

## ABSTRACT

SCHUMANN, DAVID H. Robust Variable Selection. (Under the direction of Dennis Boos and Leonard Stefanski.)

The prevalence of extreme outliers in regression data sets has led to the development of robust methods that can handle these observations. While much attention has been placed on the problem of estimating regression coefficients in the presence of outliers, few methods address variable selection. We develop and study robust versions of the forward selection algorithm, one of the most popular standard variable selection techniques. Specifically we modify the VAMS procedure, a version of forward selection tuned to control the false selection rate, to simultaneously select variables and eliminate outliers. In an alternative approach, robust versions of the forward selection algorithm are developed using the robust forward addition sequence associated with the generalized score statistic. Combining the robust forward addition sequence with robust versions of BIC and the VAMS procedure, a final model is obtained. Monte Carlo simulation compares these robust methods to current robust methods like the LSA adaptive LASSO and LAD-LASSO. Further simulation investigates the relationship between the breakdown point of the estimation methods central to each procedure and the breakdown point of the final variable selection method.

Robust Variable Selection

by  
David Schumann

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

Statistics

Raleigh, North Carolina

2009

APPROVED BY:

---

Dr. Judy Wang

---

Dr. Lexin Li

---

Dr. Dennis Boos  
Chair of Advisory Committee

---

Dr. Leonard Stefanski  
Co-chair of Advisory Committee

## DEDICATION

To my family and friends,

## BIOGRAPHY

David Schumann was born on September 26, 1980 in Everett, WA. After a short stint at Ohlone junior college in pursuit of a baseball career, he attended Cal Poly, San Luis Obispo. After graduating in 2003, he entered the North Carolina State University graduate statistics program. In May 2005, David obtained his Master of Statistics and decided to continue working towards a doctoral degree. He has conducted his research under the direction of Dr. Dennis Boos and Dr. Leonard Stefanski.

## ACKNOWLEDGMENTS

I think it is most appropriate to start by thanking my advisors Dr. Dennis Boos and Dr. Leonard Stefanski. I have a great amount of respect not only for the contributions they have made to the field of statistics, but also their ability as teachers. I am confident that many other students have benefited from their willingness to spend countless hours outside of the classroom helping with statistical problems. I would also like to thank Dr. Cavell Brownie for her help during my time as a statistical consultant, and Dr. Bibhuti Bhattacharyya for everything he has taught me throughout the years. A special thanks also goes out to Dr. Allan Rossman and Dr. Beth Chance, the professors who first introduced me to statistics when I was an undergraduate at Cal Poly, San Luis Obispo.

Aside of my professors, many other people have aided in my progress as a statistician. I would like to thank fellow classmates Dr. McKay Curtis and Dr. Hugh Crews, who were particularly helpful when I experienced computing problems in Latex or R. I'd also like to thank Aubrey Komorowski for all of her support and hours of help proof-reading my paper. Lastly, I would like to thank my parents Anita and Gerry, and my brother Karl. Their love and support has helped me grow as a statistician, and more importantly, as a person.

## TABLE OF CONTENTS

<b>LIST OF TABLES .....</b>	<b>vii</b>
<b>LIST OF FIGURES .....</b>	<b>viii</b>
<b>1 Robust Variable Selection .....</b>	<b>1</b>
1.1 Introduction . . . . .	1
1.2 Model Selection Approaches . . . . .	3
1.2.1 Subset Selection . . . . .	3
1.2.2 Penalty Methods . . . . .	6
1.3 Robust Regression . . . . .	6
1.3.1 M-estimation in Regression . . . . .	6
1.3.2 High Breakdown Methods . . . . .	8
<b>2 Variable Addition Model Selection.....</b>	<b>12</b>
2.1 Simulation-Based VAMS . . . . .	12
2.2 Fast VAMS . . . . .	15
2.3 Simultaneous Selection and Outlier Detection . . . . .	16
2.4 Outlier Detection Problems for VAMSI . . . . .	21
2.4.1 Variations of VAMSI . . . . .	23
2.4.2 Two-Stage Methods . . . . .	26
2.4.3 Summary . . . . .	27
2.5 Comparison of Methods . . . . .	28
2.5.1 Data Generation . . . . .	28
2.5.2 Simulation Design . . . . .	30
2.5.3 Results . . . . .	32
2.6 Breakdown Properties . . . . .	36
2.6.1 Criteria . . . . .	36
2.6.2 Simulation Design . . . . .	38
2.6.3 Results . . . . .	39
2.7 Conclusions . . . . .	39
<b>3 Variable Selection using Robust Regression.....</b>	<b>43</b>
3.1 Huber-Based Approach . . . . .	43
3.1.1 Score Forward Addition Sequence from M-estimation . . . . .	45
3.1.2 Robust Bayesian Information Criterion . . . . .	47

3.2	M-estimation using $t_3$ . . . . .	49
3.2.1	Comparison of Huber and $t_3$ . . . . .	50
3.2.2	Modified RBIC . . . . .	52
3.2.3	$t_3$ Score Forward Addition Sequence . . . . .	54
3.3	VAMS using Score FAS . . . . .	55
3.4	Comparison of Methods . . . . .	57
3.4.1	Simulation . . . . .	57
3.4.2	Results . . . . .	57
3.4.3	Breakdown . . . . .	59
3.4.4	Conclusion . . . . .	63
<b>4</b>	<b>Lasso-Based Methods . . . . .</b>	<b>64</b>
4.1	LASSO . . . . .	64
4.2	Least Squares Approximation to the Adaptive LASSO . . . . .	66
4.3	LAD-LASSO . . . . .	67
4.4	Comparison of Methods . . . . .	69
4.4.1	Simulation . . . . .	69
4.4.2	Results . . . . .	69
4.4.3	Breakdown . . . . .	70
4.4.4	Conclusion . . . . .	71
<b>5</b>	<b>Conclusion . . . . .</b>	<b>75</b>
5.1	False Selection Rate versus Prediction . . . . .	75
5.2	Breakdown . . . . .	77
5.3	Summary . . . . .	78
	<b>Bibliography . . . . .</b>	<b>82</b>

## LIST OF TABLES

Table 2.1 Forward Addition Sequence for $\mathbf{X}^*$ .....	19
Table 2.2 Simulations Results for $n = 100$ , $k_T = 20$ , and $k_I = 4$ .....	40
Table 2.3 Simulations Results for $n = 100$ , $k_T = 50$ , and $k_I = 10$ .....	41
Table 2.4 Simulations Results for $n = 500$ , $k_T = 50$ , and $k_I = 10$ .....	42
Table 3.1 Simulations Results for $n=100$ , $k_T=20$ , and $k_I=4$ .....	60
Table 3.2 Simulations Results for $n=100$ , $k_T=50$ , and $k_I=10$ .....	61
Table 3.3 Simulations Results for $n=500$ , $k_T=50$ , and $k_I=10$ .....	62
Table 4.1 Simulations Results for $n=100$ , $k_T=20$ , and $k_I=4$ .....	72
Table 4.2 Simulations Results for $n=100$ , $k_T=50$ , and $k_I=10$ .....	73
Table 4.3 Simulations Results for $n=500$ , $k_T=50$ , and $k_I=10$ .....	74
Table 5.1 Simulations Results for $n=100$ , $k_T=20$ , and $k_I=4$ .....	79
Table 5.2 Simulations Results for $n=100$ , $k_T=50$ , and $k_I=10$ .....	80
Table 5.3 Simulations Results for $n=500$ , $k_T=50$ , and $k_I=10$ .....	81



## LIST OF FIGURES

Figure 2.1 Plot of response versus predictor .....	21
Figure 3.1 $\psi$ functions. Left panel: Huber $\psi$ function, $k=1$ ; Right Panel: $t_3$ $\psi$ function, $v=3$ . .....	51
Figure 3.2 Huber and $t_3$ M-estimation Weight functions. Huber: dotted line; $t_3$ : solid line. ....	52

# Chapter 1

## Robust Variable Selection

### 1.1 Introduction

Consider the linear model

$$\mathbf{Y} = \beta_0 \mathbf{1} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad (1.1)$$

where  $\mathbf{Y}$  is a  $n \times 1$  response vector,  $\mathbf{1}$  is a  $n \times 1$  vector of ones,  $\beta_0$  is an intercept,  $\mathbf{X}$  is an  $n \times k_T$  design matrix,  $\boldsymbol{\beta}$  is a  $k_T \times 1$  vector of regression coefficients, and  $\boldsymbol{\epsilon}$  is an  $n \times 1$  vector of independent and identically distributed errors. A statistical problem of interest is the determination of the relationship between the response  $\mathbf{Y}$  and the  $k_T$  candidate predictors in the columns of  $\mathbf{X}$ . When  $k_T$  is not very small, say  $k_T > 10$ , it often makes sense to assume that only a fraction of the  $k_T$  variables have coefficients that are nonzero. We call those variables “informative” and use  $k_I$  to denote the number of such variables. The researcher is thus faced with two tasks, selecting the informative candidate predictors and estimating their regression coefficients.

Two broad classes of variable selection procedures are subset selection methods and penalty methods. The class of subset selection methods includes procedures such as forward selection, backward elimination, and stepwise regression. The class of penalty methods is composed of methods like the LASSO, adaptive LASSO, and SCAD. Regardless of class affiliation, standard versions of the procedures are based

on least squares methods that assume  $\epsilon$  in (1.1) are normally distributed. The sensitivity of least squares to outlying points creates a need for robust alternatives to these standard variable selection procedures when errors come from a heavier tailed distribution than the normal.

Although there are not many robust variable selection techniques, a variety of robust methods are available for estimating regression coefficients for a specified set of predictor variables. Early robust procedures obtained estimates of the regression parameters by replacing the squared error loss function with one that dampens the effects of outliers (Huber, 1973). More recent robust methods such as the least median of squares (LMS), least trimmed squares (LTS), MM-estimation, and S-estimation are focused on achieving high breakdown (introduced by Hampel, 1971; Donoho and Huber, 1983). None of these robust estimation procedures have built-in variable selection capabilities.

We propose robust variable selection procedures that combine existing variable selection and robust regression procedures. Thus, in the remainder of this chapter we give background information on subset selection and penalty methods as well as on robust regression methods. In Chapter 2 we review the VAMS procedure, a tuned version of forward selection that controls the false selection of unimportant predictors. Robust versions of the procedure are developed that allow for variable selection and the elimination of potential outliers to occur simultaneously. With outliers eliminated, the forward selection algorithm operates in a situation where it is known to perform well. In Chapter 3, existing robust regression procedures are combined with variable selection. This requires the development of robust methods for both obtaining and selecting from a group of candidate models. A new forward addition sequence based on the generalized score statistic (Boos, 1992) is introduced, and a robust BIC is developed. Two LASSO-based robust variable selection methods are outlined in Chapter 4. In Chapter 5 comparisons are made between the procedures from Chapters 2-4 exhibiting the best selection characteristics.

## 1.2 Model Selection Approaches

### 1.2.1 Subset Selection

Subset selection methods often follow a two-step process of obtaining a list of candidate models and using a selection criterion to determine which is optimal. One possibility is to consider all subsets of the candidate predictors as potential models. For such an approach, the number of models,  $2^{k_T}$ , increases exponentially with the number of candidate predictors. While the all-inclusive nature of the approach is appealing, the computational power needed for larger data sets is unreasonable. Methods such as forward selection, backward elimination, and stepwise regression are alternative procedures that provide a more manageable set of candidate models. Although these procedures can be viewed strictly as methods for obtaining candidate models, the classical form of each is identified with a test-based stopping (or selection) criteria. It is also common to select from the set of candidate models generated by each procedure using an information criteria.

### Forward Selection

The forward selection algorithm generates candidate models by sequentially adding predictors to the starting model containing just an intercept. At each step, a new candidate model is formed by adding the predictor with the most significant one degree of freedom F-test for entry into the model (most informative). Following this step-by-step process until all candidate predictors have entered the model leads to a total of  $k_T$  candidate models (in addition to the intercept-only model). Note that the  $j^{th}$  candidate model is composed of the first  $j$  variables in what we call the forward addition sequence (FAS), a list specifying the order in which the  $k_T$  candidate predictors enter the model.

The traditional forward selection algorithm incorporates a stopping criteria directly into the model building process by using an alpha-to-enter value  $\alpha$ . At each step the most informative predictor is only added to the model if the  $p$ -value associ-

ated with the F-statistic for entry is less than or equal to  $\alpha$ . At the first step where the criteria is not met, the model building process stops and the current model is deemed optimal.

## Backward Elimination

Similar to forward selection, the backward elimination procedure specifies  $k_T$  candidate models. Starting from a full model containing all  $k_T$  candidate predictors, the algorithm generates candidate models by successively removing the predictor with the least significant one degree of freedom F-test for removal from the model (most uninformative predictor). For this procedure the  $j^{th}$  candidate model is composed of all variables except the first  $j$  in the backward elimination sequence (BES), a list specifying the order in which the  $k_T$  variables were removed from the model. The requirement of a full model fit limits the use of the procedure to cases where the number of candidate predictors is less than the sample size.

Much like forward selection, the standard backward elimination algorithm incorporates a significance-based stopping criteria into the model building process using an alpha-to-delete value  $\alpha$ . At each step, the least informative predictor is only removed from the model if the  $p$ -value associated with the F-statistic for removal is larger than  $\alpha$ . The final model is obtained when either no variables remain in the model or the criteria for removal is not met.

## Stepwise Regression

The stepwise regression method combines the forward selection and backward elimination algorithms. The method builds a model from one that contains just an intercept. With an empty model the procedure starts like forward selection, with the addition of the most informative candidate predictor. However, unlike forward selection, in each step the procedure is capable of removing predictors no longer considered to be informative. Therefore, tests are performed on all  $k_T$  candidate

predictors. The predictors that are currently in the model are tested for removal while the omitted variables are tested for inclusion.

Following the notation of forward selection and backward elimination, consider using an alpha-to-enter value  $\alpha_{in}$  as a criteria for inclusion and an alpha-to-delete value  $\alpha_{out}$  as a criteria for removal. After each step that a candidate predictor is added to the model, all variables previously in the model are tested for removal. The process continues until no variables meet the criteria to either enter or be removed from the model. Note that unlike both the forward selection and backward elimination algorithms, a single list of candidate models for stepwise regression is not obtainable. The list depends on the values  $\alpha_{in}$  and  $\alpha_{out}$ .

## Information Criteria

Information criteria are often used to select from a set of candidate models. The general form of an information criteria is

$$-2 \log L(\hat{\beta}_0, \hat{\beta}, \hat{\sigma}, \mathbf{Y}) + q(p, n), \quad (1.2)$$

where  $L(\hat{\beta}_0, \hat{\beta}, \hat{\sigma}, \mathbf{Y})$  is a likelihood function evaluated at the maximum likelihood estimates of the parameters and  $q(p, n)$  is a penalty for model complexity with  $p$  equal to the dimension of  $(\hat{\beta}_0, \hat{\beta})$ . Typical choices include the Bayesian information criterion (BIC) with  $q(p, n) = p \log n$  (Schwarz, 1978), Akaike information criterion (AIC) with  $q(p, n) = 2p$ , and Mallows's  $C_p$  (Mallows, 1973), which is approximately equal to the AIC. The common definition of each is developed assuming (1.1) is a Gaussian linear model. The first step to determining an optimal model is to calculate the chosen information criterion for all candidate models. For either forward selection or backward elimination this includes all models represented by the forward addition or backward elimination sequences. The model with the minimum value for the information criteria is then declared optimal. There are other approaches to choosing from a list of candidate models, such as leave-one-out and  $k$ -fold cross-validation, that are related to the information-based approaches discussed here (for example, Shao,

1993; Zhang, 1993).

### 1.2.2 Penalty Methods

Penalty-based variable selection methods obtain a model by solving a penalized version of the least squares problem

$$\hat{\boldsymbol{\beta}} = \arg \min \left\{ \sum_{i=1}^n (Y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2 + p(\boldsymbol{\lambda}, \boldsymbol{\beta}) \right\}. \quad (1.3)$$

where  $\mathbf{x}_i^T$  is the  $i^{th}$  row of the design matrix  $\mathbf{X}$ ,  $p(\cdot)$  is a specified penalty function, and  $\boldsymbol{\lambda}$  is a set of regularization parameters. Note that the  $\mathbf{Y}$  and  $\mathbf{X}$  variables are centered to have mean zero and that the columns of  $\mathbf{X}$  are scaled to have a variance of one. Thus,  $\hat{\boldsymbol{\beta}}_0 = \bar{Y}$  does not appear in (1.3). The penalty function shrinks the regression coefficients of certain variables to zero, leaving a model that contains only a subset of the original predictors. Common penalty methods include the LASSO (Tibshirani, 1996), adaptive LASSO (Zou, 2006), and SCAD (Fan and Li, 2001) procedures. Specific technical details for the LASSO and adaptive LASSO are presented in Chapter 4, with robust alternatives to the procedure.

## 1.3 Robust Regression

### 1.3.1 M-estimation in Regression

Again consider the classical linear model (1.1) with  $\boldsymbol{\epsilon}$  coming from a normal distribution with mean zero. Under these assumptions, the optimal method of estimating regression coefficients is the method of least squares. However, many real data sets contain outlying observations that violate these assumptions. The method of least squares is not optimal for outlier contaminated data sets, failing to provide stable estimates of the regression coefficients. This failure of least squares in the presence of outliers led to the development of a general M-estimation approach to regression.

Consider the problem of obtaining regression coefficients that minimize some function  $\rho(\cdot)$  of the residuals (Huber, 1973),

$$\hat{\boldsymbol{\beta}} = \arg \min \sum_{i=1}^n \rho[Y_i - (\beta_0 + \mathbf{x}_i^T \boldsymbol{\beta})]. \quad (1.4)$$

Differentiating (1.4) with respect to  $\boldsymbol{\beta}$  and setting the result equal to zero yields the estimating equations

$$\sum_{i=1}^n \psi[Y_i - (\beta_0 + \mathbf{x}_i^T \boldsymbol{\beta})] \begin{pmatrix} 1 \\ \mathbf{x}_i \end{pmatrix} = \mathbf{0}, \quad (1.5)$$

where  $\psi$  is the derivative of  $\rho$ . For a general  $\psi$  function, the regression coefficients that solve (1.5) are not invariant to changes in the scale of the dependent variable,  $\mathbf{Y}$ . For this reason, coefficients are obtained using

$$\sum_{i=1}^n \psi \left[ \frac{Y_i - (\beta_0 + \mathbf{x}_i^T \boldsymbol{\beta})}{\hat{\sigma}} \right] \begin{pmatrix} 1 \\ \mathbf{x}_i \end{pmatrix} = \mathbf{0}, \quad (1.6)$$

where  $\hat{\sigma}$  is an estimate of scale used to standardize the residuals. Appropriate methods exist for estimating the scale parameter both prior to and simultaneously with the regression coefficients. In either case an iterative algorithm is generally required to solve (1.6). The final M-estimates of the regression coefficients are approximately normally distributed

$$\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \nu(\mathbf{X}^T \mathbf{X})^{-1}) \quad (1.7)$$

with

$$\nu = \sigma^2 \frac{E\psi(u/\sigma)^2}{[E\psi'(u/\sigma)]^2}, \quad (1.8)$$

where  $u$  follows the distribution of the regression errors and  $\psi'$  is the derivative of the  $\psi$  function. The choice of the  $\psi$  function and the method of estimating  $\hat{\sigma}$  determine the characteristics of the final M-estimates that solve (1.6). Research has shown that choosing a bounded  $\psi$  function, or even one that redescends to zero, leads to estimated regression coefficients that are robust to outlying observations. In Chapter 3 two different approaches based on M-estimation are used to develop robust variable selection procedures.



