0.18
	100	0	g LASSO	0.15	0.12
			g LAD-LASSO	0.13	0.11
			g WLAD-LASSO	0.11	0.06
		0.1	g LASSO	11.27	11.34
			g LAD-LASSO	8.13	8.38
			g WLAD-LASSO	0.08	0.05
		0.2	g LASSO	27.55	27.29
			g LAD-LASSO	26.49	26.96
			g WLAD-LASSO	0.14	0.10
		0.3	g LASSO	51.61	51.29
			g LAD-LASSO	49.62	49.30
			g WLAD-LASSO	0.20	0.11
	200	0	g LASSO	0.09	0.07
			g LAD-LASSO	0.08	0.06
			g WLAD-LASSO	0.04	0.02
		0.1	g LASSO	10.63	10.50
			g LAD-LASSO	7.78	8.09
			g WLAD-LASSO	0.04	0.03
		0.2	g LASSO	28.27	28.58
			g LAD-LASSO	27.29	27.38
			g WLAD-LASSO	0.09	0.07
		0.3	g LASSO	50.82	50.70
			g LAD-LASSO	50.86	50.61
			g WLAD-LASSO	0.13	0.12

Table 5.6: Simulation results for t_5 error for X- and Y-outliers

σ	n	ϵ	Method	Mean ME	Median ME
1.0	50	0	g LASSO	0.19	0.16
			g LAD-LASSO	0.19	0.18
			g WLAD-LASSO	0.10	0.07
		0.1	g LASSO	10.98	10.72
			g LAD-LASSO	8.21	7.06
			g WLAD-LASSO	0.13	0.09
		0.2	g LASSO	28.59	27.59
			g LAD-LASSO	25.58	24.64
			g WLAD-LASSO	0.31	0.15
		0.3	g LASSO	51.09	50.39
			g LAD-LASSO	48.10	47.81
			g WLAD-LASSO	0.49	0.13
	100	0	g LASSO	0.09	0.08
			g LAD-LASSO	0.09	0.08
			g WLAD-LASSO	0.04	0.03
		0.1	g LASSO	10.76	10.34
			g LAD-LASSO	7.78	7.87
			g WLAD-LASSO	0.05	0.04
		0.2	g LASSO	28.51	28.11
			g LAD-LASSO	26.32	26.60
			g WLAD-LASSO	0.11	0.08
		0.3	g LASSO	51.00	51.09
			g LAD-LASSO	50.42	51.29
			g WLAD-LASSO	0.22	0.09
	200	0	g LASSO	0.05	0.04
			g LAD-LASSO	0.05	0.04
			g WLAD-LASSO	0.02	0.02
		0.1	g LASSO	10.92	10.84
			g LAD-LASSO	7.70	8.13
			g WLAD-LASSO	0.03	0.02
		0.2	g LASSO	28.40	28.13
			g LAD-LASSO	26.70	27.39
			g WLAD-LASSO	0.07	0.06
		0.3	g LASSO	51.28	51.02
			g LAD-LASSO	50.94	50.79
			g WLAD-LASSO	0.12	0.09



(a) 0%

(b) 10%

(c) 20%

(d) 30%

Figure 5.2: Boxplots for Model Error for the x- and y-outlier simulation for comparing the group LASSO (gLASSO) and the group LAD-LASSO (gLAD-LASSO) to the group WLAD-LASSO (gWLAD-LASSO) for various contamination levels for $\varepsilon \sim t_3$ over 200 simulations for $\sigma = 1$ and $n = 100$.

5.1.4 Adaptive GVS Methods: Y-Direction Outliers

The adaptive group LASSO and adaptive group LAD-LASSO will be evaluated for data with outliers in the y-direction only. 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 5.7-5.9 give the resulting model errors for the various setups. Figure 5.3 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 supported visually by the box plots, indicating that the adaptive group LAD-LASSO works well for data with contaminations in the response variable.

5.1.5 Adaptive GVS Methods: X-Direction and Y-Direction Outliers

All of the adaptive group selection methods will be compared using simulated data with both x- and y-direction outliers: the adaptive group LASSO, the adaptive group LAD-LASSO, and the adaptive group WLAD-LASSO. The results can be found in Tables 5.10-5.12 for $\sigma = 1$. Results are similar for $\sigma = 0.5$ and can be found in tables B.4-B.6 in Appendix B. Box plots of model error can be found in Figure 5.4. It is clear from the table that the adaptive group WLAD-LASSO results in the smallest model error of the three methods; the adaptive group WLAD-LASSO also gives the model error closest to 0.

Table 5.7: Simulation results for $N(0,1)$ errors for strictly Y-outliers

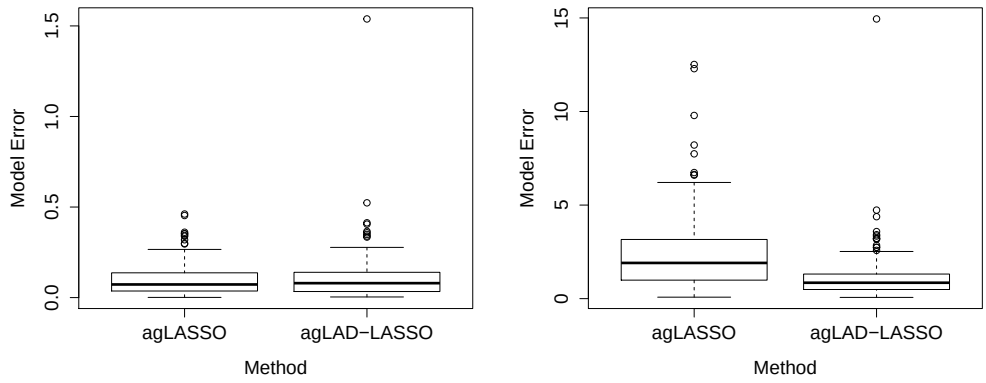
σ	n	ϵ	Method	Mean % of CZ	Mean ME	Median ME
0.5	50	0	ag LASSO	99.7	0.03	0.02
			ag LAD-LASSO	100	0.02	0.01
		0.1	ag LASSO	26.4	0.57	0.39
			ag LAD-LASSO	99.8	0.36	0.24
		0.2	ag LASSO	26.0	1.75	1.37
			ag LAD-LASSO	98.0	0.20	0.08
		0.3	ag LASSO	29.8	4.02	3.51
			ag LAD-LASSO	95.6	0.72	0.38
	100	0	ag LASSO	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 5.8: 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

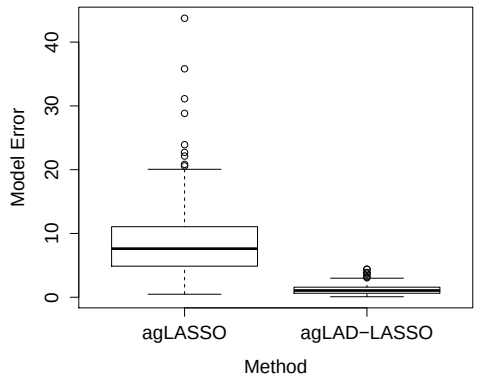
Table 5.9: Simulation results for t_5 errors for strictly Y-outliers

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

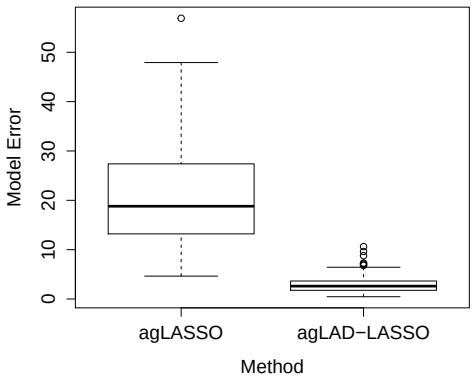


(a) 0%

(b) 10%



(c) 20%



(d) 30%

Figure 5.3: 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 $n = 100$.

Table 5.10: Simulation results for $N(0,1)$ error for X- and Y-outliers

σ	n	ϵ	Method	Mean % of CZ	Mean ME	Median ME
1.0	50	0	ag LASSO	94.3	0.09	0.07
			ag LAD-LASSO	95.8	0.08	0.07
			ag WLAD-LASSO	96.4	0.06	0.05
		0.1	ag LASSO	26.8	11.25	10.85
			ag LAD-LASSO	27.8	10.48	10.28
			ag WLAD-LASSO	92.2	0.08	0.06
		0.2	ag LASSO	23.0	29.74	29.74
			ag LAD-LASSO	21.4	28.91	28.32
			ag WLAD-LASSO	91.4	0.29	0.13
		0.3	ag LASSO	10.1	54.09	53.11
			ag LAD-LASSO	12.9	53.88	52.95
			ag WLAD-LASSO	90.5	0.60	0.18
	100	0	ag LASSO	96.6	0.04	0.03
			ag LAD-LASSO	98.4	0.04	0.03
			ag WLAD-LASSO	98.5	0.03	0.02
		0.1	ag LASSO	28.2	11.24	11.10
			ag LAD-LASSO	28.6	10.18	9.94
			ag WLAD-LASSO	92.9	0.03	0.03
		0.2	ag LASSO	23.4	29.44	28.98
			ag LAD-LASSO	23.6	28.20	28.25
			ag WLAD-LASSO	91.5	0.11	0.08
		0.3	ag LASSO	15.4	52.83	52.91
			ag LAD-LASSO	19.4	52.00	51.61
			ag WLAD-LASSO	90.9	0.23	0.11
	200	0	ag LASSO	98.7	0.02	0.01
			ag LAD-LASSO	99.0	0.02	0.02
			ag WLAD-LASSO	99.4	0.01	0.01
		0.1	ag LASSO	28.9	10.91	10.88
			ag LAD-LASSO	29.5	9.77	9.78
			ag WLAD-LASSO	93.0	0.02	0.02
		0.2	ag LASSO	19.8	28.99	28.75
			ag LAD-LASSO	20.3	28.26	28.22
			ag WLAD-LASSO	91.8	0.07	0.06
		0.3	ag LASSO	18.7	53.04	53.02
			ag LAD-LASSO	18.9	52.40	52.22
			ag WLAD-LASSO	91.2	0.12	0.10

Table 5.11: Simulation results for t_3 error for X- and Y-outliers

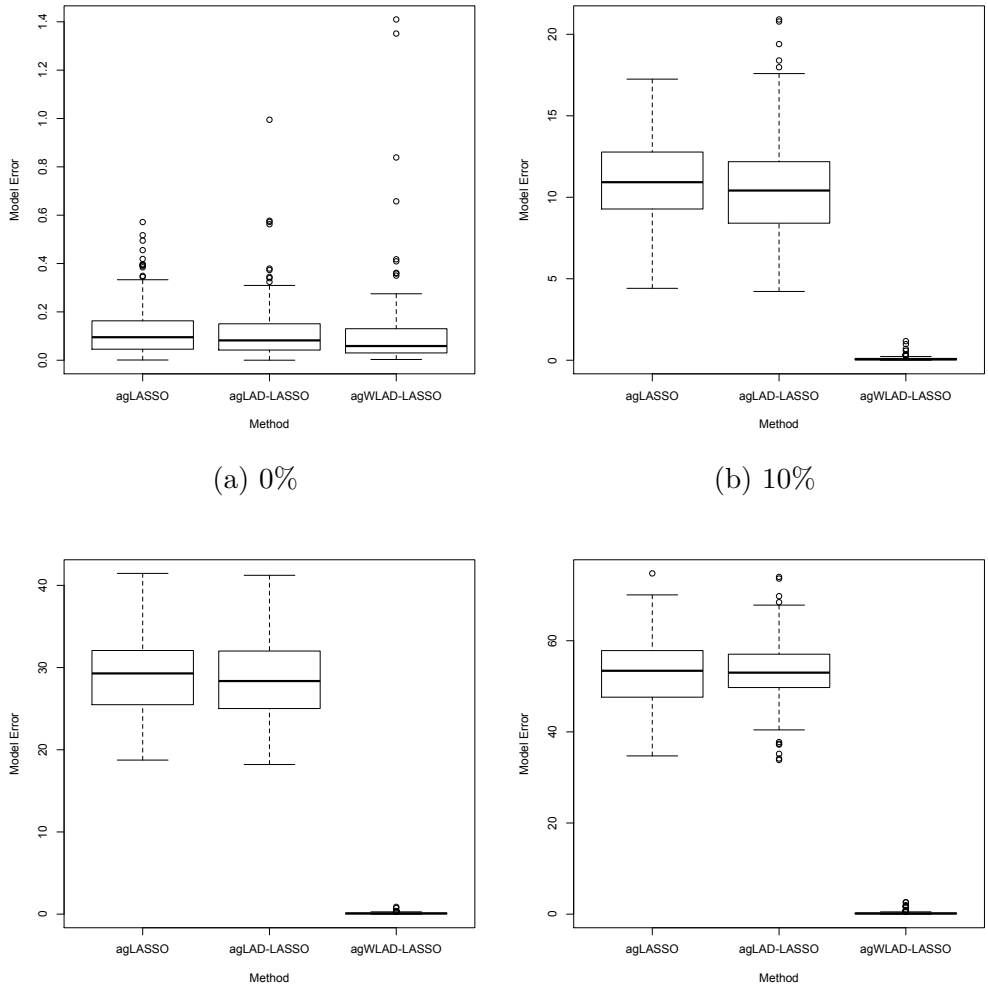
σ	n	ϵ	Method	Mean % of CZ	Mean ME	Median ME
1.0	50	0	ag LASSO	94.9	0.23	0.16
			ag LAD-LASSO	95.8	0.23	0.15
			ag WLAD-LASSO	95.9	0.15	0.09
		0.1	ag LASSO	22.1	11.53	11.21
			ag LAD-LASSO	24.8	10.51	10.06
			ag WLAD-LASSO	93.6	0.12	0.06
		0.2	ag LASSO	13.6	28.80	28.91
			ag LAD-LASSO	13.8	29.08	28.99
			ag WLAD-LASSO	91.9	0.42	0.20
		0.3	ag LASSO	10.9	52.08	52.64
			ag LAD-LASSO	11.9	54.27	53.74
			ag WLAD-LASSO	90.7	0.84	0.27
	100	0	ag LASSO	96.2	0.12	0.10
			ag LAD-LASSO	96.5	0.12	0.08
			ag WLAD-LASSO	97.3	0.10	0.06
		0.1	ag LASSO	25.0	10.99	10.92
			ag LAD-LASSO	25.4	10.49	10.41
			ag WLAD-LASSO	94.3	0.09	0.06
		0.2	ag LASSO	16.0	29.02	29.29
			ag LAD-LASSO	16.3	28.52	28.36
			ag WLAD-LASSO	92.0	0.11	0.08
		0.3	ag LASSO	12.6	53.05	53.41
			ag LAD-LASSO	13.3	53.25	53.00
			ag WLAD-LASSO	90.9	0.26	0.14
	200	0	ag LASSO	97.9	0.06	0.04
			ag LAD-LASSO	98.0	0.04	0.03
			ag WLAD-LASSO	98.5	0.03	0.02
		0.1	ag LASSO	27.4	10.92	10.80
			ag LAD-LASSO	29.8	10.06	9.81
			ag WLAD-LASSO	94.4	0.05	0.03
		0.2	ag LASSO	16.6	28.75	28.58
			ag LAD-LASSO	19.3	28.46	28.62
			ag WLAD-LASSO	92.2	0.11	0.09
		0.3	ag LASSO	13.4	52.43	52.36
			ag LAD-LASSO	13.5	52.06	51.82
			ag WLAD-LASSO	91.0	0.15	0.12

5.2 Simulation Study: 7 Groups

This simulation is designed to show the effectiveness of the adaptive group WLAD-LASSO compared to the other adaptive group variable selection methods, the adaptive group LASSO and the adaptive group LAD-LASSO. Its setup is similar to the setup for the 2-group simulation. The sample sizes n, the contamination rate ϵ, and the number of contaminated points m are defined to be the same as before. Similarly, 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 $\beta_1 = (3, 1.5, 2, 0, 0, 3, 2, 0.5, 4.5, 3.5, 0, 0, 0, 0, 1.5, 1, 0.5, 5, 3, 4.5, 0, 0, 0, 4.5, 1, 3, 2)$, such that there are seven alternating groups of predictor variables of varying sizes. The sizes of the groups are 3, 2, 5, 4, 6, 3, and 4, respectively, for groups 1, 2, 3, 4, 5, 6, and 7. 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, which are the same from before. 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 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 exactly as before. Ideally, this model error will be very close to 0, indicating the method is doing a great job of estimating the actual model.

Table 5.12: Simulation results for t_5 error for X- and Y-outliers

σ	n	ϵ	Method	Mean % of CZ	Mean ME	Median ME
1.0	50	0	ag LASSO	94.5	0.14	0.09
			ag LAD-LASSO	97.5	0.14	0.12
			ag WLAD-LASSO	97.8	0.10	0.07
		0.1	ag LASSO	23.1	11.62	11.23
			ag LAD-LASSO	23.3	10.63	10.08
			ag WLAD-LASSO	93.4	0.15	0.10
		0.2	ag LASSO	14.9	29.58	29.04
			ag LAD-LASSO	16.4	29.38	28.69
			ag WLAD-LASSO	92.8	0.36	0.18
		0.3	ag LASSO	10.2	53.13	52.55
			ag LAD-LASSO	10.6	53.17	52.38
			ag WLAD-LASSO	90.1	0.67	0.21
	100	0	ag LASSO	98.0	0.07	0.05
			ag LAD-LASSO	98.1	0.05	0.04
			ag WLAD-LASSO	98.2	0.04	0.03
		0.1	ag LASSO	23.9	10.92	10.97
			ag LAD-LASSO	25.1	10.19	9.88
			ag WLAD-LASSO	93.7	0.05	0.04
		0.2	ag LASSO	16.7	29.64	29.38
			ag LAD-LASSO	18.1	29.15	28.58
			ag WLAD-LASSO	93.0	0.15	0.11
		0.3	ag LASSO	11.8	52.53	51.60
			ag LAD-LASSO	12.8	52.43	52.40
			ag WLAD-LASSO	91.2	0.28	0.12
	200	0	ag LASSO	99.0	0.03	0.02
			ag LAD-LASSO	99.2	0.03	0.02
			ag WLAD-LASSO	99.8	0.02	0.01
		0.1	ag LASSO	28.9	11.03	10.76
			ag LAD-LASSO	29.4	9.64	9.54
			ag WLAD-LASSO	94.3	0.03	0.02
		0.2	ag LASSO	19.3	28.78	28.77
			ag LAD-LASSO	20.0	28.88	29.14
			ag WLAD-LASSO	93.1	0.08	0.07
		0.3	ag LASSO	14.1	52.44	52.01
			ag LAD-LASSO	14.3	52.81	52.55
			ag WLAD-LASSO	92.6	0.14	0.11



(a) 0%

(b) 10%

(c) 20%

(d) 30%

Figure 5.4: Boxplots for Model Error for the x- and y-outlier simulation for comparing the adaptive group LASSO (agLASSO) and the adaptive group LAD-LASSO (agLAD-LASSO) to the adaptive group WLAD-LASSO (agWLAD-LASSO) for various contamination levels for $\varepsilon \sim t_3$ over 200 simulations.