### 1.3.2 High Breakdown Methods

Hampel (1971, 1974) introduced the concept of breakdown and gross error sensitivity as measures of robustness. A finite sample version of breakdown was later developed by Donoho and Huber (1983). The finite sample breakdown point is defined to be the largest percentage of outlying points a sample can contain without completely ruining the estimate of interest. Define the maximum bias of an estimator  $T$  to be

$$b(\epsilon; \mathbf{X}, T) = \sup |T(\mathbf{X}^{(c)}) - T(\mathbf{X})| \quad (1.9)$$

where  $\mathbf{X}$  is the original sample,  $\mathbf{X}^{(c)}$  is the contaminated sample, and  $\epsilon$  is the percentage of contamination. In (1.9) the supremum is taken over all  $\epsilon$ -contaminated samples  $\mathbf{X}^{(c)}$ . As defined by (Donoho and Huber, 1983) the breakdown point of the estimator  $T$  is then

$$\epsilon^*(\mathbf{X}, T) = \inf\{\epsilon | b(\epsilon; \mathbf{X}, T) = \infty\}. \quad (1.10)$$

Note that definition (1.10) is for estimators that are unbounded. In practice  $\epsilon^*$  is often found by observing the minimum percentage of contamination that causes an estimator to take on an arbitrarily large value. For bounded estimators it is equivalent to finding the percentage of contaminated data that will force an estimator to the boundary of the parameter space (Donoho and Huber, 1983).

Consider two simple estimators of location, the mean and the median. Since a single large observation can cause the mean to become arbitrarily large, the statistic has a breakdown point of 0. The median on the other hand has a breakdown point of .5 since half the observations must be contaminated before the statistic fails to provide meaningful information about the original sample (Andrews et al., 1972). Note that a breakdown point of .5 is the largest possible since for any higher level of contamination it is impossible to distinguish between the original and contaminated sample points.

In a regression setting, a method that has good breakdown properties is able to give meaningful estimates of the regression coefficients for highly contaminated data sets. Contamination can be in the form of outlying points in either the response or independent variables. In this section we present three high breakdown regression methods that are used throughout the rest of the thesis: the LMS, LTS, and MM-estimation procedures.

## LMS and LTS procedure

The traditional least squares estimate is the one that minimizes the sum of squares of the residuals. As discussed in section (1.1), the first robust alternatives to the least squares method modified the basic procedure by replacing the square function by something else. The most obvious example is the L1 estimate that replaces the square with the absolute value, and minimizes the sum of the absolute residuals. Rousseeuw (1984) takes a different approach and instead replaces the sum with the median to get

$$\min_{\boldsymbol{\beta}} \text{med} \{r_1(\boldsymbol{\beta})^2, \dots, r_n(\boldsymbol{\beta})^2\}, \quad (1.11)$$

where  $r_i(\boldsymbol{\beta}) = [Y_i - (\beta_0 + \mathbf{x}_i^T \boldsymbol{\beta})]$  is the residual for the  $i^{th}$  observation. The solution to (1.11) is the Least Median of Squares (LMS) estimate. Rousseeuw (1984) shows that the LMS estimate that solves (1.11) is both guaranteed to exist and obtains the optimal breakdown point of fifty percent. A drawback of the LMS procedure is that the estimates have low efficiency.

Efficiency of estimates is not too important when the estimates are used as a starting point for an iterative procedure. For this reason, Simpson et al. (1992) and Rousseeuw (1984) use the LMS procedure to obtain preliminary estimates for their one-step M-estimators. Yohai (1987) takes a similar approach when constructing MM-estimates. We follow these authors and use the LMS procedure strictly to generate starting values for our robust regression variable selection procedures in Chapter 3.

The low efficiency of the LMS procedure led Rousseeuw (1984) to develop the least trimmed squares estimator (LTS) obtained by

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^h r_{i:n}^2(\boldsymbol{\beta}), \quad (1.12)$$

where  $r_{1:n}^2(\boldsymbol{\beta}) \leq r_{2:n}^2(\boldsymbol{\beta}) \leq \dots \leq r_{n:n}^2(\boldsymbol{\beta})$  are the ordered squared regression residuals. With an appropriate choice of  $h$ , the  $\hat{\boldsymbol{\beta}}$  achieving the minimum in (1.12) is a high breakdown regression estimate that also has high efficiency.

## MM-estimation

Like the LTS procedure of the last section, MM-estimation gives regression estimates that not only have high breakdown but also high efficiency at the normal distribution. Estimates are a modification of the typical M-estimates and are obtained through a three-step process (Maronna, 2006, p. 125):

1. An initial consistent estimate  $\hat{\boldsymbol{\beta}}^{(0)}$  is computed that has high breakdown but not necessarily high efficiency. For example,  $\hat{\boldsymbol{\beta}}^{(0)} = \hat{\boldsymbol{\beta}}_{LMS}^{(0)}$ , the least median of squares estimate. Although  $\hat{\boldsymbol{\beta}}_{LMS}^{(0)}$  is inefficient it has the optimal breakdown point of .5.
2. A robust scale estimate  $\hat{\sigma}$  is computed from the residuals of the high breakdown fit of the previous step. Using the residuals from the LMS fit, the robust estimate of scale,  $\hat{\sigma}$ , is found that solves the M-estimating equations

$$\frac{1}{n} \sum_{i=1}^n \rho\left(\frac{r_i}{1.56\hat{\sigma}}\right) = .5. \quad (1.13)$$

Here  $r_i$  is the  $i^{th}$  residual from the LMS fit, and the constant 1.56 is chosen to ensure consistency of the scale estimate at the normal distribution. A standard choice of  $\rho$  is the bisquare function

$$\rho_B(x) = \min \{1, 1 - (1 - x^2)^3\}. \quad (1.14)$$

Like the initial estimate of the regression coefficients  $\hat{\beta}^{(0)}$ , the scale estimate obtained from (1.13) has a breakdown point of .5 (Maronna, 2006, p. 125).

3. The final MM estimate  $\hat{\beta}$  is found using an iterative procedure with  $\hat{\beta}^{(0)}$  as a starting point. Using the estimate of scale found in the previous step the final MM-estimates are found by solving

$$\hat{\beta} = \arg \min \sum_{i=1}^n \rho \left( \frac{r_i(\beta)}{c\hat{\sigma}} \right), \quad (1.15)$$

where  $\rho(\cdot)$  is again the bi-square function and the scaling constant  $c$  is now chosen to have a desired efficiency at the normal distribution. Maronna (2006, p. 30) presents a table showing the values of  $c$  for typical target levels of efficiency. For example,  $c=4.68$  yields 95% efficiency at the normal distribution. The only restriction on  $c$  is that it is larger than 1.56, a requirement that is met for all reasonable efficiency levels.

As discussed in Section 1.3.1, solving (1.15) requires the use of an iterative procedure. Maronna (2006, p. 125) suggests the use of an iteratively re-weighted least squares algorithm. For details of this procedure see Section 3.1. The final MM-estimates output by the iterative procedure are guaranteed to not have a breakdown point less than that of  $\hat{\beta}^{(0)}$ , and therefore share the optimal breakdown point of .5 with the LMS procedure. With a good choice of  $c$  in (1.15) the MM-estimates have both high breakdown and high efficiency at the normal distribution.

## Chapter 2

# Variable Addition Model Selection

In this chapter the variable addition model selection (VAMS) procedure is modified to handle outlying points. Sections 2.1 and 2.2 outline the variable selection problem and introduce the non-robust versions of the procedure. Then we introduce the use of indicator variables and develop robust versions of the VAMS procedure that simultaneously select variables and eliminate outliers. Finally, Monte Carlo simulation is used to examine and compare the robust methods.

### 2.1 Simulation-Based VAMS

Using (1.1) as a starting point, the goal of a variable selection procedure is to determine which of the  $k_T$  explanatory variables are informative. An informative variable is defined as a variable having an associated  $\beta_i$  that is nonzero, while variables with coefficients equal to zero are deemed uninformative. Ideally, when presented with a matrix of potential predictors, a variable selection procedure would choose a model containing only informative variables. However, it is not possible for the variable selection procedure to make this selection precisely, resulting in a model that is either over-fit or under-fit. A method is characterized as over-fitting if it tends to select a significant number of uninformative variables along with the informative variables of interest. Under-fitting means the selection of a small number of uninformative

predictors along with the failure to select informative variables with small but positive regression coefficients. For a given data set  $(\mathbf{Y}, \mathbf{X})$ , the tendency of a particular method to over-fit or under-fit is related to the false selection rate (FSR), defined to be

$$\gamma_0 = E \left\{ \frac{U(\mathbf{Y}, \mathbf{X})}{1 + I(\mathbf{Y}, \mathbf{X}) + U(\mathbf{Y}, \mathbf{X})} \right\}, \quad (2.1)$$

where  $U(\mathbf{Y}, \mathbf{X})$  and  $I(\mathbf{Y}, \mathbf{X})$  are the number of selected uninformative and informative predictors, respectively. The expectation in (2.1) is taken with respect to repeated sampling of the data. With the addition of 1 in the denominator for the intercept, (2.1) is the expected proportion of uninformative predictors in the model.

Consider the case when data are entered into a variable selection procedure such as forward selection. Depending on the alpha-to-enter tuning parameter  $\alpha$ , the forward selection procedure selects models of differing size. In fact, the number of variables selected is monotone in the tuning parameter  $\alpha$ . As  $\alpha$  increases from 0 to 1, the model grows from a model containing just an intercept to one that contains every predictor in the design matrix.

For a specific data set  $(\mathbf{Y}, \mathbf{X})$ , let  $S(\alpha)$  be the total number of variables selected when applying the forward selection algorithm with a tuning parameter of  $\alpha$ . Similarly, define  $U(\alpha)$  and  $I(\alpha)$  to be the number of uninformative and informative variables, respectively, in the selected model. The VAMS procedure introduced by Wu et al. (2007) tunes the forward selection algorithm by providing an estimate of  $\alpha$  that attempts to keep the average false selection approximately equal to a target value  $\gamma_0$ .

Theoretically, this balance could be accomplished by estimating  $\alpha$  in the following way:

$$\hat{\alpha} = \sup_{\alpha} \{ \alpha : \hat{\gamma}(\alpha) \leq \gamma_0 \}, \quad (2.2)$$

where  $\hat{\gamma}(\alpha)$  is a function that estimates (2.1). By taking the supremum we obtain the largest possible model while keeping the FSR below the target rate  $\gamma_0$ . If  $U(\alpha)$  were known, a natural choice for  $\hat{\gamma}(\alpha)$  would be the empirical estimator  $U(\alpha)/\{1 + I(\alpha) + U(\alpha)\}$ . In practice, when  $U(\alpha)$  is unknown,  $\hat{\gamma}(\alpha)$  is obtained by replacing  $U(\alpha)$  in

the empirical estimator with a suitable estimate.

The VAMS procedure estimates  $U(\alpha)$  by monitoring the selection of user-created pseudo variables. These pseudo variables are uninformative variables generated so that two key assumptions are satisfied approximately. The first is that the true unimportant variables and the pseudo unimportant variables have on average the same probability of selection. The second is that the inclusion of the pseudo variables does not change the probability that an informative variable is chosen by the variable selection procedure. Wu et al. (2007) proposed the following pseudo variable generation procedure.

Let  $\mathbf{X}_p$  be the matrix formed by randomly permuting the rows of the design matrix  $\mathbf{X}$ . Then a matrix  $\mathbf{Z}$ , whose  $k_T$  columns consist of pseudo variables, can be formed using the formula

$$\mathbf{Z} = (\mathbf{I} - \mathbf{P}_{\mathbf{X}})\mathbf{X}_p. \quad (2.3)$$

Here,  $\mathbf{I}$  is a  $n \times n$  identity matrix and  $\mathbf{P}_{\mathbf{X}}$  is the projection matrix  $\mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$ . The  $i^{th}$  pseudo variable is then just the residuals from the regression of the  $i^{th}$  column of  $\mathbf{X}_p$  on the predictors contained in  $\mathbf{X}$ . The orthogonality of residuals to regressors, inherent in linear regression, guarantees that the pseudo variables have a sample correlation of zero with the original predictors.

After creation of the pseudo variables, the matrix  $\mathbf{Z}$  is appended to the original design matrix  $\mathbf{X}$ , resulting in the augmented design matrix  $\mathbf{X}_V = \mathbf{X} : \mathbf{Z}$  of dimension  $n \times 2k_T$ . The forward selection algorithm is then applied to the new data set  $(\mathbf{Y}, \mathbf{X}_V)$ . Let  $U_p(\alpha)$  be the number of pseudo variables selected when using a tuning parameter of  $\alpha$ . This process of generating pseudo variables and monitoring their selection is replicated for different permutations of the design matrix  $\mathbf{X}$ . Then, for a fixed  $\alpha$ , the rate  $\theta$  at which the pseudo variables enter the model is estimated as

$$\hat{\theta}(\alpha) = \frac{B^{-1} \sum_{i=1}^B U_p(\alpha)}{k_T} = \frac{\bar{U}_p(\alpha)}{k_T}, \quad (2.4)$$

where  $B$  is the number of Monte Carlo replicates of pseudo variable generation and  $\bar{U}_p(\alpha)$  is the average number of pseudo variables selected across these replicates. The

assumption that the real unimportant variables and pseudo unimportant variables have the same probability of being selected allows us to use (2.4) as an estimate of the rate at which the true unimportant variables are entering the model. Multiplying this rate by an estimate of the number of uninformative candidate predictors

$$\widehat{k}_u(\alpha) = k_T - S(\alpha), \quad (2.5)$$

results in an estimate of  $U(\alpha)$ ,  $\{k_T - S(\alpha)\}\widehat{\theta}(\alpha)$ . In (2.5),  $S(\alpha)$  estimates the number of candidate predictors in  $\mathbf{X}$  that are informative.

Using the estimate of  $U(\alpha)$ , we can obtain the false selection rate estimate

$$\widehat{\gamma}(\alpha) = \frac{\{k_T - S(\alpha)\}\widehat{\theta}(\alpha)}{1 + S(\alpha)}. \quad (2.6)$$

Substituting (2.6) into (2.2), results in an estimated  $\alpha$  given by

$$\widehat{\alpha} = \sup_{\alpha \leq \alpha_m} \{\alpha : \widehat{\gamma}(\alpha) \leq \gamma_0\}, \quad (2.7)$$

where  $\alpha_m$  restricts the search for  $\alpha$  to be less than a predetermined maximum possible value (generally .3). Using (2.7) as the tuning parameter for forward selection with the original data set  $(\mathbf{Y}, \mathbf{X})$  leads to a selection procedure that controls the false selection rate. In practice,  $\widehat{\gamma}(\alpha)$  is calculated for a finite grid of  $\alpha$  values, and the supremum in (2.7) is replaced by a maximum. By taking a fine enough grid there is little effect of the grid on the estimate of  $\alpha$ .

## 2.2 Fast VAMS

After development of the standard VAMS procedure in the previous section, Boos et al. (2008) proposed the simplified Fast VAMS procedure. Rather than estimating the rate at which uninformative variables enter the model using (2.4), the new procedure sets  $\widehat{\theta}(\alpha)$  equal to  $\alpha$ . The Fast VAMS procedure therefore does not require the use of simulation. The new estimate of the FSR is then

$$\widehat{\gamma}_F(\alpha) = \frac{\{k_T - S(\alpha)\}\alpha}{1 + S(\alpha)}, \quad (2.8)$$



which leads to

$$\hat{\alpha}_F = \sup_{\alpha \leq \alpha_m} \{\alpha : \hat{\gamma}_F(\alpha) \leq \gamma_0\}. \quad (2.9)$$

Like the simulation version of the procedure, the final model is obtained using  $\hat{\alpha}_F$  as the tuning parameter for forward selection with the original data  $(\mathbf{Y}, \mathbf{X})$ .

Boos et al. (2008) show that this is equivalent to selecting the model with size

$$k(\gamma_0) = \max\{i : \tilde{p}_i \leq \frac{\gamma_0[1+i]}{k_T - i} \text{ and } \tilde{p}_i \leq \alpha_m\} \quad (2.10)$$

where  $\tilde{p}$  is the sequence of monotonized  $p$ -values, that is,

$$\tilde{p}_i = \max p_k \text{ for } k = 1, \dots, i. \quad (2.11)$$

In (2.11),  $p_k$  is the  $p$ -value to enter for the  $k^{th}$  variable in the forward addition sequence. An advantage of using (2.10) is that it does not require the computation of  $\hat{\alpha}_F$ , that is, the final model can be obtained from examination of the sequential  $p$ -values of the variables in the forward addition sequence.

Simulation results show that in addition to the large reduction in computation, the Fast VAMS procedure leads to results comparable to the ones obtained from the simulation-based approach.

## 2.3 Simultaneous Selection and Outlier Detection

Mickey et al. (1967) described a modified approach to regression for detecting outliers. Let  $\mathbf{X}^* = \mathbf{X} : \mathbf{I}$  where  $\mathbf{I}$  is an  $n \times n$  identity matrix. For the moment, we assume that the  $n \times k_T$  design matrix  $\mathbf{X}$  is pre-specified and contains all regression variables of interest. Then, stepwise regression is run using  $\mathbf{X}^*$  with all of the  $\mathbf{X}$  variables already included in the model. For an indicator variable, say the  $j^{th}$  column of  $\mathbf{I}$ , we shall see that the test for entry into the model is identical to a significance test on the  $j^{th}$  PRESS residual and that inclusion in the model is equivalent to deleting the  $j^{th}$  observation from the data set. Thus, Mickey et al. (1967) proposed a simple sequential approach for identifying and deleting outliers.

Consider adding the  $j^{th}$  column of the identity matrix, denoted  $\mathbf{I}_j$ , to the regression of  $\mathbf{Y}$  on the predictors in the design matrix  $\mathbf{X}$ . The resulting regression model is

$$E(\mathbf{Y}) = \beta_0 \mathbf{1} + \mathbf{X}\boldsymbol{\beta} + \mathbf{I}_j \delta_j, \quad (2.12)$$

where  $\beta_0$ ,  $\boldsymbol{\beta}$ , and  $\delta_j$  are the regression coefficients for the intercept,  $\mathbf{X}$ , and the indicator variable  $\mathbf{I}_j$ , respectively. The standard  $n - k_T - 2$  degree of freedom  $t$ -statistic

$$t_{n-k_T-2}(\delta_j) = \frac{\widehat{\delta}_j}{SE(\widehat{\delta}_j)}, \quad (2.13)$$

tests the hypothesis  $H_0 : \delta_j = \mathbf{0}$  vs  $H_a : \delta_j \neq \mathbf{0}$ , and measures the importance of including  $\mathbf{I}_j$  in the model. In (2.13),  $\widehat{\delta}_j$  is the least squares estimate of the regression coefficient for  $\mathbf{I}_j$ , and  $SE(\widehat{\delta}_j)$  is the standard error of this estimate. Since  $\mathbf{I}_j$  is a variable which affects only a single observation,  $\widehat{\delta}_j$  is the value that causes the  $j^{th}$  residual to obtain its minimum at zero. That is,

$$\widehat{\delta}_j = \{Y_j - (\widehat{\beta}_0 + \mathbf{x}_j^T \widehat{\boldsymbol{\beta}})\}, \quad (2.14)$$

where  $\mathbf{x}_j^T$  is the  $j^{th}$  row of  $\mathbf{X}$ , and  $\widehat{\beta}_0$  and  $\widehat{\boldsymbol{\beta}}$  are the least squares estimates of  $\beta_0$  and  $\boldsymbol{\beta}$  when  $\mathbf{Y}$  is regressed on the predictors in  $\mathbf{X}$  and the indicator variable  $\mathbf{I}_j$ . Note that with the  $j^{th}$  residual necessarily equal to zero,  $\widehat{\beta}_0$  and  $\widehat{\boldsymbol{\beta}}$  are found by minimizing the sum of squares of the remaining  $n - 1$  residuals. That is

$$(\widehat{\beta}_0, \widehat{\boldsymbol{\beta}}) = \arg \min \sum_{i=1, i \neq j}^n \{Y_i - (\beta_0 + \mathbf{x}_i^T \boldsymbol{\beta})\}^2. \quad (2.15)$$

In terms of estimation, the coefficients obtained using (2.15) are the same as those obtained from the least squares regression of  $\mathbf{Y}$  on  $\mathbf{X}$  with the  $j^{th}$  observation removed from the data set. Returning to (2.14), we see that  $\widehat{\delta}_j$  is then actually the  $j^{th}$  PRESS residual, a common outlier diagnostic, equal to the  $j^{th}$  response minus the predicted response for  $\mathbf{x}_j^T$  when the  $j^{th}$  observation is excluded from the data set. An alternate formula for calculating the  $j^{th}$  PRESS residual using just the full data set is

$$PRESS_j = \frac{r_j}{1 - P_x^{jj}} \quad (2.16)$$

where  $r_j$  is the  $j^{th}$  least squares residual and  $P_x^{jj}$  is  $j^{th}$  diagonal element of the projection matrix  $\mathbf{P}_x = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$ . An advantage of using (2.16) is that all  $n$  PRESS residuals can be calculated using just a single least squares fit.

Since the stepwise regression procedure uses (2.13) to determine which indicators enter the model, the final set of indicators are those with the most significant PRESS residuals. These indicators are therefore associated with the most outlying or influential observations. It is easy to show that for a model containing  $v$  indicators, (2.15) extends to obtaining regression coefficients  $(\hat{\beta}_0, \hat{\beta})$  that minimize  $n - v$  squared residuals. Therefore, final regression coefficients are calculated with the observations associated with the set of selected indicators removed from the data set.

To move from regression with outlier detection to variable selection with outlier detection, consider running the forward selection procedure with the augmented design matrix  $\mathbf{X}^* = \mathbf{X} : \mathbf{I}$ , containing both potential predictors and indicator variables. At each step of the algorithm, a variable is added to the model if it most significantly reduces the error sum of squares. Selection of an indicator variable implies that the greatest reduction in the error sum of squares is obtained through removal of a single observation. As in the fixed regression case, only indicator variables associated with influential or outlying observations are likely to be selected. In terms of the potential predictors contained in  $\mathbf{X}$ , the amount a variable reduces the error sum of squares is directly related to the level in which its regression coefficient significantly differs from zero. This equivalence relationship implies that the variables selected from  $\mathbf{X}$  will be the ones that are deemed most informative by the data. Thus, the forward selection algorithm sequentially selects potentially informative predictors and eliminates outliers by including columns of the identity matrix. As an example, in Table 2.1 we list the first twenty variables of the forward addition sequence for  $\mathbf{X}^*$ .

Note that Mickey et al. (1967) did not suggest the use of stepwise regression on both the indicators and original variables. Their stepwise search was just for the indicator variables, and thus merely a search for outliers in a fixed regression model. By applying forward selection to the augmented design matrix  $\mathbf{X}^*$  we have

Table 2.1: Forward Addition Sequence for  $\mathbf{X}^*$ 

Order	Var	$p$ -value	Order	Var	$p$ -value
1	$X_4$	0.000035	11	$I_{67}$	0.029
2	$X_2$	0.000047	12	$I_{31}$	0.021
3	$I_1$	0.000086	13	$I_{40}$	0.022
4	$I_{98}$	0.0015	14	$I_{24}$	0.018
5	$X_3$	0.0018	15	$I_{27}$	0.020
6	$X_1$	0.0018	16	$I_{34}$	0.027
7	$I_{74}$	0.0011	17	$I_{32}$	0.015
8	$I_{10}$	0.0080	18	$I_{23}$	0.051
9	$I_{97}$	0.022	19	$X_{18}$	0.025
10	$I_{59}$	0.024	20	$I_{81}$	0.034

a procedure that not only performs variable selection but does so while adjusting for outlying points. The inclusion of indicator variables in the early steps of the selection process means that variables are being considered for entry into the model with multiple outlying points eliminated. If the eliminated outlying points are influential, their removal could significantly change the order in which the remaining variables enter the model.

As with any application of forward selection, we are faced with the choice of an appropriate stopping criteria. By incorporating the indicator variables into either the original simulation-based VAMS procedure (Section 2.1) or fast VAMS procedure (Section 2.2), we can obtain an estimated alpha-to-enter value  $\hat{\alpha}$  that controls the false selection of the original predictors in  $\mathbf{X}$ . In this section we focus on the simulation-based indicator approach, a procedure we call VAMSI, but note that all the quantities needed to use the fast VAMS procedure are also present.

Recall the key quantity to estimate is

$$\hat{\gamma}(\alpha) = \frac{\{k_T - S(\alpha)\}\hat{\theta}(\alpha)}{1 + S(\alpha)}. \quad (2.17)$$

Here the value of  $k_T$  in (2.17) remains unchanged, still equaling the number of candidate predictors in the original design matrix  $\mathbf{X}$ . The quantity  $S(\alpha)$  is still defined to

be the number of original predictors selected, however it is now the result of applying the forward selection algorithm to the data  $(\mathbf{Y}, \mathbf{X}^* = \mathbf{X} : \mathbf{I})$  instead of to  $(\mathbf{Y}, \mathbf{X})$ . In effect, we just ignore the selection of any indicators. As an example, consider the forward addition sequence for  $\mathbf{X}^*$  presented in Table 2.1. If we assume  $\alpha = .02$ , the forward selection algorithm would choose a model containing the first eight variables in the sequence ( $I_{97}$  is first variable with a  $p$ -value to enter greater than  $\alpha = .02$ ). Since only four of the eight are non-indicator variables coming from  $\mathbf{X}$ ,  $S(.02) = 4$ . Note, while indicators are ignored when calculating  $S(\alpha)$ , their inclusion in the forward selection process adjusts the selection of the variables in  $\mathbf{X}$  (and hence the value of  $S(\alpha)$ ) for the possibility of outliers.

The next step is to estimate the rate  $\hat{\theta}(\alpha)$  at which the uninformative predictors are entering the model. This rate is again estimated by monitoring the selection of user created pseudo variables. Let  $\mathbf{X}_{VI}$  be the matrix formed by appending an  $n \times n$  identity matrix to the original VAMS design matrix  $\mathbf{X}_V$ , that is,

$$\mathbf{X}_{VI} = (\mathbf{X}_V : \mathbf{I}) = (\mathbf{X} : \mathbf{Z} : \mathbf{I}). \quad (2.18)$$

The psuedo random variables in  $\mathbf{Z}$  are generated using (2.3) and the original design matrix  $\mathbf{X}$ . It should be noted that the design matrix in (2.18) contains  $2k_T + n$  candidate predictors. Since the forward selection algorithm sequentially builds from a model including only an intercept, having more predictors than observations does not pose a problem. However, if this procedure were to be extended to a variable selection algorithm that required a full model fit, such as backward selection, having more predictors than cases would be problematic.

For a particular  $\alpha$  and matrix  $\mathbf{Z}$ ,  $U_p(\alpha)$  is now the number of pseudo variables selected (ignoring the selection of indicators and original variables) when applying the forward selection algorithm to the data set  $(\mathbf{Y}, \mathbf{X}_{VI})$ . The rate at which uninformative predictors enter the model  $\hat{\theta}(\alpha)$ , is then estimated using (2.4). With indicator adjusted definitions of  $k_T$ ,  $S(\alpha)$  and  $\hat{\theta}(\alpha)$  we have all the quantities needed to use (2.17) to obtain an estimate  $\hat{\alpha}$  of the forward selection tuning parameter. By applying forward selection with a tuning parameter of  $\hat{\alpha}$  to the data set  $(\mathbf{Y}, \mathbf{X}^* = \mathbf{X} : \mathbf{I})$ , the

VAMSI procedure not only controls the false selection rate of the candidate predictors but eliminates potentially outlying points. This is advantageous since traditionally the two tasks have been treated separately.

## 2.4 Outlier Detection Problems for VAMSI

In this section we examine the merits of the VAMSI procedure as an outlier elimination method. In particular we find that a single  $\hat{\alpha}$  cannot be used for selecting both original variables and indicators. Consider the simple linear regression data set in Figure 2.1.

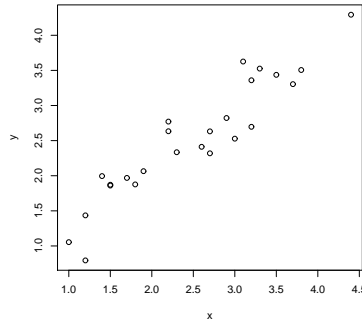


Figure 2.1: Plot of response versus predictor

This data set is one that is ideal for simple linear regression. There is a strong linear relationship between the response vector  $\mathbf{y}$  and the single predictor  $\mathbf{x}$ . With just a single predictor this is not a data set that is normally analyzed by a variable selection procedure. However, we passed the data set to the VAMSI procedure to investigate the procedure's ability to identify outlying observations. The VAMSI procedure identified every observation as an outlier, an inappropriate result due to the fact that the selection of indicators is ignored when estimating the tuning parameter.

The result obtained for the data in Figure 2.1 is an extreme case. The VAMSI

procedure fails to select an appropriate set of indicators when there is either very few candidate predictors or all candidate predictors are strongly significant. First consider the case where there are very few candidate predictors. If indicators are ignored, it is not uncommon to have two adjacent candidate predictors in the forward addition sequence requiring drastically different estimates of  $\alpha$  to enter the model. For example, consider the variables  $X_1$  and  $X_{18}$  in Table 2.1. Ignoring indicators, these two variables are adjacent in the forward addition sequence, but  $X_1$  enters the model for any value of  $\alpha$  greater than or equal to .0018 while  $X_{18}$  needs a value greater than or equal to .051. Notice, for any value of  $\alpha$  between .0018 and .051, the forward selection algorithm chooses the same set of original predictors ( $X_4, X_2, X_1, X_3$ ), making all these models equivalent in terms of the VAMSI procedure. However, two different estimates within this gap can lead to a drastically different sets of indicators being selected. Specifically, consider the models chosen by forward selection when using  $\alpha = .02$  and  $\alpha = .05$ . With  $\alpha = .02$ , four indicators are selected along with the four original predictors. For  $\alpha = .05$  the number of indicators selected grows to thirteen. A similar issue develops in the second case where all the candidate predictors are highly significant. The VAMSI gives an inflated estimate of  $\alpha$  to ensure that all the original predictors enter the model. Unfortunately, as we saw in the simple linear regression example, the inflated  $\alpha$  can be large enough to also let in all the indicator variables.

To fix these problems we propose a dual estimation approach where tuning parameter estimate  $\hat{\alpha}_X$  is used to select original variables and a separate estimate  $\hat{\alpha}_{\text{IND}}$  is used to select indicators. Tuning the selection of the indicator variables should result in a procedure with improved ability to properly identify outliers. In Section 2.4.1 we present three variations of the original VAMSI procedure. In Section 2.4.2 we outline a two step approach that estimates the tuning parameters for the two types of variables in different stages.

### 2.4.1 Variations of VAMSI

All three methods presented in this section use the VAMSI procedure (Section 2.3) to obtain  $\hat{\alpha}_X$ . Remember that this method uses the simulation-based estimate  $\hat{\theta}(\alpha)$  of the rate at which true unimportant variables enter the model to get an estimated tuning parameter that controls of the false selection of the original predictors. Where the methods differ is in their approach to obtaining  $\hat{\alpha}_{\text{IND}}$ , the tuning parameter for the indicators. First we present the methods for obtaining  $\hat{\alpha}_{\text{IND}}$  and then show how these estimates are combined with the  $\hat{\alpha}_X$  to select the final model.

#### **VAMSI<sub>1</sub>**

The VASMI<sub>1</sub> procedure takes the simplest approach and obtains  $\hat{\alpha}_{\text{IND}}$  using the simulation-free fast VAMS procedure. Let  $k_{\text{IND}}$  be the number of observations  $n$ , that is, the number of indicator variables. Also, let  $S_{\text{IND}}(\alpha)$  be the number of indicators that enter the model when applying the forward selection algorithm with a tuning parameter of  $\alpha$  to the data  $(\mathbf{Y}, \mathbf{X}^* = \mathbf{X} : \mathbf{I})$ . With  $k_{\text{IND}}$  replacing  $k_T$  and  $S_{\text{IND}}(\alpha)$  replacing  $S(\alpha)$  in (2.8), appropriate estimates of  $\gamma$  and  $\alpha$  for the indicators are

$$\hat{\gamma}_{\text{IND}}(\alpha) = \frac{\{k_{\text{IND}} - S_{\text{IND}}(\alpha)\}\alpha}{1 + S_{\text{IND}}(\alpha)} = \frac{\{n - S_{\text{IND}}(\alpha)\}\alpha}{1 + S_{\text{IND}}(\alpha)} \quad (2.19)$$

and

$$\hat{\alpha}_{\text{IND}} = \sup_{\alpha \leq \alpha_m} \{\alpha : \hat{\gamma}_{\text{IND}} \leq \gamma_0\} \quad (2.20)$$

As with the previous VAMS methods  $\alpha_m$  is a restriction placed on the maximum value of  $\alpha$  (generally equal to .3) and  $\gamma_0$  is the target false selection rate.

#### **VAMSI<sub>2</sub>**

The VAMSI<sub>2</sub> method obtains an alternate fast VAMS estimate  $\hat{\alpha}_{\text{IND}}$  by replacing  $S_{\text{IND}}(\alpha)$  in (2.19) with the simulation-based estimate of the number of informative indicators. For the same simulation data sets used by the VAMSI procedure to



estimate  $\widehat{\theta}(\alpha)$ , we calculate

$$S_{\text{IND2}}(\alpha) = \frac{1}{B} \sum_{i=1}^B S_{\text{INDP}}(\alpha), \quad (2.21)$$

where  $B$  is again the number of Monte Carlo replicates of pseudo variable generation and  $S_{\text{INDP}}(\alpha)$  is the number of indicators selected when the forward selection algorithm is applied to the data  $(\mathbf{Y}, \mathbf{X}_{VI} = \mathbf{X} : \mathbf{Z} : \mathbf{I})$ . Note that the value of  $S_{\text{INDP}}(\alpha)$  in (2.21) differs from  $S_{\text{IND}}(\alpha)$  for the VASMI<sub>1</sub> method. In that case,  $S_{\text{IND}}(\alpha)$  was the number of indicators selected when the forward selection algorithm was applied to the data set  $(\mathbf{Y}, \mathbf{X}^* = \mathbf{X} : \mathbf{I})$  containing no pseudo variables. Averaging across the bootstrap replicates to obtain (2.21) gives an estimate of the number of informative indicators that takes into account the variation across the different simulated data sets. Also, since the calculation of  $\widehat{\alpha}_X$  already requires forward selection to be applied to each simulated data set, the VAMSI<sub>2</sub> procedure is computationally equivalent to either the VAMSI or VASMI<sub>1</sub> procedures.

Plugging (2.21) in to (2.19) gives

$$\widehat{\gamma}_{\text{IND}}(\alpha) = \frac{\{n - S_{\text{IND2}}(\alpha)\}\alpha}{1 + S_{\text{IND2}}(\alpha)}, \quad (2.22)$$

a new estimate of the false selection rate that can be used to calculate  $\widehat{\alpha}_{\text{IND}}$ .

### **VAMSI<sub>3</sub>**

Both the VASMI<sub>1</sub> and VAMSI<sub>2</sub> use some form of the fast VAMS procedure to obtain an estimated forward selection entry criterion for the indicator variables. The VAMSI<sub>2</sub> method used simulation to obtain an alternate estimate of the number of informative indicators, but  $\widehat{\alpha}_{\text{IND}}$  was still found using (2.19) and (2.20). In this section a procedure is developed that uses the simulation version of the VAMS procedure to obtain  $\widehat{\alpha}_{\text{IND}}$ .

When dealing with the indicator variables the generation of pseudo variables is not straightforward. For the original simulation-based VAMS procedure (Section 2.1),

pseudo variables are generated for the candidate predictors through permutation of the design matrix  $\mathbf{X}$ . When considering a matrix of indicators the permutation method is no longer appropriate. Random permutation of the rows of  $\mathbf{I}$  would result in a matrix containing the same set of variables. Wu et al. (2007) outlines alternative approaches to pseudo variable generation. In one such approach, psuedo variables are generated by simply taking a random sample from a standard normal distribution. As with the permutation-based method, the resulting pseudo variables are stochastically uncorrelated with the true predictors (Wu et al., 2007).

Using the new approach, a matrix  $\mathbf{Z}_I$  consisting of  $n$  standard normal pseudo variables is formed. For a chosen alpha, let  $U_z(\alpha)$  be the number of these standard normal pseudo variables selected when the forward selection algorithm is applied to the data  $(\mathbf{Y}, [\mathbf{X} : \mathbf{Z}_I : \mathbf{I}])$ . Redefining (2.4), the rate which unimportant indicators are entering the model can be estimated as

$$\hat{\theta}_{IND}(\alpha) = \frac{B^{-1} \sum_{i=1}^B U_z(\alpha)}{n} = \frac{\overline{U}_z(\alpha)}{n}, \quad (2.23)$$

leading to

$$\hat{\gamma}_{IND}(\alpha) = \frac{\{n - S_{IND}(\alpha)\} \hat{\theta}_{IND}(\alpha)}{1 + S_{IND}(\alpha)}. \quad (2.24)$$

We now have all the quantities needed to obtain the VAMSI<sub>3</sub> simulation-based estimate  $\hat{\alpha}_{IND}$  for the indicator variables.

## Determining the Final Model

Regardless of the method used to find  $\hat{\alpha}_{IND}$ , the final model is determined through two separate applications of the forward selection algorithm to the data  $(\mathbf{Y}, \mathbf{X}^* = \mathbf{X} : \mathbf{I})$ . Specifically, the set of indicators are selected using a forward selection tuning parameter of  $\hat{\alpha}_{IND}$ , and the original predictors are selected using the tuning parameter  $\hat{\alpha}_X$ . To illustrate this process with Table 2.1, suppose that for one of the three procedures above  $\hat{\alpha}_X = .028$  and  $\hat{\alpha}_{IND} = .001$ . The result of applying forward selection with a tuning parameter of .028 would be a model consisting of the

variables  $X_4, X_2, I_1, I_{98}, X_3, X_1, I_{74}, I_{10}, I_{97}$  and  $I_{59}$ . Thus the original predictors  $(X_4, X_2, X_3, X_1)$  enter the final model. The second application of forward selection with  $\hat{\alpha}_{\text{IND}} = .001$  gives a model with just three variables  $X_4, X_2$  and  $I_1$ . Thus the single indicator  $I_1$  enters the model. The chosen method in this case therefore selects a model consisting of the variables  $X_4, X_2, X_3, X_1$  and  $I_1$ . Note that since all three methods in this section use the same  $\hat{\alpha}_X$  estimate, they select the same set of original predictors. However with different estimates of  $\hat{\alpha}_{\text{IND}}$ , the set of indicators selected can be different.