Additionally, the mean percentages of correct zero coefficients and incorrect zero coefficients, which are averaged over the 200 simulations, are recorded. They are denoted as the mean % of CZ and the mean % of IZ, respectively, in the table. These results are shown in Tables 5.13-5.15 for $\sigma = 1$ for standard normal, t_3, and t_5, respectively. The results for $\sigma = 0.5$ are similar and are shown in tables B.7-B.9 in appendix B. Figure 5.5 presents the box plots for the model error for the 7 group simulation with t_3 errors for a sample size of 100 where $\sigma = 1$. In all cases of contamination, the adaptive group WLAD-LASSO (agWLAD-LASSO) has the smallest model error.

In all cases with no contamination, the oracle property holds true, since the zero coefficients are estimated as such, given that the mean % of correct zeros is very close to 100% for all the adaptive methods in the simulation. Similarly, the mean % of incorrect zero coefficients is also close to 0%. For all sample sizes and all contamination levels, the adaptive group WLAD-LASSO has the smallest model error, which is also the model error closest to zero. Therefore, when there are grouped predictors in regression with outliers in both the response and the predictors, the adaptive group WLAD-LASSO is the best method in terms of model error.

5.3 Real Data Example

In order to show the effectiveness of the four proposed 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 [20]. The response is the expression level of gene TRIM32, which causes Bardet-Biedl syndrome. There are 100 predictors, which

Table 5.13: Simulation results for $\sigma = 1$ for $N(0,1)$ error for X- and Y-outliers for 7 groups

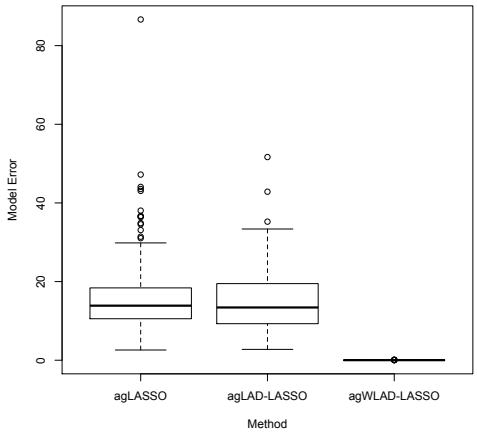
σ	n	ϵ	Method	Mean % of CZ	Mean % of IZ	Mean ME	Median ME
1.0	50	0	ag LASSO	97.3	3.0	0.40	0.38
			ag LAD-LASSO	99.4	1.8	0.37	0.36
			ag WLAD-LASSO	99.1	1.0	0.39	0.36
		0.1	ag LASSO	45.4	56.1	13.19	10.12
			ag LAD-LASSO	48.6	50.7	12.72	10.38
			ag WLAD-LASSO	98.5	4.8	0.01	0.01
		0.2	ag LASSO	36.5	64.0	32.92	29.07
			ag LAD-LASSO	44.2	57.6	29.77	26.42
			ag WLAD-LASSO	96.3	6.2	0.03	0.03
		0.3	ag LASSO	15.7	83.3	53.50	48.21
			ag LAD-LASSO	19.4	79.8	56.14	51.49
			ag WLAD-LASSO	98.8	6.5	0.06	0.05
	100	0	ag LASSO	96.7	2.4	0.18	0.18
			ag LAD-LASSO	98.2	1.9	0.18	0.17
			ag WLAD-LASSO	99.3	1.0	0.04	0.04
		0.1	ag LASSO	37.4	61.3	8.88	7.91
			ag LAD-LASSO	40.8	52.8	9.40	7.98
			ag WLAD-LASSO	95.9	3.1	0.03	0.03
		0.2	ag LASSO	32.6	68.1	23.07	21.63
			ag LAD-LASSO	35.1	65.0	23.65	21.60
			ag WLAD-LASSO	96.8	5.1	0.01	0.01
		0.3	ag LASSO	13.5	78.5	45.98	44.28
			ag LAD-LASSO	17.2	72.7	47.49	46.44
			ag WLAD-LASSO	94.1	6.1	0.03	0.02
	200	0	ag LASSO	96.0	2.9	0.09	0.08
			ag LAD-LASSO	98.7	1.4	0.09	0.09
			ag WLAD-LASSO	98.9	0.7	0.10	0.09
		0.1	ag LASSO	41.5	59.4	6.21	5.91
			ag LAD-LASSO	46.3	55.7	6.18	5.82
			ag WLAD-LASSO	94.4	6.2	0.02	0.02
		0.2	ag LASSO	33.9	64.6	18.29	18.07
			ag LAD-LASSO	37.1	62.8	18.63	17.83
			ag WLAD-LASSO	94.9	6.4	0.00	0.01
		0.3	ag LASSO	11.0	88.3	39.28	38.90
			ag LAD-LASSO	28.4	76.2	38.58	36.81
			ag WLAD-LASSO	93.1	6.7	0.02	0.01

Table 5.14: Simulation results for $\sigma = 1$ for t_3 error for X- and Y-outliers for 7 groups

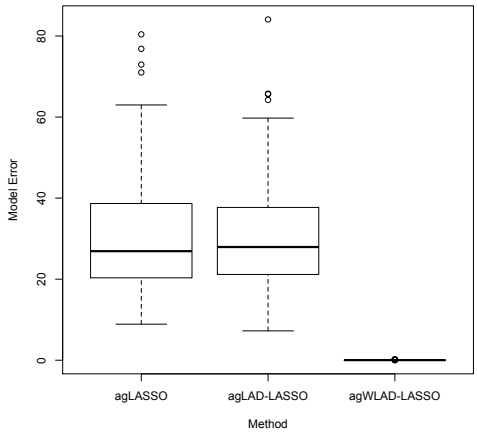
σ	n	ϵ	Method	Mean % of CZ	Mean % of IZ	Mean ME	Median ME
1.0	50	0	ag LASSO	97.8	2.2	0.14	0.16
			ag LAD-LASSO	98.2	1.5	0.12	0.17
			ag WLAD-LASSO	99.2	0.6	0.10	0.09
		0.1	ag LASSO	47.7	51.8	15.96	13.71
			ag LAD-LASSO	48.3	51.1	14.83	12.25
			ag WLAD-LASSO	97.6	2.3	0.06	0.03
		0.2	ag LASSO	32.4	67.7	32.41	28.55
			ag LAD-LASSO	45.5	53.5	32.02	28.73
			ag WLAD-LASSO	96.3	5.6	0.03	0.03
		0.3	ag LASSO	12.4	87.1	59.70	56.43
			ag LAD-LASSO	20.9	84.5	57.85	53.98
			ag WLAD-LASSO	95.0	6.7	0.01	0.01
	100	0	ag LASSO	97.4	2.1	0.06	0.05
			ag LAD-LASSO	98.6	2.0	0.06	0.04
			ag WLAD-LASSO	99.2	0.6	0.01	0.01
		0.1	ag LASSO	45.7	55.0	16.07	13.87
			ag LAD-LASSO	47.0	52.4	15.11	13.41
			ag WLAD-LASSO	96.2	4.2	0.02	0.02
		0.2	ag LASSO	24.4	74.8	30.48	26.91
			ag LAD-LASSO	25.3	73.3	30.32	27.93
			ag WLAD-LASSO	95.2	6.2	0.03	0.02
		0.3	ag LASSO	17.0	83.2	53.81	51.63
			ag LAD-LASSO	25.4	77.2	55.91	51.96
			ag WLAD-LASSO	94.5	6.4	0.04	0.04
	200	0	ag LASSO	95.9	4.6	0.03	0.03
			ag LAD-LASSO	96.9	3.8	0.03	0.02
			ag WLAD-LASSO	99.3	2.3	0.01	0.01
		0.1	ag LASSO	48.7	53.4	6.80	6.36
			ag LAD-LASSO	49.5	51.6	6.76	6.24
			ag WLAD-LASSO	97.4	4.7	0.01	0.01
		0.2	ag LASSO	34.5	78.3	19.66	18.76
			ag LAD-LASSO	37.5	57.4	19.69	19.09
			ag WLAD-LASSO	96.3	5.0	0.01	0.01
		0.3	ag LASSO	14.8	88.4	39.73	37.73
			ag LAD-LASSO	20.2	81.2	39.95	38.52
			ag WLAD-LASSO	94.7	6.8	0.02	0.02

Table 5.15: Simulation results for $\sigma = 1$ for t_5 error for X- and Y-outliers for 7 groups