### 2.4.2 Two-Stage Methods

Another approach to obtaining estimates  $\hat{\alpha}_{\text{IND}}$  and  $\hat{\alpha}_X$  for the indicators and original predictors is to carry out the estimation in two stages. In the first stage an estimate of  $\alpha$  for the indicators is obtained. Like the VASMI<sub>1</sub> procedure we use the fast VAMS estimates (2.19) and (2.20).

The estimated tuning parameter  $\hat{\alpha}_{\text{IND}}$  is used to perform forward selection on the augmented data  $(\mathbf{Y}, \mathbf{X}^*)$ . Although, the forward selection algorithm selects both original predictors and indicators, in the first stage we are only concerned about the set of indicators selected. Returning to the sample forward addition sequence in Table 2.1, assume that  $\hat{\alpha}_{\text{IND}}$  equals .025. Forward selection chooses a model consisting of the first 10 variables in the sequence. Under these circumstances, the variables of interest are the six indicators  $I_1, I_{98}, I_{74}, I_{10}, I_{97}$  and  $I_{59}$ . The first stage concludes with the elimination of the observations associated with the set of selected indicators. We would then delete observations 1,98,74,10,97, and 59 from the data set. In general, this leaves a regression data set with  $n - S_{\text{IND}}(\hat{\alpha}_{\text{IND}})$  observations.

With the outliers removed from the data set, a “clean” subset of the data remains. There is no longer any need to use indicator variables. The least squares methods inherent in the original VAMS procedures are now in a situation where they are known to perform well. Therefore, in the second stage either the simulation-based VAMS procedure (Section 2.1) or the fast VAMS procedures (Section 2.2) can be used on the

data that remains. The  $\hat{\alpha}_X$  estimate obtained controls the false selection rate of the original variables contained in  $\mathbf{X}$ . To distinguish between the two types of two-stage methods, we will refer to the procedure that uses simulation in the second stage as “TWO SIM.” The procedure that uses fast VAMS for the second stage will be called “TWO FAST.” Using the two stage approach gives a robust procedure that controls the false selection rate of both the original predictors and the indicators variables.

### 2.4.3 Summary

In this section five dual estimation approaches were introduced VASMI<sub>1</sub>, VAMSI<sub>2</sub>, VAMSI<sub>3</sub>, TWO SIM, and TWO FAST. Each of these procedures are designed to obtain separate tuning parameter estimates for the original predictors and indicator variables. With a common goal, it is not surprising that many of these procedures are only slight variations of one another. VASMI<sub>1</sub>, VAMSI<sub>2</sub>, and VAMSI<sub>3</sub> all use the same estimate of  $\hat{\alpha}_X$ . In terms of estimating the tuning parameter for the indicators, the TWO SIM, TWO FAST, and VASMI<sub>1</sub> procedures all use exactly the same version of the fast VAMS procedure. The VAMSI<sub>2</sub> is then a slight variation of this approach, using the fast VAMS procedure but with a modified estimate of the number of informative indicators. With the TWO SIM and TWO FAST procedures declaring the same set of observations as outliers, the difference in the estimated tuning parameters for the original predictors is directly attributable to differences in performance of the original simulation and fast VAMS procedures. Previous simulation results suggest that this difference should be minimal. In the next section, Monte Carlo simulation is used to further investigate the differences between these five variable selection procedures.

## 2.5 Comparison of Methods

### 2.5.1 Data Generation

To study the performance of the robust variable selection procedures presented thus far, we use Monte Carlo Simulation. Each of the one-hundred Monte Carlo replicates of the variable selection process requires the generation of data from the linear regression model (1.1). The data generation process presented here follows closely that of Luo et al. (2006).

Each row of the  $n \times k_T$  design matrix  $\mathbf{X}$  is generated from an AR(1) process with parameter  $\theta$ , giving the relationship

$$X[i, j] = \theta(X[i, j - 1]) + e_{ij} \quad j = 2, \dots, k_T, \quad (2.25)$$

with  $X[i, 1]$  randomly generated from a standard normal distribution. In (2.25), the  $e_{ij}$  are independent errors taken from a standard normal distribution leading to a variance-covariance structure of the form

$$V(X[i, j]) = \frac{1}{1 - \theta^2} \quad (2.26)$$

$$Cov(X[i, j], X[i, j + l]) = \frac{1}{1 - \theta^2} \theta^{|l|}. \quad (2.27)$$

Scaling each row of the design matrix by the normalizing constant

$$c_\rho = \sqrt{1 - \theta^2} \quad (2.28)$$

forces the variance of  $X[i, j]$  to be one and the covariance between  $X[i, j]$  and  $X[i, j + l]$  to be  $\theta^{|l|}$ .

In an attempt to make the design matrix have a variety of predictor variables, a proportion of the columns are transformed. The first half of the predictor variables are not transformed and keep their original values. The next one-fourth of the initial predictors are chosen to be dichotomous variables. These columns are redefined so that the  $j^{th}$  entry,  $x_j$  becomes  $x_j^*$  where

$$x_j^* = \begin{cases} 1 & \text{when } x_j \geq 0 \\ 0 & \text{when } x_j < 0. \end{cases} \quad (2.29)$$

The last fourth of the predictors are transformed via the absolute value function. Our final design matrix is obtained by centering, scaling, and randomly permuting the columns of the transformed matrix. The random permutation of the columns has two benefits. It mixes up the original and transformed variables and avoids having only adjacent variables with high correlation.

Through construction of our  $\beta$  vector, we specify the proportion of informative predictors. For all situations, one fifth of the  $k_T$  predictor variables are chosen to be informative and given initial regression coefficients equal to 1, while the remaining entries in the  $\beta$  vector are set to zero.

After the initial  $\beta$  vector is created, the coefficients of the informative predictors are scaled to control the strength of the linear relationship between  $\mathbf{Y}$  and  $\mathbf{X}$ . The variable selection procedures are more likely to identify informative variables when the relationship between  $\mathbf{Y}$  and  $\mathbf{X}$  is strong. In the context of regression models, the coefficient of determination,  $R^2$ , is commonly used to quantify the strength of the linear relationship. The formula for theoretical  $R^2$  with random predictors is

$$R^2 = \frac{\text{var}(\mathbf{X}^T \beta)}{\text{var}(\mathbf{X}^T \beta + \epsilon)}. \quad (2.30)$$

A drawback of using the standard version of  $R^2$  is that it requires the existence of moments, a feature that eliminates error distributions such as the Cauchy from consideration. Also, since the variance is a non-robust measure of variability, using (2.30) is not appropriate for data containing outlying points. Replacing the variance with the square of the median absolute deviation (MAD), results in a less restrictive, robust version of  $R^2$  defined to be

$$R_{rob}^2 = \frac{\{\text{MAD}(\mathbf{X}^T \beta)\}^2}{\{\text{MAD}(\mathbf{X}^T \beta + \epsilon)\}^2}. \quad (2.31)$$

Now the robust coefficient of determination can be used to control the strength of the relationship between  $\mathbf{Y}$  and  $\mathbf{X}$ . By implementing an iterative search, it is possible to find the scaled version of  $\beta$  that makes (2.31) obtain a target value. Note that when the errors are normally distributed, (2.31) reduces to (2.30).

Using the scaled version of  $\beta$  along with the final design matrix  $\mathbf{X}$ , we get the corresponding regression mean vector  $\mu$  using the relationship

$$\mu = E(\mathbf{Y}|\mathbf{X}) = E(\beta_0 \mathbf{1} + \mathbf{X}^T \beta + \epsilon) = \beta_0 + \mathbf{X}^T \beta. \quad (2.32)$$

with the true intercept  $\beta_0$  set equal to one. The response vector  $\mathbf{Y}$  is obtained via

$$\mathbf{Y} = \mu + \epsilon. \quad (2.33)$$

Here  $\mathbf{1}$  is a vector with each entry equal to one, that specifies the value of the regression models intercept and  $\epsilon$  is a vector containing  $n$  independent identically distributed errors. When choosing the distribution of  $\epsilon$ , one must consider the effect that the distribution has on the data and hence the variable selection procedures. For example, errors sampled from a standard normal distribution result in data relatively free of outliers. However, choosing heavier tailed distributions like the  $t$  or Cauchy changes both the frequency and magnitude of the outlying points in the data set. Creating simulation data sets using a variety of error distributions allows for insight into how outliers affect the variable selection procedures.

## 2.5.2 Simulation Design

In this section, we present the Monte Carlo Simulation design used to study the performance of the VAMS procedures developed in this chapter. These procedures include the non-robust version of the VAMS presented in Section 2.1, VAMS with indicators (VAMSI), and all the dual estimation approaches with the exception of VAMSI<sub>3</sub>. In preliminary studies the VAMSI<sub>3</sub> procedure proved to be computationally intensive and performed no better than the other dual estimation methods.

Our design uses simulated data sets generated using the method described in Section 2.5.1, with an AR(1) parameter of  $\theta=.5$  and a target  $R_{rob}^2$  of .5. The number of observations, number of candidate predictors, and the distribution of the regression errors then follow a  $2 \times 2 \times 3$  design. Data sets are generated with either  $n = 100$  or  $n = 500$  observations. The number of candidate predictors also has two levels,  $k_T = 20$

and  $k_T = 50$ . Finally, the regression errors are generated from three distributions: standard normal,  $t$  with three degrees of freedom, and Cauchy. For comparison purposes the  $t_3$  error distribution is scaled by the square root of the degrees of freedom (three) to get a variance of one.

While the Cauchy distribution does not have a finite second moment, it is possible to select a scale parameter that will make the Cauchy distribution mimic a standardized distribution in terms of the percentage of nonoutlying points. One way to do this is to match up quantiles. We wanted the Cauchy distribution to mimic the standard normal so we chose the scale parameter that made 95 percent of the distributions mass fall between  $-1.96$  and  $1.96$ . This amounts to solving

$$\{s : F_{c,s}(1.96) - F_{c,s}(-1.96) = .95\} \quad (2.34)$$

where  $F_{c,s}$  is the Cauchy distribution function with a scale parameter of  $s$ . Solving (2.34) gives a scale parameter of .154. While 95% of the mass for the scaled Cauchy distribution is between  $-1.96$  and  $1.96$  the heavy tail allows for the possibility of outlying points uncommon to the standard normal distribution. We thus have our three error distributions: standard normal, scaled  $t$  with three degrees of freedom and Cauchy with a scale parameter of .154.

The performance of the VAMS procedures under these conditions is compared in terms of the following criteria:

- Ratio of Med. ME = median model error of “optimal” model divided by the median model error of selected models
- Ratio of Mean ME = mean model error of “optimal” model divided by the mean model error of selected models
- CS = average number of informative predictors selected
- FS = average number of uninformative predictors selected
- Outliers = average number of indicator variables selected
- Size = CS+FS= average number of original predictors selected
- FSR = average false selection rate=average of FS/(1+Size)



- $\hat{\alpha}_X$  = average estimated  $\alpha$  for the original predictors
- $\hat{\alpha}_{\text{IND}}$  = average estimated  $\alpha$  for the indicator variables

For the Ratio of Median ME and Ratio of Mean ME, the model error for a single data set is defined to be

$$\text{ME} = \frac{1}{n} \sum_{i=1}^n (\hat{y}_i - \mu_i)^2 \quad (2.35)$$

where  $\hat{y}_i$  is the  $i^{\text{th}}$  predicted response and  $\mu_i$  is the  $i^{\text{th}}$  value in the mean vector  $\boldsymbol{\mu}$ . For each data set a model error is also calculated for the “optimal” model. The optimal model is one consisting of all the informative predictors and no uninformative predictors. The method of obtaining fitted values for this model depends on the distribution of the errors. For normal distributed errors, least squares regression is used. For  $t_3$  and Cauchy errors, the method of maximum likelihood is used to obtain estimated regression coefficients for each of the informative predictors.

### 2.5.3 Results

Tables 2.2-2.4 contain results for the VAMS procedures. There is a consistent difference in the number of outliers selected by the procedures across each of the simulation conditions. First consider the case when the errors are normally distributed. Each of the dual estimation approaches are eliminating significantly fewer outliers than the original VAMSI procedure. In Table 2.2, when  $n = 100$  and  $k_T = 20$ , the number of selected outliers is 1.67 for VAMSI compared to .13 or less for the dual estimation methods. A drop in the number of declared outliers is also seen when the errors come from the  $t_3$  distribution. However, when we move to Cauchy errors the dual estimation methods start to select more outliers than the VAMSI procedure. Specifically, we see an increase from 15.04 outliers for the VAMSI procedure to 18 or more for the dual estimation approaches. By tuning the selection of the indicator variables the dual estimation approaches are able to select fewer outliers than the VAMSI method when the errors are well behaved and more when the errors come

from a heavy tailed distribution.

In Table 2.2, we see that both of the model error ratios are similar across the five robust VAMS procedures. This is a result we expect, since many of these procedures are only slight variations of one another. Remember that the VAMSI, VASMI<sub>1</sub>, and VAMSI<sub>2</sub> procedures all select original predictors by the same method. In spite of this, the fact that the VASMI<sub>1</sub> and VAMSI<sub>2</sub> procedures select fewer outliers than the VAMSI procedure leads to a slight improvement in terms of model error when the errors are normally distributed. Staying with normal errors, the two-stage methods have a slight edge in terms of selection characteristics. Both the TWO FAST and TWO SIM procedures select on average more informative predictors and less uninformative than the other VAMS procedures. For  $t_3$  errors the relationship is reversed. Now the two-stage methods are doing slightly worse than the other dual estimation methods both in terms of model error and selection characteristics. With the heavier-tailed errors  $t_3$  errors, we see a clear separation between the non-robust and robust versions of the VAMS procedures. With Cauchy errors the effect is even more pronounced. The original VAMS procedure selects on average more uninformative than informative predictors. Another issue is the failure of the TWO SIM procedure for Cauchy distributed errors. Upon further inspection the procedure fails if enough outliers are removed from the data set, in the first stage of the procedure, to make the number of observations less than the number of predictors. With the number of observations less than the number of predictors generation of (2.3), the matrix of pseudo variables, is impossible due to the singularity of  $\mathbf{X}^T \mathbf{X}$ . Since the issue is with the generation of the pseudo variables, the TWO FAST procedure is unaffected.

Another important criterion is the false selection rate. Each of these methods are designed to obtain a false selection rate of .05. In Table 2.2 all five of the robust methods do an excellent job, hitting the target on average for normally distributed errors. The procedures have a little more difficulty for the heavier tailed errors but do not exceed a false selection rate of .08.

In Table 2.3 the conditions shift to  $n = 100$  and  $k_T = 50$ . These conditions are much more difficult for the variable selection procedures to handle. For normally distributed errors, the procedures are selecting on average about forty percent of the informative predictors compared to ninety-three percent under the previous set of conditions. An even smaller proportion are selected for the heavier tailed error distributions. In terms of model error, the TWO FAST procedure does worse than the other procedures. This is most evident when the errors have a Normal or  $t_3$  distribution. The procedure is fitting a significantly smaller model than the others. The TWO SIM procedure now fails for both  $t_3$  and Cauchy errors. The large number of candidate predictors makes it easier for the number of observations to fall below the number of candidate predictors. Compared to the results seen in Table 2.2, each of the dual estimation procedures are doing a poor job of keeping the false selection rate near the target rate of .05. Even for normally distributed errors the false selection rate is in the range of .12-.15.

For Table 2.4 the number of observations is raised to five hundred while  $k_T$  remains equal to fifty. With the number of observations being ten times the number of candidate predictors the variable selection procedures do a good job of distinguishing between the informative and uninformative predictors. The four dual estimation approaches have comparable model error ratios for all error distributions, with a slight edge going to the VAMSI<sub>2</sub> procedure. For Normal errors the VAMSI procedure does worse than the competition in terms of model error. Again this can be attributed to the procedure not tuning the selection of the indicator variables. With the increase in the number of observations, the false selection rates are again close to their target value of .05. Only for the Cauchy errors are they slightly inflated to around .08. The similarity of results to those seen in Table 2.2 suggests that the abnormalities exhibited under the second set of conditions were not simply due to an increase in the number of candidate predictors but the relationship between the number of candidate predictors and the number of observations. When the number of observations is large compared to the number of candidate predictors the variable selection procedures

do a good job of both separating the informative and noninformative predictors and controlling the false selection rate.

## 2.6 Breakdown Properties

In this section we develop a simulation-based procedure to investigate the finite sample breakdown point of a variable selection method. As was stated in Section 1.3.2, in practice the breakdown point is often found by observing the percentage of contaminated data  $\epsilon$ , that will send an estimator to the boundary of the parameter space (Donoho and Huber, 1983). In the context of regression, this means finding the smallest proportion of outlying points that make the estimated regression coefficients become arbitrarily large. While it may seem like a natural extension to define the breakdown point of a variable selection procedure in terms of regression coefficients, other issues arise with model selection. Unlike regression, as contamination is added to the model, the set of predictors chosen by the variable selection procedure is changing. With the set of predictors constantly changing, quantifying the effect of contamination is complex. Our method determines the breakdown point in terms of two criteria: the set of selected candidate predictors and the predicted values of the final model. Note that direct comparison of the predicted values is reasonable even with different models. Monte Carlo simulation is used to determine the effect of systematically added contamination on these two criteria.

### 2.6.1 Criteria

To determine the breakdown point in terms of the set of candidate predictors and final models predicted values, we need numerical measures of each. For the set of candidate predictors, define  $k_M$  to be the number of candidate predictors chosen by variable selection procedure  $M$ . This statistic takes values from zero (intercept model) to  $k_T$  (saturated model) and is an estimate of the number of candidate predictors that are informative. Using this statistic we have three separate definitions of breakdown that depend on the underlying “true” regression model. If the true model is not the intercept or saturated model then the breakdown point is the smallest percentage of contamination that leads to a model that contains either none or all of the candidate

predictors. That is

$$\epsilon^*(\mathbf{Y}, \mathbf{X}, k_M) = \inf\{\epsilon | k_M(\epsilon; \mathbf{Y}, \mathbf{X}) = 0 \text{ or } k_T\}. \quad (2.36)$$

where  $k_M(\epsilon; \mathbf{Y}, \mathbf{X})$  is the number of candidate predictors selected when model selection method  $M$  is applied to data with  $\epsilon$  percent contamination. If the intercept model is the true model then the breakdown point is defined to be

$$\epsilon^*(\mathbf{Y}, \mathbf{X}, k_M) = \inf\{\epsilon | k_M(\epsilon; \mathbf{Y}, \mathbf{X}) = k_T\}. \quad (2.37)$$

Conversely, if the saturated model is the true model the breakdown point is defined to be

$$\epsilon^*(\mathbf{Y}, \mathbf{X}, k_M) = \inf\{\epsilon | k_M(\epsilon; \mathbf{Y}, \mathbf{X}) = 0\}. \quad (2.38)$$

In terms of predicted values, the breakdown point is determined by monitoring the effect of contamination on the set of predicted values. Given data  $(\mathbf{Y}, \mathbf{X})$ , define  $\hat{Y}_{i,M}$  to be the fitted value obtained by applying variable selection method  $M$  to the data  $(\mathbf{Y}, \mathbf{X})$ . Given an  $\epsilon$ -percent contaminated data set  $(\mathbf{Y}^{(c)}, \mathbf{X})$  define  $\hat{Y}_{i,M}^{(c)}$  to be the fitted value obtained by applying variable selection method  $M$  to the data  $(\mathbf{Y}^{(c)}, \mathbf{X})$ . Note that the contaminated response vector  $\mathbf{Y}^{(c)}$  is obtained by modifying  $\epsilon$  percent of the original entries in  $\mathbf{Y}$  to be outliers. For details on the process of contamination see Section 2.6.2. If the variable selection method is robust to contamination, then we expect  $\hat{Y}_{i,M}$  and  $\hat{Y}_{i,M}^{(c)}$  to be close. This motivates the statistic

$$D_p = \frac{\frac{1}{n} \sum_{i=1}^n (\hat{Y}_{i,M}^{(c)} - \hat{Y}_{i,M})^2}{\frac{1}{n-p-1} \sum_{i=1}^n (Y_{i,M} - \hat{Y}_{i,M})^2}, \quad (2.39)$$

as a unitless measure of the (non)robustness of the variable selection method to contamination:  $D_p$  is small(large) if the method is robust(nonrobust). Note that the numerator of (2.39) coincides with the definition of bias (1.9) with  $T$  equal to the set of predicted values. Theoretically, the breakdown point in terms of the fitted values is then

$$\epsilon^*(\mathbf{Y}, \mathbf{X}) = \inf\{\epsilon | \sup D_p(\epsilon; \mathbf{Y}, \mathbf{X}) = \infty\}. \quad (2.40)$$

where  $D_p(\epsilon; \mathbf{Y}, \mathbf{X})$  is the value of (2.39) for a  $\epsilon$  percent contaminated data set  $(\mathbf{Y}^{(c)}, \mathbf{X})$ . As in equation (1.9) the supremum is taken over all possible  $\epsilon$ -contaminated data sets.

### 2.6.2 Simulation Design

Simulation is used to investigate the breakdown point of the VAMS procedures. Each of the VAMS methods are applied to 100 randomly generated data sets. The data are generated following the method described in Section 2.5.1 with errors taken from a standard normal distribution. Before contamination is added to the data set each variable selection method is run. The number of candidate predictors selected and the fitted values for the final model are recorded. Contamination is then systematically added to the initial data set through modification of the response vector  $\mathbf{Y}$ . First, a single observation  $Y_i$  is randomly chosen to be contaminated. After the observation is contaminated each variable selection procedure is carried out and the quantities  $k_M$  and  $D_p$  are calculated. The process of contaminating an observation and running the variable selection methods is continued until half of the observations are contaminated. At each step the percentage of contamination  $\epsilon$ , is simply the number of modified observations divided by the number of observations ( $n$ ).

The amount and pattern of contamination follow a  $3 \times 2$  design. Contamination amounts of one thousand, ten million, and one hundred billion are used to contaminate the randomly selected observations. There are also two patterns of contamination. The first involves randomly adding or subtracting the contamination quantity. The second just adds the contamination to each of the randomly chosen observations. By varying the magnitude and pattern of the contamination we can access the stability of our simulation results.

By monitoring the values of  $k_M$  and  $D_p$  a simulation-based breakdown point is obtained. For the number of variables selected, the breakdown point is the smallest proportion of contaminated points that causes either zero or  $k_T$  candidate predictors to be selected. Remember, for our simulated data sets we use definition (2.36) since exactly one-fifth of the candidate predictors have nonzero regression coefficients. The

breakdown point in terms of the predicted values is found by observing the minimal amount of contamination that causes  $D_p$  to be arbitrarily large.

### 2.6.3 Results

The simulation-based estimate of the breakdown point for the VAMS procedures is .10. The result is consistent for all magnitudes and patterns of contamination. There is also agreement when the breakdown point is calculated in terms of the two different diagnostics  $k_M$  and  $D_p$ . These results coincide with those in Hampel (1985) where it was found that forward-addition, outlier-rejection methods had a breakdown point of .10. In Hampel's study the breakdown point was calculated for a fixed regression model. The carryover of the results gives some evidence that the breakdown point of a variable selection procedure is equal to the breakdown point of the regression estimation method driving the procedure.

## 2.7 Conclusions

From the simulation results, the VASMI<sub>1</sub> and VAMSI<sub>2</sub> methods have the best performance. The ability of each procedure to tune the selection of indicators and handle data set like the one presented in Figure 2.1 is a big advantage over the VAMSI procedure. There is also the issue that the TWO SIM procedure is unable to be used if the number of outliers eliminated in the first stage of the procedure causes the number of candidate predictors to exceed the number of observations. As the number of candidate predictors approaches the number of observations this is much more likely to be an issue. While the TWO FAST procedure is able to produce results as  $k_T$  approaches  $n$ , they are not consistent with the output from the other dual estimation approaches. In terms of breakdown, the VASMI<sub>1</sub> and VAMSI<sub>2</sub> methods are equivalent with a breakdown point of .10. While breakdown point does not help us compare between these two methods it will be useful when comparing to methods presented in the next two chapters.



Table 2.2: Simulations Results for  $n = 100$ ,  $k_T = 20$ , and  $k_I = 4$ 

Normal Errors								
Method	Ratio of Med. ME	Ratio of Mean ME	CS	FS	Outliers	Size	FSR	$\hat{\alpha}_X$
VAMS	0.74	0.50	3.73	0.34	0	4.07	0.06	0.02
VAMSI	0.65	0.43	3.68	0.33	1.67	4.01	0.05	0.01
VASMI <sub>1</sub>	0.72	0.47	3.68	0.33	0.13	4.01	0.05	0.01
VAMSI <sub>2</sub>	0.72	0.47	3.68	0.33	0.09	4.01	0.05	0.01
TWO SIM	0.72	0.51	3.74	0.32	0.13	4.06	0.05	0.02
TWO FAST	0.72	0.48	3.70	0.31	0.13	4.01	0.05	0.02
Max SE	0.08	0.04	0.06	0.07			0.01	
$t_3$ Errors								
Method	Ratio of Med. ME	Ratio of Mean ME	CS	FS	Outliers	Size	FSR	$\hat{\alpha}_X$
VAMS	0.17	0.18	3.08	0.47	0	4.55	0.09	0.02
VAMSI	0.52	0.33	3.61	0.45	5.49	4.06	0.08	0.01
VASMI <sub>1</sub>	0.47	0.33	3.61	0.45	2.46	4.06	0.08	0.01
VAMSI <sub>2</sub>	0.50	0.33	3.61	0.45	2.14	4.06	0.08	0.01
TWO SIM	0.38	0.29	3.52	0.48	2.46	4.00	0.08	0.02
TWO FAST	0.39	0.29	3.51	0.41	2.49	3.92	0.07	0.02
Max SE	0.08	0.03	0.08	0.07			0.01	
Cauchy								
Method	Ratio of Med. ME	Ratio of Mean ME	CS	FS	Outliers	Size	FSR	$\hat{\alpha}_X$
VAMS	0.01	0.00	0.34	0.87	0	1.21	0.41	0.00
VAMSI	0.43	0.25	3.78	0.45	15.04	4.23	0.07	0.01
VASMI <sub>1</sub>	0.46	0.26	3.78	0.45	18.61	4.23	0.07	0.01
VAMSI <sub>2</sub>	0.49	0.27	3.78	0.45	19.04	4.23	0.07	
TWO SIM	—	—	—	—	—	—	—	—
TWO FAST	0.42	0.20	3.70	0.49	18.61	4.19	0.08	0.02
Max SE	0.06	0.04	0.08	0.08			0.01	

CS=average # correctly selected, FS=average # falsely selected, Size=CS+FS

 $\hat{\alpha}_X$ =average estimated  $\alpha$  to enter for forward selection

Note:Entries are based on 100 replicates

Max SE=maximum standard error across the methods excluding non robust VAMS

Table 2.3: Simulations Results for  $n = 100$ ,  $k_T = 50$ , and  $k_I = 10$ 

Normal Errors								
Method	Ratio of Med. ME	Ratio of Mean ME	CS	FS	Outliers	Size	FSR	$\hat{\alpha}_X$
VAMS	0.22	0.25	4.51	1.10	0	5.61	0.15	0.01
VAMSI	0.21	0.23	4.14	0.96	1.41	5.10	0.14	0.01
VASMI <sub>1</sub>	0.21	0.24	4.14	0.96	0.16	5.10	0.14	0.01
VAMSI <sub>2</sub>	0.21	0.24	4.14	0.96	0.04	5.10	0.14	0.01
TWO SIM	0.22	0.24	4.41	1.11	0.16	5.52	0.15	0.01
TWO FAST	0.19	0.22	3.59	0.72	0.16	4.47	0.12	0.01
Max SE	0.02	0.01	0.19	0.12			0.02	
$t_3$ Errors								
Method	Ratio of Med. ME	Ratio of Mean ME	CS	FS	Outliers	Size	FSR	$\hat{\alpha}_X$
VAMS	0.16	0.17	2.84	0.69	0	3.53	0.14	0.01
VAMSI	0.20	0.21	3.69	0.71	3.42	4.40	0.12	0.01
VASMI <sub>1</sub>	0.20	0.21	3.69	0.71	2.08	4.40	0.12	0.01
VAMSI <sub>2</sub>	0.20	0.21	3.69	0.71	1.61	4.40	0.12	0.01
TWO SIM	—	—	—	—	—	—	—	—
TWO FAST	0.18	0.18	2.81	0.52	2.08	3.33	0.10	0.01
Max SE	0.01	0.01	0.16	0.09			0.02	
Cauchy								
Method	Ratio of Med. ME	Ratio of Mean ME	CS	FS	Outliers	Size	FSR	$\hat{\alpha}_X$
VAMS	0.04	0.00	0.38	0.89	0	1.27	0.40	0.00
VAMSI	0.10	0.11	3.69	0.95	9.43	4.64	0.15	0.01
VASMI <sub>1</sub>	0.10	0.11	3.69	0.95	21.1	4.61	0.15	0.01
VAMSI <sub>2</sub>	0.10	0.11	3.69	0.95	11.41	4.61	0.15	
TWO SIM	—	—	—	—	—	—	—	—
TWO FAST	0.10	0.11	3.74	1.42	21.1	5.16	0.16	0.01
Max SE	0.01	0.01	0.25	0.25			0.02	

CS=average # correctly selected, FS=average # falsely selected, Size=CS+FS

 $\hat{\alpha}_X$ =average estimated  $\alpha$  to enter for forward selection

Note:Entries are based on 100 replicates

Max SE=maximum standard error across the methods excluding non robust VAMS

Table 2.4: Simulations Results for  $n = 500$ ,  $k_T = 50$ , and  $k_I = 10$ 

Normal Errors								
Method	Ratio of Med. ME	Ratio of Mean ME	CS	FS	Outliers	Size	FSR	$\hat{\alpha}_X$
VAMS	0.71	0.69	9.99	0.69	0	10.68	0.05	0.01
VAMSI	0.64	0.62	9.98	0.61	8.24	10.59	0.05	0.01
VASMI <sub>1</sub>	0.71	0.72	9.98	0.61	0.03	10.59	0.05	0.01
VAMSI <sub>2</sub>	0.71	0.72	9.98	0.61	0.03	10.59	0.05	0.01
TWO SIM	0.70	0.69	9.99	0.68	0.03	10.67	0.05	0.01
TWO FAST	0.71	0.70	9.99	0.64	0.04	10.63	0.05	0.01
Max SE	0.06	0.03	0.01	0.08			0.01	
$t_3$ Errors								
Method	Ratio of Med. ME	Ratio of Mean ME	CS	FS	Outliers	Size	FSR	$\hat{\alpha}_X$
VAMS	0.34	0.28	9.55	0.66	0	10.21	0.05	0.01
VAMSI	0.61	0.57	9.94	0.63	30.36	10.47	0.05	0.01
VASMI <sub>1</sub>	0.59	0.58	9.94	0.63	14.28	10.57	0.05	0.01
VAMSI <sub>2</sub>	0.62	0.57	9.94	0.63	10.84	10.57	0.05	0.01
TWO SIM	0.58	0.56	9.95	0.64	14.28	10.59	0.05	0.01
TWO FAST	0.59	0.57	9.95	0.59	14.47	10.54	0.05	0.01
Max SE	0.04	0.03	0.03	0.09			0.01	
Cauchy								
Method	Ratio of Med. ME	Ratio of Mean ME	CS	FS	Outliers	Size	FSR	$\hat{\alpha}_X$
VAMS	0.01	0.00	0.32	0.88	0	1.20	0.42	0.00
VAMSI	0.43	0.42	9.99	0.99	81.22	10.98	0.08	0.01
VASMI <sub>1</sub>	0.46	0.46	9.99	0.99	100.14	10.98	0.08	0.01
VAMSI <sub>2</sub>	0.49	0.50	9.99	0.99	111.73	10.98	0.08	0.01
TWO SIM	0.49	0.43	10.00	1.09	100.14	11.09	0.08	0.02
TWO FAST	0.42	0.43	10.00	1.07	100.28	11.07	0.08	0.02
Max SE	0.03	0.02	0.01	0.11			0.01	

CS=average # correctly selected, FS=average # falsely selected, Size=CS+FS

$\hat{\alpha}_X$ =average estimated  $\alpha$  to enter for forward selection

Note:Entries are based on 100 replicates

Max SE=maximum standard error across the methods excluding non robust VAMS

## Chapter 3

# Variable Selection using Robust Regression

In this chapter we develop methods for robust variable selection based on robust M-estimation regression procedures. First we develop a forward-addition sequence based on M-estimation score tests and then select a model using a robust Bayesian information criterion (RBIC). In addition we develop an analogue to the least squares VAMS procedures described in Chapter 2.

### 3.1 Huber-Based Approach

Section 1.3.1 presented solutions to the M-estimating equations (1.6) that depend on the choice of  $\rho(\cdot)$ ,  $\psi(\cdot)$ , and the method used to obtain a standardizing estimate of scale. Here we provide details for the classical approach of Huber (1964, 1981), in which  $\psi(\cdot)$  is given by

$$\psi_H(x) = \begin{cases} x & \text{when } |x| \leq k_H \\ k & \text{when } |x| > k_H, \end{cases} \quad (3.1)$$

where  $k_H$  is a constant, controlled by the researcher, that affects the robustness of the final M-estimates. In this thesis we have chosen to use  $k_H = 1$ .

Huber (1964) outlines multiple approaches for obtaining an estimate of scale. Huber's Proposal 2 method entails estimating  $\sigma$  by jointly solving the M-estimating equations

$$\sum_{i=1}^n \psi \left[ \frac{Y_i - (\beta_0 + \mathbf{x}_i^T \boldsymbol{\beta})}{\sigma} \right] \begin{pmatrix} 1 \\ \mathbf{x}_i \end{pmatrix} = \mathbf{0}, \quad (3.2)$$

with

$$\frac{1}{n - k_T} \sum_{i=1}^n \psi^2 \left( \frac{r_i}{\sigma} \right) = E_{\Phi}(\psi^2) = \int_{-\infty}^{\infty} \psi^2(x) \phi(x) dx, \quad (3.3)$$

where  $\Phi$  and  $\phi$  are the standard normal distribution and density function respectively. Solving this system of equations is often accomplished by iteratively re-weighted least squares (Holland and Welsch, 1977; Street et al., 1988). The algorithm starts with initial estimates of  $\hat{\sigma}^{(0)}$ ,  $\hat{\beta}_0^{(0)}$  and  $\hat{\boldsymbol{\beta}}^{(0)}$ , calculates the residuals

$$r_i^{(0)} = \frac{Y_i - (\hat{\beta}_0^{(0)} + \mathbf{x}_i^T \hat{\boldsymbol{\beta}}^{(0)})}{\hat{\sigma}^{(0)}} \quad (3.4)$$

and then the weights

$$w_i^{(0)} = \frac{\psi \left( r_i^{(0)} \right)}{\left( r_i^{(0)} \right)}. \quad (3.5)$$

Updated estimates  $(\hat{\beta}_0^{(1)}, \hat{\boldsymbol{\beta}}^{(1)})$ , are found through minimization of the function

$$\sum_{i=1}^n w_i^{(0)} \{Y_i - (\beta_0 + \mathbf{x}_i^T \boldsymbol{\beta})\}^2, \quad (3.6)$$

and then an updated  $\hat{\sigma}^{(1)}$  is

$$\hat{\sigma}^{(1)} = \left[ \frac{1}{n - k_T} \sum_{i=1}^N \frac{w_i^{(0)} \{Y_i - (\beta_0^{(1)} + \mathbf{x}_i^T \hat{\boldsymbol{\beta}}^{(1)})\}^2}{E_{\Phi}(\psi^2)} \right]^{\frac{1}{2}}. \quad (3.7)$$

Note that  $w_i^{(0)}$  in (3.7) is found by using the initial estimate of  $\sigma$  and the updated version of  $(\beta_0, \boldsymbol{\beta})$  from (3.6). Starting with estimates  $\hat{\sigma}^{(1)}$ ,  $\hat{\beta}_0^{(1)}$  and  $\hat{\boldsymbol{\beta}}^{(1)}$  the process is repeated to get second level estimates  $\hat{\sigma}^{(2)}$ ,  $\hat{\beta}_0^{(2)}$  and  $\hat{\boldsymbol{\beta}}^{(2)}$ . The process is repeated until the regression coefficients  $(\hat{\beta}_0^{(i)}, \hat{\boldsymbol{\beta}}^{(i)})$  stabilize at the final robust regression estimates  $(\hat{\beta}_0, \hat{\boldsymbol{\beta}})$ . The final weights represent the amount of downweighting for each

observation. A weight of zero is equivalent to completely removing an observation during calculation of the regression estimates. Thus, downweighting extreme points is an alternative to removing them from the data set.

Following the process just outlined, we have a method for obtaining regression coefficients for candidate models. Using the classical Huber  $\psi$  function allows for estimates more robust to outlying points than those obtained from classical least squares methods.

### 3.1.1 Score Forward Addition Sequence from M-estimation

In this section, we develop a method to identify the set of candidate models for robust regression. One possibility is to consider all possible subsets of the potential predictor variables. As stated in Section 1.2.1 the computation required for such an approach is unreasonable for large  $k_T$ . An alternative is to apply a robust forward selection algorithm to the set of predictors, and thereby construct a robust forward addition sequence (RFAS) that specifies only  $k_T$  candidate models.