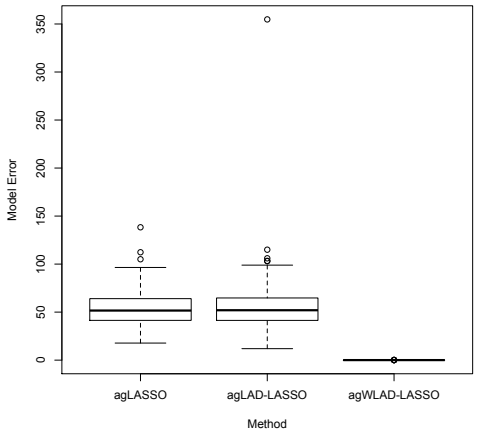
σ	n	ϵ	Method	Mean % of CZ	Mean % of IZ	Mean ME	Median ME
1.0	50	0	ag LASSO	96.2	3.1	0.06	0.06
			ag LAD-LASSO	98.3	2.3	0.05	0.04
			ag WLAD-LASSO	99.6	0.9	0.03	0.02
		0.1	ag LASSO	28.4	72.9	17.38	15.45
			ag LAD-LASSO	37.8	64.0	15.65	14.10
			ag WLAD-LASSO	95.2	3.9	0.02	0.01
		0.2	ag LASSO	25.5	76.5	34.67	31.57
			ag LAD-LASSO	24.6	75.8	35.74	32.92
			ag WLAD-LASSO	97.9	4.7	0.04	0.03
		0.3	ag LASSO	15.2	86.8	60.33	57.93
			ag LAD-LASSO	23.4	76.6	60.10	55.93
			ag WLAD-LASSO	96.1	5.5	0.03	0.02
	100	0	ag LASSO	95.6	4.5	0.03	0.03
			ag LAD-LASSO	97.5	2.9	0.03	0.03
			ag WLAD-LASSO	99.8	1.8	0.06	0.03
		0.1	ag LASSO	43.8	56.7	8.54	7.40
			ag LAD-LASSO	47.6	54.6	8.44	7.64
			ag WLAD-LASSO	96.2	5.1	0.00	0.01
		0.2	ag LASSO	25.9	67.3	23.20	22.21
			ag LAD-LASSO	37.2	62.1	24.45	22.54
			ag WLAD-LASSO	96.5	5.5	0.01	0.01
		0.3	ag LASSO	13.4	88.5	44.66	43.25
			ag LAD-LASSO	14.5	78.1	45.06	43.46
			ag WLAD-LASSO	93.9	6.7	0.02	0.01
	200	0	ag LASSO	98.0	1.5	0.04	0.04
			ag LAD-LASSO	99.5	1.1	0.04	0.03
			ag WLAD-LASSO	99.8	0.6	0.03	0.03
		0.1	ag LASSO	45.3	60.0	6.16	5.50
			ag LAD-LASSO	42.2	55.1	6.38	5.83
			ag WLAD-LASSO	98.7	3.8	0.02	0.02
		0.2	ag LASSO	34.1	65.8	18.70	18.20
			ag LAD-LASSO	38.4	60.5	18.27	17.64
			ag WLAD-LASSO	96.5	5.5	0.01	0.01
		0.3	ag LASSO	17.8	82.7	39.09	38.66
			ag LAD-LASSO	31.5	60.0	40.10	40.10
			ag WLAD-LASSO	92.7	6.9	0.02	0.01



(a) 10%



(b) 20%



(c) 30%

Figure 5.5: Boxplots for Model Error for the x- and y-outlier simulation for comparing the adaptive group LASSO (agLASSO) and the adaptive group LAD-LASSO (agLAD-LASSO) to the adaptive group WLAD-LASSO (agWLAD-LASSO) for various contamination levels for $\varepsilon \sim t_3$ over 200 simulations with 7 groups and a sample size of 100.

are the expression levels of 20 genes, which were expanded using 5 basis B-splines [25]. That is, each 5 consecutive columns corresponds to a grouped gene.

Preliminary analyses of the data indicate there is some 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 [6]. 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. However, this is not enough to truly show how well the proposed methods work in comparison to the non-adaptive and adaptive group LASSO methods. As a result, 24 observations in the response are randomly chosen and shifted to become outliers, and similarly for the x-matrix, such that each column in the predictor will have 24 observations randomly shifted to simulate outliers (24 observations are 20% of the overall 120 observations, indicating there will be 20% contamination of outliers).

In order to compute the robust distances, special considerations are made, due to the high dimensionality of the predictor matrix, which is 120 x 100. The covariance matrix will be found using the R-MCD (regularized MCD estimator) [10]. The R-MCD is similar to the MCD described before for the group WLAD-LASSO; however, exactly half of the observations are used such that $h = \frac{n}{1}$ and, to compensate for the shortage of data, regularization is applied to the MCD estimator resulting in the R-MCD estimator. In particular, the authors suggest using ridge regularization.

All three 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.

$$\text{MSE} = \frac{1}{n-p}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{5.2}$$

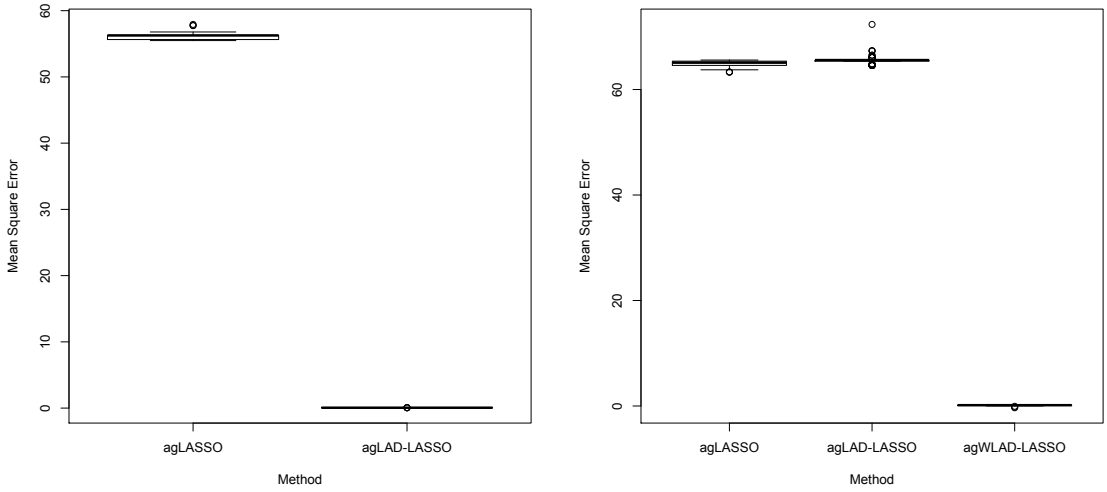
Table 5.16: MSE for the application on the Bardet data set.

Method	Contamination	Mean MSE	Median MSE
gLASSO	0%	0.023	0.021
gLAD-LASSO		0.029	0.028
gWLAD-LASSO		0.222	0.034
agLASSO		0.010	0.010
agLAD-LASSO		0.010	0.010
agWLAD-LASSO		0.010	0.011
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
gLASSO	20% x- and y-outliers	94.057	94.432
gLAD-LASSO		94.173	93.841
gWLAD-LASSO		0.216	0.247
agLASSO		64.926	65.095
agLAD-LASSO		65.572	65.583
agWLAD-LASSO		0.038	0.050

We find the mean square error (MSE) 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. The results are below.

5.3.1 Results

It is clear from the table that, with contamination, the group LAD-LASSO and the group WLAD-LASSO and their adaptive counterparts perform much better than the group LASSO and the adaptive group LASSO. In particular, the group LAD-LASSO and the adaptive group LAD-LASSO do well by having the smallest MSE when there is contamination in the response, while the group WLAD-LASSO and the adaptive group WLAD-LASSO do well and have the smallest MSE when there is contamination in both the predictors and response. It is also of note that the adaptive group variable selection methods have a smaller MSE than the regular group variable selection methods; this can be attributed to the adaptive



(a) 20% y-outliers (b) 20% x- and y-outliers

Figure 5.6: Boxplots for Mean Square Error for the adaptive group LASSO (agLASSO), the adaptive group LAD-LASSO (agLAD-LASSO), and the adaptive group WLAD-LASSO (agWLAD-LASSO) for various conditions over 100 fittings on the Bardet data set.

shrinkage parameter, which gives the adaptive group variable selection methods their nice properties of consistency, sparsity, and the oracle property.

Chapter 6

Conclusion

Variable selection in regression is still a very important problem in statistics. A new twist on the variable selection idea is the notion of grouped predictors. With this added assumption on the structure of the predictors, this creates a new interesting problem to the already intriguing variable selection topic. Still, there are concerns on the tradeoffs of prediction accuracy and interpretability. On the one hand, we'd like to be able to have as accurate predictions as possible by including as many groups of predictor variables as needed. However, on the other, we'd like to be able to interpret the final model in order to understand the relationship underlying the groups of predictors and the response. Herein lies the conundrum of how many groups of variables is too many and how many groups of variables is too few.

Several variable selection methods have been proposed, each well suited to specific data situations. In our case, we are interested in data with outliers. In particular, we are interested in two cases: outliers in the response, and outliers in the response and predictor space. In addition, most of the variable selection methods have been adapted to the group variable selection problem, but most of them have not been shown to be robust to outliers in any direction. In fact, many of them perform poorly in the case of data which exhibits outliers.

In this dissertation, we proposed four methods to perform robust group variable selection. Two of these we call regular robust group variable selection methods, the group LAD-LASSO and the group WLAD-LASSO. The first is well-suited to perform robust group variable selection with outliers in the y-direction, while the second is well-suited to perform robust group variable selection with outliers in both the x- and y-directions. The second type are called adaptive robust group variable selection methods, which are comprised of the

adaptive group LAD-LASSO and the adaptive group WLAD-LASSO. These methods are built to perform robust group variable selection in the presence of outliers in the response and both the response and predictor space, respectively. These second set of adaptive-type methods have nice properties, including the oracle property. A simulation study and a real data application show the effectiveness of all four methods in their respective situations with outliers in the data.

All of the proposed methods are based on the group LASSO, which will select important groups. It would be interesting to apply these same robust measures to other group variables selection methods that not only select important groups, but also select important individual variables within the groups. That is, the final model would have have important groups that have been selected, but not every member of the group would be nonzero. Insignificant groups would still have every variable in the group equal to 0. Exploring the properties of such methods would also be a natural next step in future research. In addition, for all methods, including the proposed methods presented, finding the empirical influence curve as an explicit measure of the robustness is a potential endeavor.

Bibliography

[1] Arslan, O. (2012), "Weighted LAD-LASSO method for robust parameter estimation and variable selection in regression," Computational Statistics & Data Analysis, 56, 1952-1965.

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

[3] Breheny, P. and Huang, J. (to appear) "Group descent algorithms for non convex penalized linear and logistic regression models with grouped predictors," Statistics and Computing,

[4] Breiman, L. (1993) "Better subset selection using the non-negative garotte," Technical Report. University of California, Berkeley.

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

[6] 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.

[7] Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004), "Least Angle Regression," Annals of Statistics, 32, 407-499.

[8] 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.

[9] Frank, I.E. and Friendman, J. H. (1993). "A statistical view of some chemometrics regression tools (with discussion)," Technometrics, 35, 928-961.

[10] Fritsch, V., Varoquaux, G., Thyreau, B., Poline, J.B., and Thirion, B. (2011), "Detecting outlying subjects in high-dimensional neuroimagine datasets with regularized minimum covariance determinant," Medical Image Computing and Computer-Assisted Intervention, 14 (Part 3), 264-71.

[11] Fu, W. J. (1998), "Penalized Regressions: The Bridge Versus the Lasso," Journal of Computational and Graphical Statistics, 7, 3, 397-416.

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

[13] Harrell, F. E. (2001), "Regression modeling strategies: With applications to linear models, logistic regression, and survival analysis," Spring-Verlag, New York.

[14] Hoerl, A.E., and Kennard, R.W. (1970), "Ridge Regression: Biased Estimation for Nonorthogonal Problems," Technometrics, 12, 55-67.

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

[16] Kim, J. Y. (2015), "A Lasso-type Robust Variable Selection for Time-Course Microarray Data," Communications in Statistics-Theory and Methods, 44, 1411-1425.

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

[18] Koenker, R., and Zhao, Q. (1996), "Conditional Quantile Estimation and Inference for ARCH Models," Econometric Theory, 12, 793-813.

[19] Park, C., and Yoon, Y.-J. (2011), "Bridge Regression: Adaptivity and Group Variable Selection," Journal of Statistical Planning and Inference, 141, 3506-3519.

[20] 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

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

[22] Todorov, V., and Filzmoser, P. (2009), "An Object-Oriented Framework for Robust Multivariate Analysis," Journal of Statistical Software, 32 (3), 1-47.

[23] 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.

[24] Wang, H., and Leng, C. (2008), "A note on the adaptive group LASSO," Computational Statistics & Data Analysis, 52, 5277-5286.

[25] Yang, Y. and Zou, H. (2013), "A Fast Unified Algorithm for Computing Group-Lasso Penalized Learning Problems," Statistics and Computing, Accepted.

[26] 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.

[27] Zou, H. (2006), "The adaptive LASSO and its oracle properties," Journal of the American Statistical Association, 101, 1418-1429.

[28] Zou, H., and Hastie, T. (2005), "Regularization and variable selection via the elastic net," Journal of the Royal Statistical Society, Series B, 67, 301-320.

Appendix A

Proofs of Theorems

A.1 Proof of Theorem 1

Before the proof, first assume all of the conditions presented before the theorem in section 5.1.2. 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 $Q(\boldsymbol{\beta})$ (4.1) 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

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

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,

$$
\begin{aligned}
D_n(\boldsymbol{u}) = & \sum_{i=1}^{n} \frac{1}{2}|y_i - \sum_{k=1}^{K} \boldsymbol{x}_{ik}(\boldsymbol{\beta}_k + n^{-1/2}\boldsymbol{u}_k)| + n\sum_{k=1}^{K} \lambda_k ||\boldsymbol{\beta}_k + n^{-1/2}\boldsymbol{u}_k||_2 \\
& - \sum_{i=1}^{n} \frac{1}{2}|y_i - \sum_{k=1}^{K} \boldsymbol{x}_{ik}\boldsymbol{\beta}_k| + n\sum_{i=1}^{K} \lambda_k ||\boldsymbol{\beta}_k||_2
\end{aligned} \tag{A.2}
$$

$$
\begin{aligned}
= & \sum_{i=1}^{n} \frac{1}{2}\{|y_i - \sum_{k=1}^{K} \boldsymbol{x}_{ik}(\boldsymbol{\beta}_k + n^{-1/2}\boldsymbol{u}_k)| - |y_i - \sum_{k=1}^{K} \boldsymbol{x}_{ik}\boldsymbol{\beta}_k|\} \\
& + n\sum_{k=1}^{K} \lambda_k ||\boldsymbol{\beta}_k + n^{-1/2}\boldsymbol{u}_k||_2 - n\sum_{i=1}^{K} \lambda_k ||\boldsymbol{\beta}_k||_2
\end{aligned} \tag{A.3}
$$

$$
\begin{aligned}
= & \sum_{i=1}^{n} \frac{1}{2}\{|y_i - \sum_{k=1}^{K} \boldsymbol{x}_{ik}(\boldsymbol{\beta}_k + n^{-1/2}\boldsymbol{u}_k)| - |y_i - \sum_{k=1}^{K} \boldsymbol{x}_{ik}\boldsymbol{\beta}_k|\} \\
& n\sum_{k=1}^{K} \lambda_k ||\boldsymbol{\beta}_k + n^{-1/2}\boldsymbol{u}_k||_2 - n\sum_{k=1}^{k_0} \lambda_k ||\boldsymbol{\beta}_k||_2
\end{aligned} \tag{A.4}
$$

$$
\begin{aligned}
\geq & \sum_{i=1}^{n} \frac{1}{2}\{|y_i - \sum_{k=1}^{K} \boldsymbol{x}_{ik}(\boldsymbol{\beta}_k + n^{-1/2}\boldsymbol{u}_k)| - |y_i - \sum_{k=1}^{K} \boldsymbol{x}_{ik}\boldsymbol{\beta}_k|\} \\
& n\sum_{k=1}^{k_0} \lambda_k (||\boldsymbol{\beta}_k + n^{-1/2}\boldsymbol{u}_k||_2 - ||\boldsymbol{\beta}||_2)
\end{aligned} \tag{A.5}
$$

$$
\begin{aligned}
\geq & \sum_{i=1}^{n} \frac{1}{2}\{|y_i - \sum_{k=1}^{K} \boldsymbol{x}_{ik}(\boldsymbol{\beta}_k + n^{-1/2}\boldsymbol{u}_k)| - |y_i - \sum_{k=1}^{K} \boldsymbol{x}_{ik}\boldsymbol{\beta}_k|\} \\
& + p_0\sqrt{n}a_n \sum_{k=1}^{k_0} ||\boldsymbol{u}_k||_2
\end{aligned} \tag{A.6}
$$

Line (A.4) follows from (A.3), because $\boldsymbol{\beta}_k = 0$ for any $j > p_0$. Separate equation (A.6) into two parts, divided 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:

$$\sum_{i=1}^{n}\{|y_i - \sum_{k=1}^{K}\boldsymbol{x}_{ik}\boldsymbol{\beta}_k - \sum_{k=1}^{K}\boldsymbol{x}_{ik}n^{-1/2}\boldsymbol{u}_k| - |y_i - \sum_{k=1}^{K}\boldsymbol{x}_{ik}\boldsymbol{\beta}_k| \tag{A.7}$$

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

$$-n^{-1/2}\boldsymbol{u}\sum_{i=1}^{n}\boldsymbol{x}_i[I(\epsilon_i > 0) - I(\epsilon_i < 0)] + 2\sum_{i=1}^{n}\int_0^{n^{-1/2}\boldsymbol{u}^T\boldsymbol{x}_i}[I(\epsilon \leq s) - I(\epsilon \leq 0)]ds \tag{A.8}$$

By the Central Limit Theorem, the first term of (A.8) 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 (A.8), denote the c.d.f. of ϵ_i by F and $\int_0^{n^{-1/2}\boldsymbol{u}^T\boldsymbol{x}_i}[I(\epsilon \leq s) - I(\epsilon \leq 0)]ds$ by $Z_{ni}(\boldsymbol{u})$. Hence,

$$nE[Z_{ni}(\boldsymbol{u})I(n^{-1/2}|\boldsymbol{u}^T\boldsymbol{x}_i| \geq \eta)] \leq nE\{(\int_0^{n^{-1/2}|\boldsymbol{u}^T\boldsymbol{x}_i|} 2ds)^2 I(n^{-1/2}|\boldsymbol{u}^T\boldsymbol{x}_i| \geq \eta)\} \tag{A.9}$$

$$= 4E[|\boldsymbol{u}^T\boldsymbol{x}|^2 I(|\boldsymbol{u}^T\boldsymbol{x}| \geq \sqrt{n}\eta)] \tag{A.10}$$

$$= o(1) \tag{A.11}$$

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,

$$R \leq 2n\eta E\{\int_0^{n^{-1/2}|\boldsymbol{u}^T\boldsymbol{x}_i|} |I(\epsilon_i \leq s) - I(\epsilon_i \leq 0)ds * I(n^{-1/2}|\boldsymbol{u}^T\boldsymbol{x}_i| < \eta)\} \tag{A.12}$$

$$\leq 2n\eta E\{\int_0^{n^{-1/2}|\boldsymbol{u}^T\boldsymbol{x}_i|} [F(s) - F(0)]ds * I(n^{-1/2}|\boldsymbol{u}^T\boldsymbol{x}_i| < \eta)\} \tag{A.13}$$

$$\leq 2n\eta\{f(0) + \kappa\}E\{\int_0^{n^{-1/2}|\boldsymbol{u}^T\boldsymbol{x}_i|} sds * I(n^{-1/2}|\boldsymbol{u}^T\boldsymbol{x}_i| < \eta)\} \tag{A.14}$$