We consider a modified forward addition sequence based on the robust generalized score statistic described in Boos (1992). Let  $\boldsymbol{\beta}^T = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T)$  be a partition of the vector containing regression coefficients. There is a corresponding partition of the design matrix  $\mathbf{X} = \mathbf{X}_1 : \mathbf{X}_2$  where  $\mathbf{X}_1$  is the matrix of predictor variables identified with  $\boldsymbol{\beta}_1$ , and  $\mathbf{X}_2$  is the matrix of predictor variable associated with  $\boldsymbol{\beta}_2$ . The generalized score statistic tests the hypothesis

$$H_0 : \boldsymbol{\beta}_2 = \mathbf{0} \quad \text{vs} \quad H_a : \boldsymbol{\beta}_2 \neq \mathbf{0}, \quad (3.8)$$

and is developed using M-estimation techniques. Restricted estimates  $\tilde{\beta}_0$  and  $\tilde{\boldsymbol{\beta}}$  are calculated assuming the null hypothesis in (3.8) is true, and therefore are the solutions to

$$\sum_{i=1}^n \psi \left\{ \frac{Y_i - (\beta_0 + \mathbf{x}_{1i}^T \boldsymbol{\beta})}{\tilde{\sigma}} \right\} \begin{pmatrix} 1 \\ \mathbf{x}_{1i} \end{pmatrix} = 0 \quad (3.9)$$

where  $\mathbf{x}_{1i}^T$  is the  $i^{th}$  row of  $\mathbf{X}_1$  and  $\tilde{\sigma}$  is an estimate of scale calculated under the

the null hypothesis. Note that (3.9) is the M-estimating equations for the robust regression of  $\mathbf{Y}$  on  $\mathbf{X}_1$ .

Taking the derivative of

$$f(\boldsymbol{\beta}; Y_1, \dots, Y_n) = \sum_{i=1}^n \rho \left\{ \frac{Y_i - (\beta_0 + \mathbf{x}_i^T \boldsymbol{\beta})}{\sigma} \right\}, \quad (3.10)$$

with respect to  $\boldsymbol{\beta}_2$  and plugging in the restricted estimates from (3.9), we obtain the score function

$$S_2(\tilde{\boldsymbol{\beta}}) = \sum_{i=1}^n \psi \left\{ \frac{Y_i - (\tilde{\beta}_0 + \mathbf{x}_{1i}^T \tilde{\boldsymbol{\beta}})}{\tilde{\sigma}} \right\} \mathbf{x}_{2i}, \quad (3.11)$$

where  $\mathbf{x}_{2i}^T$  is the  $i^{th}$  row of the  $\mathbf{X}_2$  matrix. The asymptotic covariance matrix for (3.11) as presented in Boos (1992), is

$$V_{S_2} = \psi_{G,\sigma}^2 [\mathbf{X}_2^T \mathbf{X}_2 - (\mathbf{X}_2^T \mathbf{X}_1)(\mathbf{X}_1^T \mathbf{X}_1)^{-1}(\mathbf{X}_1^T \mathbf{X}_2)], \quad (3.12)$$

with  $\psi_{G,\sigma}^2$  estimated by

$$\tilde{\psi}_{G,\sigma}^2 = \frac{1}{n-q} \sum_{i=1}^n \psi^2 \left\{ \frac{Y_i - (\tilde{\beta}_0 + \mathbf{x}_{1i}^T \tilde{\boldsymbol{\beta}})}{\tilde{\sigma}} \right\}. \quad (3.13)$$

In (3.13)  $q = \text{rank}(\mathbf{X}_1) + 1$ . The addition of one takes into account that we are fitting a model with an intercept. The generalized score statistic that tests (3.8) is then

$$T_{GS} = S_2(\tilde{\boldsymbol{\beta}})^T \tilde{\mathbf{V}}_{S_2}^{-1} S_2(\tilde{\boldsymbol{\beta}}). \quad (3.14)$$

Through proper specification of  $\boldsymbol{\beta}_1$ ,  $\boldsymbol{\beta}_2$ , and  $\psi(\cdot)$ , we can develop a score forward-addition sequence based on (3.14). As in the classical least squares case, the sequence is constructed by sequentially testing the effect of adding a single predictor to the model. Therefore for all tests  $\beta_2$  is a scalar, equal to the regression coefficient of the single variable being tested for entry into the model.

Starting from a model containing just an intercept,  $\beta_2$  iterates through the regression coefficients of the  $k_T$  potential predictors. The  $k_T$  resulting score statistics measure the strength of the relationship between each predictor and the response

variable. The variable with the largest test statistic is entered into the model and becomes the first variable in the robust score forward addition sequence. At the next step, the  $\beta_1$  vector is updated to include the regression coefficient of the variable that entered the model in the first step. Similarly,  $\beta_2$  then iterates through the coefficients of the remaining predictors to find the second variable chosen to enter the model. The process continues until every variable has entered the model. The resulting score forward-addition sequence specifies the  $k_T$  candidate models from which the final model is chosen.

The level of robustness of the resulting forward addition sequence depends on the choice of the  $\psi$  function. By using  $\psi_H(x)$  we can obtain a Huber-based forward addition sequence that corresponds to the M-estimation approach in the previous section. Note that the usual least squares forward addition sequence is obtained when  $\psi(x) = x$ .

### 3.1.2 Robust Bayesian Information Criterion

The Bayesian information criterion (BIC) is often used to select between a number of models of differing dimension. The basic formula (Schwarz, 1978) is

$$BIC = -2 \log L(\hat{\boldsymbol{\theta}}, \mathbf{Y}) + p \log(n), \quad (3.15)$$

where  $\log L(\hat{\boldsymbol{\theta}}, \mathbf{Y})$  is the log likelihood of the response evaluated at the maximum likelihood estimate of  $\boldsymbol{\theta}$ . The second term,  $p \log(n)$ , is a measure of model complexity, where  $p$  represents the number of predictors in the current model. Note that the penalty  $p \log(n)$  increases with the number of parameters in the model. As defined in (3.15) preferred models have a smaller value of BIC.

Developing (3.15) in terms of the normal likelihood and dropping constants leads to

$$BIC = n \log \left[ \frac{\sum_{i=1}^n \{Y_i - (\hat{\beta}_0 + \mathbf{x}_i^T \hat{\boldsymbol{\beta}})\}^2}{n} \right] + p \log(n), \quad (3.16)$$

for the Gaussian linear model. Since (3.16) is based on the normal errors model, choosing the model with minimum BIC can perform poorly when the regression errors



stray from normality (Wang et al., 2007). Therefore, it is advantageous to develop a robust version of the BIC that adjusts for heavier tailed error distributions. Consider (3.15) in terms of the Huber density (Huber, 1981)

$$f(x) = \frac{1 - \epsilon}{\sqrt{2\pi}\sigma} e^{-\rho_H(x)}, \quad (3.17)$$

where  $\rho_H$  is the Huber  $\rho$  function defined to be

$$\rho_H(x) = \begin{cases} \frac{1}{2}x^2 & \text{when } |x| < k_H \\ k|x| - \frac{1}{2}k^2 & \text{when } |x| \geq k_H \end{cases} \quad (3.18)$$

and  $\epsilon$  satisfies

$$\frac{2\phi(k_H)}{k_H} - 2\Phi(-k_H) = \frac{\epsilon}{1 - \epsilon}. \quad (3.19)$$

where  $\phi$  and  $\Phi$  are the normal density and cumulative density function, respectively. Note that relationship (3.19) between  $k_H$  and  $\epsilon$  ensures (3.17) is a proper probability density function. Constructing the log likelihood function in terms of the Huber density gives

$$\begin{aligned} \log L_H(\beta_0, \boldsymbol{\beta}, \sigma, \mathbf{Y}) &= -n \log \sigma + n \log(1 - \epsilon) - \frac{1}{2}n \log(2\pi) \\ &\quad - \sum_{i=1}^n \rho_H \left\{ \frac{Y_i - (\beta_0 + \mathbf{x}_i^T \boldsymbol{\beta})}{\sigma} \right\}. \end{aligned} \quad (3.20)$$

The value of the log likelihood is then calculated using estimates of  $\beta_0$ ,  $\boldsymbol{\beta}$ , and  $\sigma$ . When the errors are normally distributed, as in the case of (3.16), the least squares estimates  $\hat{\beta}_0$ ,  $\hat{\boldsymbol{\beta}}$ , and  $\hat{\sigma}$  are used. For the robust version of BIC, we use the estimates  $\hat{\beta}_0$ ,  $\hat{\boldsymbol{\beta}}$ , and  $\hat{\sigma}$  obtained by the robust regression procedure outlined in Section 3.1. Note that this is not actually a true BIC because  $\hat{\sigma}$  is not the mle. Thus, the RBIC is defined as

$$\text{RBIC}_H = -2 \log L_H(\hat{\beta}_0, \hat{\boldsymbol{\beta}}, \hat{\sigma}, \mathbf{Y}) + p \log n. \quad (3.21)$$

$\text{RBIC}_H$  is an appropriate version of BIC for use when data contain outlying points. As with the original version of BIC, when distinguishing between a group of candidate models it is appropriate to select the model with the minimum value of the criterion.

## 3.2 M-estimation using $t_3$

The robust regression procedure defined in Section 3.1 uses M-estimation coupled with the  $\rho$  and  $\psi$  functions associated with the Huber density (3.17). The Huber  $\rho$  and  $\psi$  functions also play pivotal roles in the construction of the score forward addition sequence and the calculation of RBIC, the two methods that enable robust regression to have variable selection capabilities. The dependence of each of these methods on the starting density suggests that the performance of the variable selection procedure can change with the selection of a new starting density.

Preliminary simulation results indicated that the Huber-based approach is vulnerable to outlying points with particularly large residuals. We would like to choose a starting density that leads to a procedure that can handle these high-magnitude outliers. Examination of the Huber density can guide our selection of a new density. The Huber density is a mixed distribution with a Normal center and exponential tail. Having a Normal center allows for efficient results when most or all of the data meet the assumptions of a Gaussian linear model. The wider than normal exponential tail enables robust estimation of the regression coefficient when there are observations that deviate from this model.

Consider redefining the procedure in terms of the  $t$  distribution (Lange et al., 1989),

$$f_{t_v}(t) = \frac{\Gamma(\frac{v+1}{2})}{\sqrt{v\pi}\Gamma(\frac{v}{2})} \left(1 + \frac{t^2}{v}\right)^{-\frac{v+1}{2}} \quad (3.22)$$

where  $v$  is the degrees of freedom and  $\Gamma(\cdot)$  is the function  $\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t} dt$ . A benefit of using (3.22) is that specification of the degrees of freedom directly controls the width of the tails. The  $t$  distribution also shares the favorable properties of the Huber density, a symmetric center and heavy tails.

Using (3.22) we can obtain associated  $\rho$  and  $\psi$  functions necessary for M-estimation. Remember that M-estimates of the regression parameters are found by solving

$$(\hat{\beta}_0, \hat{\boldsymbol{\beta}}, \hat{\sigma}) = \arg \min \sum_{i=1}^n \rho \left\{ \frac{Y_i - (\beta_0 + \mathbf{x}_i^T \boldsymbol{\beta})}{\sigma} \right\}. \quad (3.23)$$

We take  $\rho_t(t) = -\log f_{t_v}(t)$  to get

$$\rho_t(t) = \log \left\{ \frac{\Gamma(\frac{v+1}{2})}{\sqrt{v\pi}\Gamma(\frac{v}{2})} \right\} - \frac{v+1}{2} \log\left(1 + \frac{t^2}{v}\right). \quad (3.24)$$

The M-estimates obtained using (3.24) are those that minimize the negative of the log likelihood, or equivalently maximize the likelihood. Note that in contrast to the Huber-based approach, combining (3.24) with (3.23) results in maximum likelihood estimate not just of the regression coefficients but of scale. The  $\psi$  function corresponding to (3.24) is proportional to  $\rho'(t)$  and is given by

$$\psi_t(t) = \frac{t}{t^2 + v} \quad (3.25)$$

To this point the results have been derived for a non-specific  $v$  degrees of freedom. For the purposes of this paper a  $t$  distribution with 3 degrees of freedom is used. This allows for the existence of both first and second moments but gives a much heavier tail than a  $t$  distribution with a larger number of degrees of freedom. With a heavier tail than even the Huber density, we have a procedure better equipped to handle large outliers. Note that the existence of moments is not required for this procedure, and there has been some evidence that using a  $t$  distribution with one or two degrees of freedom will also give favorable results (He et al., 2000).

### 3.2.1 Comparison of Huber and $t_3$

Examination of the Huber and  $t_3$   $\psi$  functions gives insight into the expected difference in performance for the two variable selection procedures. Graphs of the  $\psi$  functions are shown in Figure (3.1). We see in Figure 3.1 that the Huber  $\psi$  function is a monotone function that levels off when standardized residuals are larger in absolute value than  $k_H$ . An advantage of the Huber  $\psi$  function is that the monotonicity guarantees a unique solution to the M-estimating equations (1.6). We also see in Figure 3.1 that the  $t_3$   $\psi$  function is not monotone, with values approaching zero for inputs large in absolute value. While a unique solution is not guaranteed, M-estimates based on non-monotone or “redescending”  $\psi$  functions are suggested for

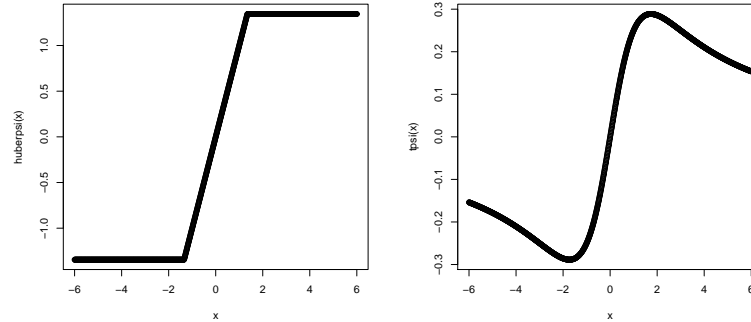


Figure 3.1:  $\psi$  functions. Left panel: Huber  $\psi$  function,  $k=1$ ; Right Panel:  $t_3$   $\psi$  function,  $v=3$ .

use with errors from a symmetric heavy-tailed distribution. The estimates obtained using these  $\psi$  functions are more robust to large outliers (Maronna, 2006, p. 30).

The easiest way to see the advantage of using a redescending  $\psi$  function like the  $t_3$ , in terms of outliers, is through examination of the M-estimation weight function. We saw in Section 3.1 that the weight of an observation is determined by the function

$$W(x) = \begin{cases} \frac{\psi(x)}{x} & \text{if } x \neq 0 \\ \psi'(0) & \text{when } x = 0. \end{cases} \quad (3.26)$$

When using a Huber  $\psi$  function, (3.26) leads to weights that are inversely proportional to the magnitude of the extreme value. In contrast, for the  $t_3$  method the weights  $1/(x^2+3)$ , are approximately inversely proportional to the square of the magnitude of the outlier. As a result, when using the  $t_3$   $\psi$  function, the weights of outlying points approach zero more quickly and thus have less of an effect on estimates. Figure 3.2 displays the weight functions for both the Huber and  $t_3$  methods, standardized so that the maximum weight is equal to one. As stated earlier the drawback of using a redescending  $\psi$  function is that a unique solution to the M-estimating equations (1.6) is not guaranteed. One can avoid inappropriate solutions by providing the iterative procedure used to solve the M-estimating equations with a good starting point. A number of methods are available for obtaining these initial estimates. One possibility

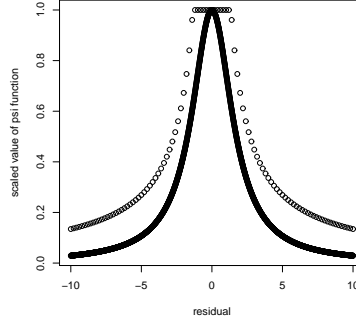


Figure 3.2: Huber and  $t_3$  M-estimation Weight functions. Huber: dotted line;  $t_3$ : solid line.

is to use the M-estimates obtained when using the Huber  $\psi$  function. Since the Huber  $\psi$  function is monotone, a unique solution is not only guaranteed but often a good starting point (Maronna, 2006, p. 99). Other options include robust, high-breakdown, regression methods such as the least median of squares (LMS) or the least trimmed squares (LTS) presented in Section 1.3.2. Depending on the structure of the data, one method may provide more stable starting values than another. The failure of the Huber method in the presence of large outliers suggests that the high breakdown regression methods are a better approach for obtaining starting values.

### 3.2.2 Modified RBIC

The next step in obtaining a modified robust regression variable selection procedure is incorporating the  $t_3$  distribution into our definition of RBIC. As in Section 3.1.2, the robust version of BIC is developed from the basic BIC formula

$$BIC = -2 \log L(\hat{\boldsymbol{\theta}}, \mathbf{Y}) + p \log(n). \quad (3.27)$$

For comparison purposes, rather than constructing the likelihood in terms of the most basic  $t_3$  probability density function, we first scale the distribution to have a variance of one. The standardization is accomplished by dividing by the square root of the

degrees of freedom. That is, we are interested in the distribution of

$$t_* = \frac{t}{\sqrt{3}} \quad (3.28)$$

where  $t$  has density (3.22) with  $v = 3$ . Using transformation (3.28), the scaled  $t_3$  pdf is

$$f_*(t) = \sqrt{3}f_{t_3}(\sqrt{3}t). \quad (3.29)$$

Now using (3.29) we construct our log likelihood

$$\log L_t(\beta_0, \boldsymbol{\beta}, \sigma, \mathbf{Y}) = n \log \left( \frac{\sqrt{3}}{\sigma} \right) + \sum_{i=1}^n \log \left[ f_{t_3} \left( \frac{\sqrt{3}}{\sigma} \{Y_i - (\beta_0 + \mathbf{x}_i^T \boldsymbol{\beta})\} \right) \right]. \quad (3.30)$$

The second level of scaling by  $\sigma$  is required to ensure that results are invariant to changes in scale of the dependent variable.

To obtain an RBIC value for a particular data set, robust estimates of  $\beta_0$ ,  $\boldsymbol{\beta}$ , and  $\sigma$  are needed to plug into (3.30). There are several valid methods for obtaining estimates. One possibility is to follow the method outlined in Section 3.1 and use re-weighted least squares with a Huber Proposal 2 estimate of scale. A  $t_3$  version of the procedure can be obtained by replacing the Huber  $\rho$  and  $\psi$  functions with their  $t_3$  counterparts. Another option is to obtain estimates through direct maximization of (3.30). While the maximization problem is computationally difficult, software packages such as R make obtaining a solution simple. For the purposes of this paper the `nlm` function in R is used to obtain the minimum. The algorithm is a Newton-type algorithm outlined by Dennis and Schnable (1983).

The `nlm` function requires starting values for both the regression coefficients and scale. We use the least median of squares (LMS) procedure to obtain initial estimates of the regression coefficients. The procedure is computationally simple and robust. As discussed in Section 1.3.2, the low efficiency of the resulting estimates is not of great concern since the estimates are only being used as a starting point for an optimization problem. After running the LMS procedure, the starting value for the scale estimate is calculated using the normalized MAD,

$$\text{MADN} = \frac{1}{.675} \text{Med}(|\mathbf{r} - \text{Med}(\mathbf{r})|) \quad (3.31)$$

where  $\mathbf{r}$  is a  $n \times 1$  vector of residuals from the initial least-median-of-squares fit. Dividing by the normalizing constant of .675 ensures the MAD approximates the standard deviation when the errors are normally distributed (Maronna, 2006, p. 33). Much like the LMS procedure, the normalized MAD has good robustness properties.

Plugging (3.30) and the MLE estimates obtained from the `nlm` function into (3.27), gives the modified version of the RBIC,

$$\text{RBIC}_{t_3} = -2 \log L_t(\hat{\boldsymbol{\beta}}, \hat{\sigma}, \mathbf{Y}) + p \log(n). \quad (3.32)$$

The  $\text{RBIC}_{t_3}$  criterion defined here is used in precisely the same way as the Huber version (3.21). For a group of candidate models, the model with the smallest RBIC is deemed superior.