$$\leq \{f(0) + \kappa\}E|\boldsymbol{u}^T\boldsymbol{x}_i|^2 \tag{A.15}$$

The terms in (A.15) converge to 0 as $\eta \to 0$. This implies that R is dominated by the given function. It follows that as $n \to \infty$, $Var(\sum_{i=1}^n Z_{ni}) = \sum_{i=1}^n Var(Z_{ni}) \leq nE(Z_{ni}^2(\boldsymbol{u})) \to 0$. Hence, $\sum_{i=1}^n \{Z_{ni}(\boldsymbol{u}) - E[Z_{ni}(\boldsymbol{u})]\} = o(1)$. Furthermore,

$$E(\sum_{i=1}^n Z_{ni}(\boldsymbol{u})) = nE[Z_{ni}(\boldsymbol{u})] \tag{A.16}$$

$$= nE\{\int_0^{n^{-1/2}\boldsymbol{u}^T\boldsymbol{x}_i} [F(s) - F(0)]ds\} \tag{A.17}$$

$$= E\int_0^{n^{-1/2}} \boldsymbol{u}^T\boldsymbol{x}_i sf(0)ds\} + o(1) \tag{A.18}$$

$$= 0.5f(0)\boldsymbol{u}^T(\boldsymbol{x}_i\boldsymbol{x}_i^T)\boldsymbol{u} + o(1) \tag{A.19}$$

because

$$P\{n^{-1/2}\max(|\boldsymbol{u}^T\boldsymbol{x}_1|, \ldots, |\boldsymbol{u}^T\boldsymbol{x}_n| > \eta^*)\} \leq nP\{|\boldsymbol{u}^T\boldsymbol{x}_1| > \eta^* n^{1/2}\} \tag{A.20}$$

$$\leq \frac{1}{(\eta^*)^2}E\{|\boldsymbol{u}^T\boldsymbol{x}_1|^2 I(|\boldsymbol{u}^T\boldsymbol{x}_1| > \eta^* n^{1/2})\} \to 0 \tag{A.21}$$

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 (A.8) converges to

$f(0)\boldsymbol{u}^T\boldsymbol{\Sigma}\boldsymbol{u}$ in probability. Hence, when C is sufficiently large, the second term of (A.8) dominates both the first part of (A.8) and the last term in (A.6). This implies (A.1) and completes the proof.

A.2 Proof of Theorem 2

First, assume all conditions from the proof of theorem 1 are true. Using an argument from Bloomfield and Steiger [2], 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:

$$= -n^{-1/2} \sum_{i=1}^{n} \text{sgn}(y_i - \boldsymbol{x}_i^T \tilde{\boldsymbol{\beta}}) x_{ik} + \sqrt{n}\lambda_k \frac{\hat{\boldsymbol{\beta}}_{\boldsymbol{b}}}{||\hat{\boldsymbol{\beta}}_{\boldsymbol{b}}||_2} \tag{A.22}$$

where

$$\text{sgn}(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{if } x = 0 \\ -1 & \text{if } x < 0 \end{cases} \tag{A.23}$$

.

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

$$V(\boldsymbol{\Delta}) = n^{-1/2} \sum_{i=1}^{n} \boldsymbol{x}_i \text{sgn}(\epsilon_i - n^{-1/2} \boldsymbol{x}_i^T \boldsymbol{\Delta}). \tag{A.24}$$

By the Central Limit Theorem, it follows that

$$V(\boldsymbol{0}) = n^{-1/2} \sum_{i=1}^{n} \boldsymbol{x}_i \text{sgn}(\epsilon_i) \rightarrow_d N(\boldsymbol{0}, \boldsymbol{\Sigma}), \tag{A.25}$$

where $\rightarrow_d$ means 'convergence in distribution.' Because $n^{-1/2}\text{max}\{|\boldsymbol{u}^T \boldsymbol{x}_i|\} = o(1)$ and because of lemma A.2 from Koenker and Zhao [18], it follows that

$$\sup_{||\boldsymbol{\Delta}||\leq M} |V(\boldsymbol{\Delta}) - V(0) + f(0)\boldsymbol{\Sigma}\boldsymbol{\Delta}| = o(1) \tag{A.26}$$

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 $|\tilde{\boldsymbol{\beta}}_{\boldsymbol{b}} - \boldsymbol{\beta}_b| \leq \epsilon_n = Mn^{-1/2}$,

$$n^{-1/2}\sum_{i=1}^{n} \text{sgn}(y_i - \boldsymbol{x}_i^T\tilde{\boldsymbol{\beta}}) - n^{-1/2}\sum_{i=1}^{n} \boldsymbol{x}_i \text{sgn}(\epsilon) + f(0)\boldsymbol{\Sigma}\boldsymbol{\Delta}^* = o(1) \tag{A.27}$$

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

$$n^{-1/2}\sum_{i=1}^{n} \boldsymbol{x}_i^T \text{sgn}(y_1 - \boldsymbol{x}_i^T\tilde{\boldsymbol{\beta}}) = o(1), \tag{A.28}$$

which, in turn, implies that the first term of (22) is $o(1)$. As for the second term of (22), 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

$$\frac{|\hat{\boldsymbol{\beta}}_{\boldsymbol{b1}}|}{||\hat{\boldsymbol{\beta}}_{\boldsymbol{b}}||_2} \geq \frac{1}{\sqrt{p_k}} > 0. \tag{A.29}$$

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 (22) with probability tending to 1. This means (22) 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.

A.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,

$$S_n(\boldsymbol{v}) = \sum_{i=1}^{n}\{|y_i - \boldsymbol{x}_{ia}\boldsymbol{\beta}_a - n^{-1/2}\boldsymbol{v}^T\boldsymbol{x}_{ia}| - |y_i - \boldsymbol{x}_{ia}^T\boldsymbol{\beta}_a| + n\sum_{j=1}^{p_0}\lambda_j\{|\beta_j + n^{-1/2}\boldsymbol{v}_j| - |\boldsymbol{\beta}_j|\} \quad \text{(A.30)}$$

where $\boldsymbol{x}_{ia} = (x_{i1}, \ldots, x_{ip_0})^T$. Similar to the proof of theorem 1, the first term of (30), such that (30) 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 $\mathbf{0}$ and variance matrix $\boldsymbol{\Sigma}_a$. Also, the absolute value of the second term of (30), which can be denoted by $**$ is constrained by the following

$$|**| \leq \sqrt{n}a_n\sum_{j=1}^{p_0}|\boldsymbol{v}_j| \to 0 \quad \text{(A.31)}$$

Using the results from theorem 2 and remark 1 from Davis [5], the central limit theorem follows, which completes the proof of theorem 3.

A.4 Proof of Theorem 4

The proof of theorem 4, which establishes the estimation consistency of the adaptive group WLAD-LASSO estimators, follows from the proof of theorem 1 and from the proof in Arslan [1] with the additional assumption on the weights w_i such that $0 < w_i \leq 1$. Therefore, it will be omitted.

A.5 Proof of Theorem 5

The proof of theorem 5, which establishes the selection consistency of the adaptive group WLAD-LASSO estimators, follows from the proof of theorem 2 and from the proof in Arslan

[1] with the additional assumption on the weights w_i such that $0 < w_i \leq 1$. Therefore, it will be omitted.

A.6 Proof of Theorem 6

The proof of theorem 5, which establishes the oracle property of the adaptive group WLAD-LASSO estimators, follows from the proof of theorem 3 and from the proof in Arslan [1] with the additional assumption on the weights w_i such that $0 < w_i \leq 1$. Therefore, it will be omitted.

Appendix B

Simulation Results

Table B.1: Simulation results for regular methods when $\sigma = 0.5$ for $N(0,1)$ error for X- and Y-outliers for 2 groups

σ	n	ϵ	Method	Mean ME	Median ME
0.5	50	0	g LASSO	0.03	0.02
			g LAD-LASSO	0.03	0.02
			g WLAD-LASSO	0.05	0.04
		0.1	g LASSO	10.52	10.42
			g LAD-LASSO	7.66	7.01
			g WLAD-LASSO	0.09	0.07
		0.2	g LASSO	27.30	27.11
			g LAD-LASSO	25.99	25.75
			g WLAD-LASSO	0.44	0.16
		0.3	g LASSO	50.71	50.00
			g LAD-LASSO	46.52	46.27
			g WLAD-LASSO	0.59	0.07
	100	0	g LASSO	0.02	0.01
			g LAD-LASSO	0.01	0.01
			g WLAD-LASSO	0.03	0.02
		0.1	g LASSO	10.90	10.76
			g LAD-LASSO	7.41	6.99
			g WLAD-LASSO	0.04	0.03
		0.2	g LASSO	28.02	28.45
			g LAD-LASSO	26.42	26.19
			g WLAD-LASSO	0.17	0.10
		0.3	g LASSO	51.77	51.37
			g LAD-LASSO	49.64	49.87
			g WLAD-LASSO	0.20	0.08
	200	0	g LASSO	0.01	0.01
			g LAD-LASSO	0.01	0.01
			g WLAD-LASSO	0.01	0.01
		0.1	g LASSO	10.74	10.66
			g LAD-LASSO	6.78	6.89
			g WLAD-LASSO	0.02	0.02
		0.2	g LASSO	28.44	28.10
			g LAD-LASSO	27.09	27.32
			g WLAD-LASSO	0.15	0.11
		0.3	g LASSO	51.80	51.76
			g LAD-LASSO	51.38	51.62
			g WLAD-LASSO	0.12	0.08