### 3.2.3 $t_3$ Score Forward Addition Sequence

The final step is to generate a score forward addition sequence based on the  $t_3$   $\psi$  function. We saw in Section 3.1.1 that the generalized score statistic used to create the forward sequence has the form

$$T_{GS} = S_2(\tilde{\boldsymbol{\beta}})^T \tilde{\mathbf{V}}_{S_2}^{-1} S_2(\tilde{\boldsymbol{\beta}}), \quad (3.33)$$

where both  $S_2(\tilde{\boldsymbol{\beta}})$  and  $\tilde{\mathbf{V}}_{S_2}^{-1}$  are calculated using the Huber  $\psi$  function (3.1). Replacing the Huber  $\psi$  function with (3.25) gives a generalized score statistic based on the  $t_3$   $\psi$  function. The estimates  $\tilde{\beta}_0$ ,  $\tilde{\boldsymbol{\beta}}$ , and  $\tilde{\sigma}$  in (3.11) and (3.13) are still restricted estimates calculated assuming the null hypothesis  $\beta_2 = 0$  is true. As with the  $t_3$  version of RBIC, the regression slope and scale estimates are determined by direct maximization of the scaled  $t_3$  likelihood. We again use the LMS procedure and the normalized MAD as starting values.

With a robust forward addition sequence of models and an RBIC criterion based on the  $t$  distribution to choose a model, we have a robust regression variable selection procedure that is more resistant to large outlying observations. In a later section, simulation is used to see how much an advantage the  $t$  version of the procedure has over the Huber.

### 3.3 VAMS using Score FAS

The VAMS procedures in Chapter 2 are built upon least squares methods. The sensitivity of the least squares methods to outlying points led to the robust VAMSI and dual estimation methods. By appending a matrix of indicators to the original design matrix, these procedures had the ability to remove the outlying points that would make the use of least squares methods inappropriate.

An alternative VAMS type approach is to use the robust forward addition sequence based on the generalized score statistic (3.14) and the fast VAMS method for estimating  $\alpha$ . The distribution of (3.14) has been irrelevant to this point, as the forward addition sequence has been used simply as a way of obtaining candidate models for the RBIC. However, if we are to combine a robust forward addition sequence with the fast VAMS procedure, we need sequential  $p$ -values for the variables as they enter the model. For a general  $\psi$  and  $\rho$  function the generalized score statistic (3.14) has an asymptotic  $\chi^2$  distribution with one degree of freedom under the null hypothesis. The sequential  $p$ -value of the  $i^{th}$  variable in a robust forward addition sequence can therefore be obtained using

$$p_i = P(Z^2 > TGS_i), \quad (3.34)$$

where  $TGS_i$  is the generalized score statistic for the  $i^{th}$  variable in the forward addition sequence and  $Z$  is a standard normal random variable.

In this section we consider applying the fast VAMS procedure to two specific robust forward addition sequences. The first is the  $t_3$  forward addition sequence presented in Section 3.2.3. Using this forward addition sequence, we expect the final procedure to have the favorable robustness properties discussed for the  $t_3$  RBIC method. The second robust forward addition sequence is developed using the generalized score statistic based on the bi-square  $\rho$  function (1.14), and corresponding  $\psi$  function

$$\psi_B(x) = 6x(1 - x^2)^2 I(|x| \leq 1). \quad (3.35)$$

In (3.35),  $I(\cdot)$  is an indicator function that is equal to 1 if  $x$  is less than or equal to 1 in absolute value and zero otherwise. The restricted estimates  $\tilde{\beta}_0$ ,  $\tilde{\beta}$ , and  $\tilde{\sigma}$



required to calculate the generalized score statistic are found using MM-estimation. This approach coincides with the one used to construct the Huber and  $t_3$  sequences since aside from high breakdown starting estimates, MM-estimates are found using (3.35) and the basic M-estimation procedure. Note that the development of an RBIC approach using MM-estimation is less straightforward, because unlike the  $t_3$  and Huber  $\rho$  functions, (1.14) is not associated with a particular distribution (Maronna, 2006, p. 29).

The robust VAMS procedure is as follows. For both the  $t_3$  M-estimation and bi-square MM-estimation approaches, the score forward addition sequence and the corresponding sequential  $p$ -values found using (3.34) are passed to the fast VAMS procedure (Section 2.2). An estimate of the false selection rate is calculated using the formula

$$\hat{\gamma}_F(\alpha) = \frac{\{k_T - S(\alpha)\}\alpha}{1 + S(\alpha)}, \quad (3.36)$$

where  $S(\alpha)$  is the number of variables that enter the model when applying the robust forward selection algorithm with tuning parameter  $\alpha$ . An easy way to find  $S(\alpha)$  is to use (2.11) to obtain  $\tilde{p}$  from the robust forward addition sequence's sequential  $p$ -values, and simply count the number of monotonized  $p$ -values less  $\alpha$ . Using (3.36), the fast VAMS estimate of the forward selection tuning parameter is

$$\hat{\alpha}_F = \sup_{\alpha \leq \alpha_m} \{\alpha : \hat{\gamma}_F(\alpha) \leq \gamma_0\}. \quad (3.37)$$

We call the fast VAMS procedure based on the  $t_3$   $\psi$  function  $\text{VAMS}_{t_3}$  and that based on the bisquare  $\psi$  function  $\text{VAMS}_{Bi}$ . This gives two robust versions of the VAMS procedure to compete with the dual estimation indicator approaches of Chapter 2. The robustness properties and selection characteristics of each approach is examined in the next section.

## 3.4 Comparison of Methods

### 3.4.1 Simulation

The simulation design used in Section 2.5.2 is used to study the robust regression variable selection procedures, including the Huber and  $t_3$  RBIC methods,  $\text{RBIC}_H$  and  $\text{RBIC}_{t_3}$  respectively, along with the VAMS methods of Section 3.3. The results are presented in Tables 3.1-3.3.

### 3.4.2 Results

Starting with the  $\text{RBIC}_H$  and  $\text{RBIC}_{t_3}$  methods, in Table 3.1 we see that for the Normal and  $t_3$  errors the Huber-based approach has a slight advantage in terms of the ratio of median model errors. This result is somewhat surprising since the  $\text{RBIC}_{t_3}$  is designed to handle  $t_3$  errors. However, with a standard error of .11 and very similar selection characteristics the difference is not significant. This conclusion is further supported by the equality of the ratio of mean model errors for both the Normal and  $t_3$  errors. The real difference between the two procedures is seen when the errors have a Cauchy distribution. The models selected by the  $\text{RBIC}_H$  procedure tend to be drastically under-fit. The procedure selects on average just 1.24 of the four informative predictors compared to 3.93 out of four for the  $\text{RBIC}_{t_3}$  procedure. The drastic difference in selected models is due to the high-magnitude outliers produced by the Cauchy distribution. As discussed in Section 3.2, the  $\text{RBIC}_{t_3}$  procedure is better able to dampen the effects of gross outliers.

When  $n = 100$  and  $k_T = 50$  we see similar results. Again the Huber and  $t_3$  methods have comparable results when the errors are distributed Normal or  $t_3$ . The  $\text{RBIC}_H$  method does slightly better than its  $t_3$  counterpart in terms of both model error diagnostics when the errors are Normally distributed. The  $\text{RBIC}_H$  procedure also selects on average 1.14 fewer unimportant predictors than the  $\text{RBIC}_{t_3}$  method. The  $\text{RBIC}_{t_3}$  method again shows drastic improvement when the errors are sampled from a Cauchy distribution.

For the final set of conditions,  $n = 500$  and  $k_T = 50$ . Much like we saw for the VAMS procedures in the previous chapter, the relationships that hold when  $n = 100$  and  $k_T = 20$  continue to hold under these conditions. The Huber and  $t_3$  methods provide very similar results for the first two error distributions but the  $t_3$  method does a better job when the errors are Cauchy.

Now that we have compared the two RBIC methods we turn our focus to the  $\text{VAMS}_{t_3}$  and  $\text{VAMS}_{Bi}$  methods. These two methods have similar results across all simulation conditions. The  $\text{VAMS}_{Bi}$  method has a slight edge over the  $\text{VAMS}_{t_3}$  in terms of model error when the errors are normally distributed. The procedure also tends to have better average selection characteristics when the errors have a Cauchy distribution. Similar to the VAMS procedures in Chapter 2 these procedures do an excellent job of controlling the false selection rate when the number of observations is large compared to the number of candidate predictors. Only when  $n = 100$  and  $k_T = 50$  do the procedures fail to keep the average false selection rate very near the target value of .05, see Table 3.2.

Comparing between the RBIC and VAMS procedures, the biggest differences occur under the second set of conditions. There is clearly a tradeoff between controlling the false selection rate and having a low model error. The RBIC procedures do a significantly better job in terms of model error while the  $\text{VAMS}_{t_3}$  and  $\text{VAMS}_{Bi}$  methods have significantly lower false selection rates. While the  $\text{VAMS}_{t_3}$  and  $\text{VAMS}_{Bi}$  methods have a false selection rate above the target value of .05, it is still well below that of the RBIC methods. For normal errors the average false selection rate of the RBIC procedures gets up to three times the average false selection rate of the VAMS procedures. Under the other two sets of conditions, when the number of observations is large compared to the number of predictors, the separation between the two classes of procedures is less pronounced. When  $n = 100$  and  $k_T = 20$  the VAMS procedures do a better job both in terms of model error and false selection rate for normally distributed errors. For  $t_3$  errors, the RBIC methods have a higher ratio of mean model errors while the VAMS methods have a high ratio of median model

errors. The disagreement between the two model error diagnostics is due to a few large model errors for the VAMS procedures. Procedures that tend to select smaller models are more likely to have a few large model errors since they have a better chance of omitting informative predictors from the final model. Omitting informative predictors has a much more negative effect on model error than the inclusion of uninformative variables. For Cauchy errors the two model error ratios are again in agreement with the  $\text{RBIC}_{t_3}$  procedure doing slightly better in terms of each. For  $n = 500$  and  $k_T = 50$ , both the RBIC and VAMS procedures do an excellent job of distinguishing between the informative and uninformative predictors. For normally distributed errors the VAMS procedures have an edge in terms of both model error and false selection. However, for  $t_3$  and Cauchy errors the  $\text{RBIC}_{t_3}$  method does better than both the VAMS procedures. The ability of the  $\text{RBIC}_{t_3}$  procedure to control the false selection rate in these cases is due to the strength of the relationship between the response and the informative predictors. With this strong of a relationship the  $\text{RBIC}_{t_3}$  method has the advantage of being able to have a very small false selection rate while the VAMS methods still attempt to obtain an average false selection rate of .05. Note that in practice it is unlikely to have a data set where the separation between informative and uninformative predictors is so clear.

### 3.4.3 Breakdown

The Breakdown of the Robust Regression variable selection methods are calculated following the process outlined in Section 2.6. The  $\text{RBIC}_H$  approach has a breakdown point of zero, a fact that explains the poor simulation results seen in the last section for Cauchy errors. The  $t_3$  RBIC and VAMS methods on the other hand, have a breakdown point of .25. The result coincides with one found in He et al. (2000) where it was shown that M-estimates of regression coefficients developed from the  $t$ -distribution have a breakdown point of  $1/(v + 1)$ , where  $v$  is the degrees of freedom of the  $t$ -distribution. The  $\text{VAMS}_{B_i}$  procedure exhibits a breakdown point of .5. Again this coincides with the breakdown point of the LMS and MM-estimation procedures

Table 3.1: Simulations Results for  $n=100$ ,  $k_T=20$ , and  $k_I=4$ 

Normal Errors						
Method	Ratio of Med. ME	Ratio of Mean ME	CS	FS	Size	FSR
RBIC <sub>H</sub>	0.51	0.43	3.77	0.71	4.48	0.11
RBIC <sub>t<sub>3</sub></sub>	0.49	0.43	3.85	0.81	4.66	0.12
VAMS <sub>t<sub>3</sub></sub>	0.68	0.45	3.68	0.20	3.88	0.04
VAMS <sub>Bi</sub>	0.77	0.50	3.74	0.23	3.97	0.04
Max SE	0.10	0.04	0.06	0.10		0.01
$t_3$ Errors						
Method	Ratio of Med. ME	Ratio of Mean ME	CS	FS	Size	FSR
RBIC <sub>H</sub>	0.52	0.42	3.84	0.76	4.60	0.12
RBIC <sub>t<sub>3</sub></sub>	0.47	0.42	3.85	0.77	4.62	0.12
VAMS <sub>t<sub>3</sub></sub>	0.64	0.40	3.68	0.33	4.01	0.05
VAMS <sub>Bi</sub>	0.68	0.36	3.61	0.31	3.92	0.05
Max SE	0.11	0.04	0.07	0.10		0.01
Cauchy						
Method	Ratio of Med. ME	Ratio of Mean ME	CS	FS	Size	FSR
RBIC <sub>H</sub>	0.03	0.03	1.24	0.30	1.54	0.03
RBIC <sub>t<sub>3</sub></sub>	0.57	0.42	3.93	0.33	4.25	0.05
VAMS <sub>t<sub>3</sub></sub>	0.54	0.27	3.78	0.34	4.12	0.06
VAMS <sub>Bi</sub>	0.46	0.30	3.84	0.34	4.18	0.06
Max SE	0.09	0.05	0.06	0.07		0.01

CS=average # correctly selected, FS=average # falsely selected, Size=CS+FS

Note: Entries are based on 100 replicates

Max SE: maximum standard error across the methods

Table 3.2: Simulations Results for  $n=100$ ,  $k_T=50$ , and  $k_I=10$ 

Normal Errors						
Method	Ratio of Med. ME	Ratio of Mean ME	CS	FS	Size	FSR
RBIC <sub>H</sub>	0.25	0.27	6.09	2.91	9.00	0.27
RBIC <sub>t<sub>3</sub></sub>	0.23	0.26	6.69	4.05	10.74	0.33
VAMS <sub>t<sub>3</sub></sub>	0.18	0.20	3.12	0.55	3.67	0.10
VAMS <sub>Bi</sub>	0.18	0.20	3.12	0.62	3.74	0.11
Max SE	0.02	0.01	0.19	0.22		0.02
$t_3$ Errors						
Method	Ratio of Med. ME	Ratio of Mean ME	CS	FS	Size	FSR
RBIC <sub>H</sub>	0.26	0.27	5.74	2.32	8.06	0.22
RBIC <sub>t<sub>3</sub></sub>	0.26	0.28	6.39	3.21	9.60	0.27
VAMS <sub>t<sub>3</sub></sub>	0.19	0.20	3.12	0.47	3.59	0.10
VAMS <sub>Bi</sub>	0.17	0.19	2.87	0.51	3.38	0.12
Max SE	0.02	0.01	0.18	0.19		0.02
Cauchy						
Method	Ratio of Med. ME	Ratio of Mean ME	CS	FS	Size	FSR
RBIC <sub>H</sub>	0.07	0.075	2.17	1.68	3.85	0.12
RBIC <sub>t<sub>3</sub></sub>	0.14	0.15	5.52	2.08	7.60	0.22
VAMS <sub>t<sub>3</sub></sub>	0.09	0.10	2.85	0.56	3.41	0.13
VAMS <sub>Bi</sub>	0.10	0.11	3.45	0.55	4.00	0.10
Max SE	0.01	0.01	0.23	0.17		0.02

CS=average # correctly selected, FS=average # falsely selected, Size=CS+FS

Note: Entries are based on 100 replicates

Max SE: maximum standard error across the methods

Table 3.3: Simulations Results for  $n=500$ ,  $k_T=50$ , and  $k_I=10$ 

Normal Errors						
Method	Ratio of Med. ME	Ratio of Mean ME	CS	FS	Size	FSR
$RBIC_H$	0.62	0.56	9.89	0.96	10.85	0.08
$RBIC_{t_3}$	0.53	0.53	9.96	1.00	10.96	0.08
$VAMS_{t_3}$	0.60	0.61	9.97	0.60	10.57	0.05
$VAMS_{Bi}$	0.68	0.66	10.00	0.66	10.66	0.05
Max SE	0.05	0.03	0.04	0.10		0.01
$t_3$ Errors						
Method	Ratio of Med. ME	Ratio of Mean ME	CS	FS	Size	FSR
$RBIC_H$	0.83	0.71	9.97	0.50	10.47	0.04
$RBIC_{t_3}$	0.82	0.73	9.97	0.47	10.44	0.04
$VAMS_{t_3}$	0.75	0.71	9.97	0.59	10.56	0.05
$VAMS_{Bi}$	0.75	0.70	9.98	0.49	10.47	0.05
Max SE	0.06	0.04	0.02	0.09		0.01
Cauchy						
Method	Ratio of Med. ME	Ratio of Mean ME	CS	FS	Size	FSR
$RBIC_H$	0.012	0.011	1.38	0.48	1.86	0.003
$RBIC_{t_3}$	0.75	0.71	10.0	0.12	10.12	0.01
$VAMS_{t_3}$	0.52	0.55	10.00	0.66	10.66	0.05
$VAMS_{Bi}$	0.52	0.52	10.00	0.66	10.66	0.05
Max SE	0.06	0.03	—	0.10		0.01

CS=average # correctly selected, FS=average # falsely selected, Size=CS+FS

Note: Entries are based on 100 replicates

Max SE: maximum standard error across the methods

central to the procedure. Therefore, much like for the VAMS procedures in the previous chapter, there is a carryover of the breakdown point seen in regression.

### 3.4.4 Conclusion

The failure of the Huber-based RBIC method for Cauchy errors indicates that it is not appropriate for use as a robust variable selection procedure. With the breakdown point of zero, the procedure is sensitive to even a single large magnitude outlier. The  $\text{RBIC}_{t_3}$  procedure has favorable results across each of the error distributions. In terms of model error the procedure is competitive with the dual estimation VAMS approaches chosen in Chapter 2. The procedure does an excellent job for heavy-tailed errors when the number of observations greatly exceed the number of predictors. With a breakdown point of .25 the procedure can handle the typical amount of outlying points seen in real data sets.

The simulation results provide evidence that the  $\text{VAMS}_{t_3}$  and  $\text{VAMS}_{Bi}$  procedures are also reasonable choices for a robust variable selection procedure. Each procedure is competitive in terms of model error and there is evidence that these methods do the best job of controlling the false selection rate. The breakdown points of .25 and .50 are also quite a bit larger than the .10 seen for the other false selection rate methods. The  $\text{VAMS}_{t_3}$  and  $\text{VAMS}_{Bi}$  procedures do have small model error ratios for the case when  $n = 100$  and  $k_T = 50$ . Generally small model error ratios are a result of a selection method failing to select enough of the informative predictors. For false selection rate methods like  $\text{VAMS}_{t_3}$  and  $\text{VAMS}_{Bi}$  there are two possible explanations for failing to select these variables. The first is that these methods are doing a poor job of separating the informative and uninformative variables. The second is that the failure of the methods to select enough informative predictors is directly related to each methods ability to control the false selection rate. While the data seems to support the second of these two explanations, future research can give further insight on this issue.



# Chapter 4

## Lasso-Based Methods

In this chapter we introduce a new class of robust variable selection procedures. Unlike the subset selection methods studied previously, the procedures presented here are variations of the LASSO. As such, we begin with an introduction to the L1 penalized version of the least squares problem. We then summarize the development of robust versions of the procedure through replacement of squared error loss function.