Table B.2: Simulation results for regular methods when $\sigma = 0.5$ for t_3 error for X- and Y-outliers for 2 groups

σ	n	ϵ	Method	Mean ME	Median ME
0.5	50	0	g LASSO	0.07	0.05
			g LAD-LASSO	0.07	0.05
			g WLAD-LASSO	0.15	0.08
		0.1	g LASSO	10.70	10.51
			g LAD-LASSO	8.06	8.02
			g WLAD-LASSO	0.20	0.13
		0.2	g LASSO	27.88	28.02
			g LAD-LASSO	25.36	25.11
			g WLAD-LASSO	0.60	0.24
		0.3	g LASSO	51.03	49.34
			g LAD-LASSO	48.12	48.03
			g WLAD-LASSO	0.62	0.10
	100	0	g LASSO	0.04	0.03
			g LAD-LASSO	0.04	0.03
			g WLAD-LASSO	0.02	0.01
		0.1	g LASSO	10.77	10.56
			g LAD-LASSO	7.20	6.95
			g WLAD-LASSO	0.02	0.02
		0.2	g LASSO	27.50	27.08
			g LAD-LASSO	25.93	25.77
			g WLAD-LASSO	0.07	0.04
		0.3	g LASSO	51.45	51.40
			g LAD-LASSO	50.26	50.12
			g WLAD-LASSO	0.24	0.09
	200	0	g LASSO	0.02	0.02
			g LAD-LASSO	0.02	0.01
			g WLAD-LASSO	0.01	0.01
		0.1	g LASSO	10.68	10.71
			g LAD-LASSO	7.06	7.22
			g WLAD-LASSO	0.01	0.01
		0.2	g LASSO	28.00	27.81
			g LAD-LASSO	26.86	26.91
			g WLAD-LASSO	0.05	0.04
		0.3	g LASSO	51.55	51.04
			g LAD-LASSO	50.79	50.70
			g WLAD-LASSO	0.11	0.08

Table B.3: Simulation results for regular methods when $\sigma = 0.5$ for t_5 error for X- and Y-outliers for 2 groups

σ	n	ϵ	Method	Mean ME	Median ME
0.5	50	0	g LASSO	0.04	0.03
			g LAD-LASSO	0.05	0.04
			g WLAD-LASSO	0.03	0.02
		0.1	g LASSO	10.61	10.30
			g LAD-LASSO	8.26	7.78
			g WLAD-LASSO	0.04	0.03
		0.2	g LASSO	26.89	26.41
			g LAD-LASSO	25.02	24.06
			g WLAD-LASSO	0.16	0.06
		0.3	g LASSO	51.01	50.40
			g LAD-LASSO	48.74	48.55
			g WLAD-LASSO	0.55	0.11
	100	0	g LASSO	0.02	0.02
			g LAD-LASSO	0.02	0.02
			g WLAD-LASSO	0.01	0.01
		0.1	g LASSO	11.04	10.84
			g LAD-LASSO	7.37	7.36
			g WLAD-LASSO	0.02	0.01
		0.2	g LASSO	27.95	27.80
			g LAD-LASSO	25.62	25.82
			g WLAD-LASSO	0.06	0.05
		0.3	g LASSO	51.61	50.89
			g LAD-LASSO	49.02	49.48
			g WLAD-LASSO	0.19	0.08
	200	0	g LASSO	0.01	0.01
			g LAD-LASSO	0.01	0.01
			g WLAD-LASSO	0.01	0.01
		0.1	g LASSO	10.73	10.71
			g LAD-LASSO	7.11	7.22
			g WLAD-LASSO	0.01	0.01
		0.2	g LASSO	28.51	28.22
			g LAD-LASSO	26.58	26.39
			g WLAD-LASSO	0.05	0.04
		0.3	g LASSO	51.94	51.69
			g LAD-LASSO	51.55	51.28
			g WLAD-LASSO	0.11	0.08

Table B.4: Simulation results for adaptive methods when $\sigma = 0.5$ for $N(0,1)$ error for X- and Y-outliers for 2 groups

σ	n	ϵ	Method	Mean % of CZ	Mean ME	Median ME
0.5	50	0	ag LASSO	95.0	0.03	0.03
			ag LAD-LASSO	95.1	0.02	0.02
			ag WLAD-LASSO	95.2	0.01	0.01
		0.1	ag LASSO	26.6	11.81	11.60
			ag LAD-LASSO	26.7	10.42	9.79
			ag WLAD-LASSO	93.9	0.03	0.02
		0.2	ag LASSO	18.7	28.45	27.88
			ag LAD-LASSO	20.3	28.87	27.78
			ag WLAD-LASSO	92.7	0.28	0.08
		0.3	ag LASSO	11.7	53.52	52.58
			ag LAD-LASSO	13.2	52.79	51.94
			ag WLAD-LASSO	90.3	0.77	0.12
	100	0	ag LASSO	95.3	0.01	0.01
			ag LAD-LASSO	95.4	0.01	0.01
			ag WLAD-LASSO	95.6	0.01	0.00
		0.1	ag LASSO	27.4	11.04	11.11
			ag LAD-LASSO	28.4	9.87	9.43
			ag WLAD-LASSO	94.8	0.01	0.01
		0.2	ag LASSO	20.9	28.88	28.77
			ag LAD-LASSO	23.3	28.52	28.22
			ag WLAD-LASSO	93.3	0.07	0.05
		0.3	ag LASSO	14.1	53.52	53.16
			ag LAD-LASSO	14.8	52.58	52.36
			ag WLAD-LASSO	91.5	0.20	0.09
	200	0	ag LASSO	97.1	0.01	0.01
			ag LAD-LASSO	97.8	0.00	0.00
			ag WLAD-LASSO	99.3	0.00	0.00
		0.1	ag LASSO	28.5	10.80	10.76
			ag LAD-LASSO	29.1	9.46	9.29
			ag WLAD-LASSO	94.9	0.01	0.01
		0.2	ag LASSO	23.8	28.75	28.63
			ag LAD-LASSO	24.3	28.44	28.67
			ag WLAD-LASSO	93.8	0.06	0.04
		0.3	ag LASSO	15.9	52.31	51.92
			ag LAD-LASSO	17.2	52.01	51.90
			ag WLAD-LASSO	92.5	0.12	0.09

Table B.5: Simulation results for adaptive methods when $\sigma = 0.5$ for t_3 error for X- and Y-outliers for 2 groups

σ	n	ϵ	Method	Mean % of CZ	Mean ME	Median ME
0.5	50	0	ag LASSO	95.0	0.08	0.06
			ag LAD-LASSO	96.0	0.05	0.03
			ag WLAD-LASSO	96.7	0.04	0.02
		0.1	ag LASSO	21.9	11.91	11.54
			ag LAD-LASSO	23.1	10.90	10.62
			ag WLAD-LASSO	92.8	0.06	0.03
		0.2	ag LASSO	13.6	29.33	29.42
			ag LAD-LASSO	15.1	29.01	28.43
			ag WLAD-LASSO	92.0	0.20	0.08
		0.3	ag LASSO	10.4	53.54	53.19
			ag LAD-LASSO	12.3	51.92	52.11
			ag WLAD-LASSO	90.5	0.84	0.23
	100	0	ag LASSO	96.9	0.04	0.03
			ag LAD-LASSO	97.0	0.02	0.02
			ag WLAD-LASSO	98.2	0.02	0.01
		0.1	ag LASSO	23.5	11.15	10.91
			ag LAD-LASSO	25.3	10.09	9.77
			ag WLAD-LASSO	93.4	0.02	0.02
		0.2	ag LASSO	18.4	29.49	29.17
			ag LAD-LASSO	19.3	29.45	29.33
			ag WLAD-LASSO	92.5	0.09	0.06
		0.3	ag LASSO	12.8	53.30	54.04
			ag LAD-LASSO	12.9	52.58	52.75
			ag WLAD-LASSO	90.6	0.25	0.11
	200	0	ag LASSO	98.4	0.02	0.02
			ag LAD-LASSO	98.9	0.01	0.01
			ag WLAD-LASSO	99.7	0.01	0.00
		0.1	ag LASSO	26.4	10.95	10.95
			ag LAD-LASSO	28.4	9.65	9.59
			ag WLAD-LASSO	94.2	0.01	0.01
		0.2	ag LASSO	21.7	28.70	28.32
			ag LAD-LASSO	21.8	28.53	28.40
			ag WLAD-LASSO	92.7	0.07	0.06
		0.3	ag LASSO	13.1	52.81	52.79
			ag LAD-LASSO	13.3	52.60	52.51
			ag WLAD-LASSO	90.7	0.13	0.09

Table B.6: Simulation results for adaptive methods when $\sigma = 0.5$ for t_5 error for X- and Y-outliers for 2 groups

σ	n	ϵ	Method	Mean % of CZ	Mean ME	Median ME
0.5	50	0	ag LASSO	96.0	0.04	0.04
			ag LAD-LASSO	96.5	0.03	0.02
			ag WLAD-LASSO	96.6	0.02	0.01
		0.1	ag LASSO	22.6	11.41	11.12
			ag LAD-LASSO	22.8	10.54	10.05
			ag WLAD-LASSO	94.5	0.04	0.03
		0.2	ag LASSO	16.7	29.66	29.82
			ag LAD-LASSO	19.3	29.16	28.82
			ag WLAD-LASSO	93.5	0.21	0.07
		0.3	ag LASSO	10.2	54.12	53.47
			ag LAD-LASSO	10.6	52.63	52.56
			ag WLAD-LASSO	90.0	0.73	0.16
	100	0	ag LASSO	96.9	0.02	0.02
			ag LAD-LASSO	97.7	0.01	0.01
			ag WLAD-LASSO	99.1	0.01	0.01
		0.1	ag LASSO	27.0	11.01	10.95
			ag LAD-LASSO	27.2	10.08	9.87
			ag WLAD-LASSO	95.6	0.02	0.01
		0.2	ag LASSO	19.9	29.07	28.97
			ag LAD-LASSO	21.1	29.15	28.71
			ag WLAD-LASSO	93.6	0.09	0.06
		0.3	ag LASSO	10.8	53.17	52.40
			ag LAD-LASSO	15.3	52.44	52.93
			ag WLAD-LASSO	90.7	0.24	0.08
	200	0	ag LASSO	99.2	0.01	0.01
			ag LAD-LASSO	99.4	0.01	0.01
			ag WLAD-LASSO	99.7	0.00	0.00
		0.1	ag LASSO	29.4	10.74	10.74
			ag LAD-LASSO	29.6	9.62	9.43
			ag WLAD-LASSO	95.9	0.01	0.01
		0.2	ag LASSO	21.4	28.58	28.46
			ag LAD-LASSO	21.9	28.54	28.82
			ag WLAD-LASSO	93.7	0.06	0.05
		0.3	ag LASSO	15.7	52.40	52.46
			ag LAD-LASSO	16.3	52.66	52.93
			ag WLAD-LASSO	92.0	0.14	0.09

Table B.7: Simulation results for $\sigma = 0.5$ for $N(0,1)$ error for X- and Y-outliers for 7 groups

σ	n	ϵ	Method	Mean % of CZ	Mean % of IZ	Mean ME	Median ME
0.5	50	0	ag LASSO	97.8	4.1	0.09	0.09
			ag LAD-LASSO	97.9	3.8	0.09	0.09
			ag WLAD-LASSO	98.4	3.6	0.10	0.11
		0.1	ag LASSO	45.4	56.2	13.71	11.64
			ag LAD-LASSO	47.7	54.6	12.35	10.76
			ag WLAD-LASSO	97.0	4.1	0.01	0.01
		0.2	ag LASSO	24.6	67.8	31.25	28.21
			ag LAD-LASSO	24.7	67.3	31.06	26.26
			ag WLAD-LASSO	95.8	5.0	0.03	0.03
		0.3	ag LASSO	18.5	81.3	59.90	56.46
			ag LAD-LASSO	19.4	80.2	57.51	53. 91
			ag WLAD-LASSO	95.0	6.4	0.06	0.06
	100	0	ag LASSO	98.5	2.3	0.05	0.05
			ag LAD-LASSO	99.2	0.7	0.05	0.05
			ag WLAD-LASSO	99.7	0.6	0.10	0.06
		0.1	ag LASSO	37.0	65.2	9.05	7.82
			ag LAD-LASSO	47.3	53.0	9.01	8.15
			ag WLAD-LASSO	98.3	2.7	0.01	0.01
		0.2	ag LASSO	17.3	84.8	22.34	21.37
			ag LAD-LASSO	29.5	82.7	22.81	21.74
			ag WLAD-LASSO	95.0	5.6	0.01	0.01
		0.3	ag LASSO	12.3	88.1	46.56	44.35
			ag LAD-LASSO	16.1	85.4	49.39	46.84
			ag WLAD-LASSO	95.3	5.6	0.03	0.02
	200	0	ag LASSO	96.4	4.6	0.02	0.02
			ag LAD-LASSO	98.0	1.9	0.02	0.02
			ag WLAD-LASSO	100	0.1	0.04	0.05
		0.1	ag LASSO	25.8	76.2	6.27	5.71
			ag LAD-LASSO	48.1	52.0	6.10	5.61
			ag WLAD-LASSO	96.4	5.2	0.01	0.01
		0.2	ag LASSO	21.8	79.8	18.66	18.30
			ag LAD-LASSO	23.0	78.3	18.28	17.49
			ag WLAD-LASSO	95.9	6.0	0.00	0.00
		0.3	ag LASSO	14.5	87.5	38.74	38.07
			ag LAD-LASSO	28.2	81.7	38.38	36.94
			ag WLAD-LASSO	93.7	6.1	0.01	0.01

Table B.8: Simulation results for $\sigma = 0.5$ for t_3 error for X- and Y-outliers for 7 groups

σ	n	ϵ	Method	Mean % of CZ	Mean % of IZ	Mean ME	Median ME
0.5	50	0	ag LASSO	99.3	2.6	0.31	0.24
			ag LAD-LASSO	99.5	1.2	0.29	0.22
			ag WLAD-LASSO	100	0.0	0.36	0.33
		0.1	ag LASSO	47.4	52.7	14.65	12.56
			ag LAD-LASSO	48.7	52.0	15.97	13.78
			ag WLAD-LASSO	98.4	3.1	0.01	0.01
		0.2	ag LASSO	45.7	56.0	33.88	31.14
			ag LAD-LASSO	49.3	52.8	33.19	30.70
			ag WLAD-LASSO	96.1	5.7	0.04	0.03
		0.3	ag LASSO	17.9	85.1	63.51	59.32
			ag LAD-LASSO	20.9	79.0	59.58	55.98
			ag WLAD-LASSO	95.7	6.6	0.01	0.01
	100	0	ag LASSO	96.2	3.7	0.15	0.12
			ag LAD-LASSO	99.6	0.9	0.14	0.11
			ag WLAD-LASSO	99.8	0.3	0.13	0.12
		0.1	ag LASSO	41.7	60.7	10.38	9.39
			ag LAD-LASSO	46.6	55.2	8.98	7.92
			ag WLAD-LASSO	96.3	4.1	0.01	0.01
		0.2	ag LASSO	22.4	69.3	24.58	22.78
			ag LAD-LASSO	36.5	65.5	24.03	22.76
			ag WLAD-LASSO	95.7	6.3	0.01	0.01
		0.3	ag LASSO	18.9	83.7	44.18	42.57
			ag LAD-LASSO	27.2	75.1	47.04	47.13
			ag WLAD-LASSO	94.5	6.9	0.03	0.03
	200	0	ag LASSO	96.4	2.9	0.07	0.06
			ag LAD-LASSO	97.5	2.7	0.07	0.06
			ag WLAD-LASSO	98.3	2.1	0.00	0.00
		0.1	ag LASSO	40.7	58.7	7.13	6.76
			ag LAD-LASSO	50.8	51.4	6.80	6.11
			ag WLAD-LASSO	96.6	5.2	0.01	0.01
		0.2	ag LASSO	40.6	62.8	19.13	18.61
			ag LAD-LASSO	39.3	59.9	19.81	18.81
			ag WLAD-LASSO	94.4	5.7	0.04	0.01
		0.3	ag LASSO	26.5	75.9	40.77	39.24
			ag LAD-LASSO	40.8	63.6	40.54	39.96
			ag WLAD-LASSO	94.1	6.5	0.02	0.02

Table B.9: Simulation results for $\sigma = 0.5$ for t_5 error for X- and Y-outliers for 7 groups

σ	n	ϵ	Method	Mean % of CZ	Mean % of IZ	Mean ME	Median ME
0.5	50	0	ag LASSO	97.1	3.0	0.03	0.02
			ag LAD-LASSO	98.7	2.8	0.03	0.02
			ag WLAD-LASSO	99.8	0.2	0.01	0.01
		0.1	ag LASSO	45.8	56.2	16.45	14.52
			ag LAD-LASSO	49.3	50.2	15.71	13.18
			ag WLAD-LASSO	97.1	4.7	0.02	0.01
		0.2	ag LASSO	22.1	79.9	35.70	33.38
			ag LAD-LASSO	26.7	76.2	36.71	33.32
			ag WLAD-LASSO	96.6	5.0	0.04	0.03
		0.3	ag LASSO	17.6	86.5	62.71	59.29
			ag LAD-LASSO	20.9	80.3	59.08	54.19
			ag WLAD-LASSO	94.1	6.8	0.04	0.01
	100	0	ag LASSO	97.7	3.9	0.08	0.07
			ag LAD-LASSO	98.8	2.2	0.08	0.07
			ag WLAD-LASSO	99.6	2.1	0.07	0.06
		0.1	ag LASSO	18.3	83.2	8.70	7.24
			ag LAD-LASSO	22.1	80.4	8.32	6.82
			ag WLAD-LASSO	97.5	4.1	0.01	0.00
		0.2	ag LASSO	17.7	84.5	22.65	21.95
			ag LAD-LASSO	18.3	83.6	22.76	20.83
			ag WLAD-LASSO	95.7	5.0	0.01	0.01
		0.3	ag LASSO	15.0	87.6	43.70	42.23
			ag LAD-LASSO	17.0	84.9	46.62	44.61
			ag WLAD-LASSO	94.2	6.7	0.01	0.00
	200	0	ag LASSO	96.5	3.9	0.04	0.04
			ag LAD-LASSO	98.8	2.3	0.04	0.04
			ag WLAD-LASSO	98.9	2.2	0.05	0.03
		0.1	ag LASSO	46.0	55.8	6.74	6.21
			ag LAD-LASSO	49.2	52.5	6.32	5.99
			ag WLAD-LASSO	96.9	4.4	0.02	0.02
		0.2	ag LASSO	30.4	71.0	17.93	16.74
			ag LAD-LASSO	39.0	61.1	18.40	17.31
			ag WLAD-LASSO	95.7	4.5	0.06	0.05
		0.3	ag LASSO	11.7	89.8	40.46	40.21
			ag LAD-LASSO	19.1	73.1	38.93	37.58
			ag WLAD-LASSO	94.6	5.8	0.02	0.01