### 4.1 LASSO

Tibshirani (1996) proposed the least absolute shrinkage and selection operator or LASSO variable selection method. For the remainder of this chapter we consider the linear model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{4.1}$$

where the response vector  $\mathbf{Y}$  and the columns of the design matrix  $\mathbf{X}$  are centered through subtraction of each vectors sample mean. Each column of  $\mathbf{X}$  is also scaled to have a sample variance of one. Centering and scaling the data removes the need for an intercept and allows the results in this chapter to be presented in the standard format seen throughout the LASSO literature. The LASSO method looks to improve

prediction accuracy and interpretability over traditional methods through shrinkage of the regression coefficients. In the most basic setting, LASSO estimates are obtained by minimizing the sum of squared residuals subject to the constraint  $\sum_j |\beta_j| \leq t$ . The non-negative tuning parameter  $t$ , controls the amount of shrinkage applied to the regression coefficients. The LASSO problem has the dual representation

$$\hat{\boldsymbol{\beta}} = \arg \min \sum_{i=1}^N (Y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2 + \lambda \sum_j |\beta_j|, \quad (4.2)$$

where  $\lambda$  is a regularization parameter. The regularization parameter  $\lambda$  and the L1 bound  $t$  are related in a one-to-one fashion with larger values of  $\lambda$  corresponding to smaller  $t$ . As a consequence, the regression coefficients shrink towards zero as  $\lambda$  increases. The ability of the LASSO to shrink coefficients to zero allows for simultaneous variable selection and parameter estimation. For a particular choice of  $\lambda$ , variables with coefficients equal to zero are removed from the model and the resulting subset of initial predictors determined by LASSO are considered informative.

The LASSO procedure was later extended to allow each regression coefficient to have its own shrinkage parameter (Zou, 2006). In this new method, the adaptive LASSO, parameter estimates are defined via

$$\hat{\boldsymbol{\beta}} = \arg \min \sum_{i=1}^N (Y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2 + \lambda \sum_j \hat{w}_j |\beta_j|, \quad (4.3)$$

where weights  $\hat{w}_j = 1/|\hat{\beta}_j(ols)|$  are a typical choice and  $\lambda$  is again a regularization parameter. This modification allows for differential shrinkage and leads to more sparse solution sets, helping to avoid biases sometimes seen in regression coefficients obtained when using the original LASSO (Wang and Leng, 2007). In the next two sections we discuss variable selection methods that are modifications of the adaptive LASSO.

## 4.2 Least Squares Approximation to the Adaptive LASSO

In the previous section, we saw that the adaptive LASSO procedure obtained regression estimates through minimization of the sum of squared residuals subject to a constraint. Alternatively, we can look at the adaptive LASSO as a more general problem of minimizing the objective function

$$\frac{1}{n}L_n(\beta) + \lambda \sum_j \hat{w}_j |\beta_j|, \quad j = 1, \dots, p \quad (4.4)$$

where  $L_n(\beta)$  is a particular loss function. In (4.4) it is common to use weights  $\hat{w}_j = 1/|\tilde{\beta}_j|$  where  $\tilde{\beta}$  is the unrestricted minimizer of  $L_n(\beta)$ . The classical adaptive LASSO is then a special case of this general approach with  $L_n(\beta) = n \sum (\mathbf{Y}_i - \mathbf{x}_i^T \beta)^2$ ,  $\hat{w}_j = 1/|\hat{\beta}_j(ols)|$ , and the optimal  $\lambda$  found using cross validation.

Wang and Leng (2007) propose a method of least squares approximation (LSA) for this general adaptive LASSO problem. The LSA adaptive LASSO procedure replaces the adaptive LASSO objective function with an asymptotically equivalent least squares problem. The procedure is developed using the Taylor series expansion of  $n^{-1}L_n(\beta)$  around  $\tilde{\beta}$ . Following the notation of Wang and Leng (2007), this expansion gives

$$\frac{1}{n}L_n(\beta) \approx \frac{1}{n}L_n(\tilde{\beta}) + \frac{1}{n}\dot{L}_n(\tilde{\beta})^T(\beta - \tilde{\beta}) + \frac{1}{2}(\beta - \tilde{\beta})^T \left\{ \frac{1}{n}\ddot{L}_n(\tilde{\beta}) \right\} (\beta - \tilde{\beta}), \quad (4.5)$$

where  $\dot{L}_n(\tilde{\beta})$  and  $\ddot{L}_n(\tilde{\beta})$  are the first and second derivative of the loss function evaluated at  $\tilde{\beta}$ . Since  $\tilde{\beta}$  is the minimizer of  $L_n(\beta)$  we know that  $\dot{L}_n(\tilde{\beta}) = 0$ . Therefore

$$\frac{1}{n}L_n(\beta) \approx \frac{1}{n}L_n(\tilde{\beta}) + \frac{1}{2}(\beta - \tilde{\beta})^T \left\{ \frac{1}{n}\ddot{L}_n(\tilde{\beta}) \right\} (\beta - \tilde{\beta}). \quad (4.6)$$

Ignoring the constant  $n^{-1}L_n(\tilde{\beta})$  and the coefficient of one half in front of the second term the result is an approximation of  $n^{-1}L_n(\beta)$  that is in the form of a least squares function. That is,

$$\frac{1}{n}L_n(\beta) \approx (\beta - \tilde{\beta})^T \hat{\Sigma}^{-1}(\beta - \tilde{\beta}) \quad (4.7)$$

where  $\widehat{\Sigma}^{-1} = n^{-1}\ddot{L}_n(\tilde{\beta})$ .

Rewriting the original adaptive LASSO objective function given in (4.4), with the least squares function approximating  $n^{-1}L_n(\beta)$ , we have the new objective function,

$$Q(\beta) = (\beta - \tilde{\beta})^T \widehat{\Sigma}^{-1}(\beta - \tilde{\beta}) + \lambda \sum_{j=1}^d \widehat{w}_j |\beta_j|. \quad (4.8)$$

The value of  $\beta$  minimizing (4.8) is the LSA adaptive LASSO estimate. This estimate differs slightly from the original adaptive LASSO estimate found through minimization of (4.4). The only requirement of (4.8) is the existence of a consistent covariance matrix estimate  $\widehat{\Sigma}$ , which allows for this general approach to be extended to most regression methods. Wang and Leng (2007) provide R code to obtain LSA adaptive LASSO estimates for a variety of loss functions. The code finds the optimal  $\lambda$  in (4.8) using the Bayesian information criteria (Wang and Leng, 2007)

$$BIC_\lambda = (\widehat{\beta}_\lambda - \tilde{\beta})^T \widehat{\Sigma}^{-1}(\widehat{\beta}_\lambda - \tilde{\beta}) + \frac{p \log(n)}{n}. \quad (4.9)$$

where  $\widehat{\beta}_\lambda$  are the shrinkage estimates associated with the regularization parameter  $\lambda$  and  $p$  is the number of these estimates that are non-zero.

Using the `r1m` function in R we can combine the LSA adaptive LASSO with the Huber and  $t_3$  robust regression approaches outlined in Sections 3.1 and 3.2. By doing so we have two more methods that simultaneously selects variables and adjusts for potential outliers. Note that the standard LSA adaptive LASSO code provided by Wang and Leng (2007) uses the the sandwich covariance matrix (1.7) output by the `r1m` procedure in place of the second derivative matrix in the loss function approximation (4.7). Also, following the advice of Dr. Wang and Dr. Leng slight modifications were made to the original LSA adaptive LASSO procedure to better handle the objects output by the `r1m` procedure.

### 4.3 LAD-LASSO

The LAD-LASSO procedure is a combination of the least absolute deviation (LAD) regression and the LASSO procedures. LAD regression, also referred to as

median regression, is a robust alternative to least squares. This procedure obtains regression parameter estimates via minimization of the sum of the absolute residuals. Combining the LAD with LASSO penalization, one can retain the robustness provided by LAD regression while allowing for simultaneous variable selection.

Much like the LSA procedure, the LAD-LASSO is developed using the general adaptive LASSO objective function. Plugging the LAD absolute error loss function into (4.4) results in,

$$\sum_{i=1}^n |Y_i - \mathbf{x}_i^T \boldsymbol{\beta}| + n \sum_{j=1}^p \lambda_j |\beta_j|. \quad (4.10)$$

Before we can find the regression coefficients that minimize (4.10), we need to obtain a set of optimal regularization parameters. For the basic adaptive LASSO procedure with a squared error loss function, the  $\lambda_j$ 's are found via cross validation (Tibshirani, 1996; Zou, 2006). Cross validation is computationally time consuming for absolute error loss. To avoid this issue, Wang and Leng (2007) propose estimating the  $\lambda_j$ s via minimization of the BIC-type objective function

$$\sum_{i=1}^n |Y_i - \mathbf{x}_i^T \boldsymbol{\beta}| + n \sum_{j=1}^p \lambda_j |\beta_j| - \log(.5n\lambda_j) \log(n). \quad (4.11)$$

This objective function differs from (4.10) only by the last term which is penalizing for model complexity. This penalty is a factor of  $\log(n)$ , much like we see in the general BIC formula (3.15). Minimization of (4.11) leads to

$$\lambda_j = \frac{\log(n)}{n|\beta_j|}.$$

Since  $\beta_j$  is an unknown regression parameter, we plug in the unpenalized LAD estimates,  $\tilde{\boldsymbol{\beta}}$ , to get the regularization parameter estimates

$$\hat{\lambda}_j = \frac{\log(n)}{n|\tilde{\beta}_j|}$$

The LAD-LASSO solution is then the beta vector that minimizes (4.10) with  $\hat{\lambda}_j$  plugged in for  $\lambda_j$ .

An appealing feature of this method is that it has a relatively simple implementation. All that is needed is a statistical software package capable of performing LAD regression. First, one conducts a LAD regression fit on the original data set. The coefficients from this fit are the  $\tilde{\beta}_j$  which are used to calculate the tuning parameter estimates  $\hat{\lambda}_j$ . Then one simply appends  $p$  new observations to the data set of the form,  $(Y_{n+j}, \mathbf{x}_{n+j}^T) = (0, n\hat{\lambda}_j \mathbf{e}_j)$  for  $j = 1, \dots, p$ . Here  $\mathbf{e}_j$  is a  $p$ -dimensional row vector with the  $j^{th}$  entry equal to 1 and all others 0. Conducting a LAD fit on this augmented data results in regression coefficients  $\hat{\beta}_j$  minimizing

$$\sum_{i=1}^{n+p} |Y_i - \mathbf{x}_i^T \boldsymbol{\beta}| = \sum_{i=1}^n |Y_i - \mathbf{x}_i^T \boldsymbol{\beta}| + n \sum_j \hat{\lambda}_j |\beta_j| \quad (4.12)$$

Estimates obtained from minimization of formula (4.12) are equal to the LAD-LASSO estimates.

## 4.4 Comparision of Methods

### 4.4.1 Simulation

The simulation design used in Section 2.5.2 is used to study the LAD-LASSO and the two variations of the LSA adaptive LASSO procedure. The LSA adaptive LASSO procedure using the Huber loss function is designated  $\text{LSA}_H$ , while the  $t_3$  version of the procedure is labeled  $\text{LSA}_{t_3}$ . The results are presented in Tables 4.1-4.3.

### 4.4.2 Results

Under each of the simulation conditions the LAD-LASSO has smallest model error ratios and fewest number of correctly selected variables across all the robust procedures presented in this paper. The procedure has a tendency to under-fit for even the most well behaved simulation conditions. Therefore, we exclude the LAD-LASSO from the rest of this discussion, declaring it uncompetitive.

In Table 4.1 the  $\text{LSA}_H$  and  $\text{LSA}_{t_3}$  procedures perform adequately at all error distributions. It is interesting to note that the  $\text{LSA}_H$  procedure does not tend to drastically under-fit for Cauchy errors like the Huber-based  $\text{RBIC}_H$  procedure of Chapter 3. However, like the RBIC methods the  $t_3$  version of the LSA procedure shows significant improvement over the Huber-based approach for Cauchy errors. In Table 4.3 we see that the ratio of median model errors is .29 for the  $\text{LSA}_H$  compared to .43 for the  $\text{LSA}_{t_3}$ . While both methods are able to select all  $k_I=10$  of the informative variables the  $\text{LSA}_{t_3}$  procedure does a better job of limiting the number of falsely selected predictors. For normal and  $t_3$  errors the two variations of the LSA adaptive LASSO perform comparably. The  $\text{LSA}_{t_3}$  procedure had slightly better selection characteristics for  $t_3$  distributed errors but the difference is not significant.

The high false selection rates in Tables 4.1 and 4.2 suggest that like the RBIC procedures in Chapter 3, the LSA adaptive LASSO is best for prediction accuracy. This is further supported by the relatively high ratio of median model errors in Table 4.2 when  $n = 100$  and  $k_T = 50$ .

### 4.4.3 Breakdown

The results for the  $\text{LSA}_H$  and  $\text{LSA}_{t_3}$  procedures were obtained using R code provided by Wang and Leng (2007). To use this code the Huber and  $t_3$  robust regression fits were found using the M-estimation approach to regression with a Proposal 2 estimate of scale. As defined, each of the resulting methods has a breakdown point of zero. However, it is reasonable to believe that the LSA adaptive LASSO code can be adjusted to incorporate high breakdown methods such as MM-estimation and LMS. While the complete modification of the LSA adaptive LASSO procedure is beyond the scope of this thesis, preliminary testing suggests that with these adjustments the LSA adaptive LASSO method would maintain selection characteristics similar to those seen in Tables 4.1-4.3 while obtaining a breakdown point that coincides with the other robust variable selection methods. Specifically, there was evidence that the  $t_3$  version of the procedure would again have a breakdown point of .25.

#### 4.4.4 Conclusion

With the LAD-LASSO drastically under-fitting, the two versions of the LSA adaptive LASSO procedure are the most competitive LASSO-based methods. The ability of the  $\text{LSA}_{t_3}$  procedure to better handle Cauchy errors gives it a distinct advantage over the  $\text{LSA}_H$  method. Despite the increase in efficiency there are still signs that the LSA adaptive LASSO does not perform as well for Cauchy distributed errors as the methods presented in the previous two chapters. Further comparison to these other methods is done in Chapter 5.



Table 4.1: Simulations Results for  $n=100$ ,  $k_T=20$ , and  $k_I=4$ 

Method	Ratio of Med. ME	Normal Errors				
		Ratio of Mean ME	CS	FS	Size	FSR
$LSA_H$	0.43	0.42	3.88	1.10	4.98	0.14
$LSA_{t_3}$	0.44	0.42	3.87	1.14	5.01	0.15
LAD-LASSO	0.24	0.25	3.71	0.59	4.30	0.10
Max SE	0.04	0.02	0.07	0.16		0.02
Method	Ratio of Med. ME	$t_3$ Errors				
		Ratio of Mean ME	CS	FS	Size	FSR
$LSA_H$	0.42	0.41	3.90	1.22	5.12	0.17
$LSA_{t_3}$	0.40	0.42	3.91	1.16	5.07	0.16
LAD-LASSO	0.20	0.21	3.69	0.32	4.01	0.06
Max SE	0.05	0.02	0.10	0.14		0.02
Method	Ratio of Med. ME	Cauchy				
		Ratio of Mean ME	CS	FS	Size	FSR
$LSA_H$	0.21	0.18	3.84	0.93	4.77	0.13
$LSA_{t_3}$	0.28	0.23	3.90	0.89	4.79	0.12
LAD-LASSO	0.04	0.04	2.43	0.02	2.45	0.01
Max SE	0.04	0.03	0.05	0.12		0.02

CS=average # correctly selected, FS=average # falsely selected, Size=CS+FS

Note: Entries are based on 100 replicates

Max SE: Maximum standard error across methods excluding LAD-LASSO

Table 4.2: Simulations Results for  $n=100$ ,  $k_T=50$ , and  $k_I=10$ 

Method	Ratio of Med. ME	Normal Errors				
		Ratio of Mean ME	CS	FS	Size	FSR
$LSA_H$	0.30	0.32	6.89	4.51	11.4	0.32
$LSA_{t_3}$	0.30	0.32	6.86	4.41	11.27	0.32
LAD-LASSO	0.20	0.23	5.34	1.87	7.21	0.22
Max SE	0.03	0.02	0.16	0.33		0.02
Method	Ratio of Med. ME	$t_3$ Errors				
		Ratio of Mean ME	CS	FS	Size	FSR
$LSA_H$	0.25	0.26	5.70	3.47	9.17	0.31
$LSA_{t_3}$	0.26	0.27	5.86	3.75	9.61	0.32
LAD-LASSO	0.20	0.21	4.52	1.07	5.59	0.16
Max SE	0.02	0.01	0.20	0.29		0.02
Method	Ratio of Med. ME	Cauchy				
		Ratio of Mean ME	CS	FS	Size	FSR
$LSA_H$	0.08	0.08	3.20	2.34	5.54	0.23
$LSA_{t_3}$	0.11	0.10	4.63	3.23	7.86	0.28
LAD-LASSO	0.06	0.07	1.53	0.22	1.75	0.06
Max SE	0.01	0.01	0.26	0.32		0.02

CS=average # correctly selected, FS=average # falsely selected, Size=CS+FS

Note: Entries are based on 100 replicates

Max SE: Maximum standard error across methods excluding LAD-LASSO

Table 4.3: Simulations Results for  $n=500$ ,  $k_T=50$ , and  $k_I=10$ 

Method	Ratio of Med. ME	Normal Errors				
		Ratio of Mean ME	CS	FS	Size	FSR
$LSA_H$	0.45	0.49	10.0	1.16	11.16	0.09
$LSA_{t_3}$	0.46	0.49	10.0	1.12	11.12	0.09
LAD-LASSO	0.27	0.27	9.93	0.56	10.49	0.05
Max SE	0.03	0.02	—	0.12		0.01
Method	Ratio of Med. ME	$t_3$ Errors				
		Ratio of Mean ME	CS	FS	Size	FSR
$LSA_H$	0.48	0.48	9.99	1.02	11.01	0.08
$LSA_{t_3}$	0.52	0.48	9.98	0.99	10.97	0.07
LAD-LASSO	0.24	0.24	9.88	0.13	10.01	0.01
Max SE	0.03	0.02	0.01	0.13		0.01
Method	Ratio of Med. ME	Cauchy				
		Ratio of Mean ME	CS	FS	Size	FSR
$LSA_H$	0.29	0.31	10.0	0.98	10.98	0.07
$LSA_{t_3}$	0.43	0.43	10.0	0.74	10.74	0.06
LAD-LASSO	0.05	0.05	9.68	0	9.68	0.00
Max SE	0.03	0.02	—	0.12		0.01

CS=average # correctly selected, FS=average # falsely selected, Size=CS+FS

Note: Entries are based on 100 replicates

Max SE: Maximum standard error across methods excluding LAD-LASSO

# Chapter 5

## Conclusion

### 5.1 False Selection Rate versus Prediction

The variable selection methods presented in this thesis can be divided into two types of procedures: those that are designed to control the false selection rate and those designed for prediction. The four VAMS-based procedures:  $VASMI_1$ ,  $VAMSI_2$ ,  $VAMS_{t_3}$ , and  $VAMS_{Bi}$  fall into the first type, and the  $RBIC_{t_3}$  and LSA adaptive LASSO methods in the second. The difference in the basic design of the two types of procedures leads to different selection characteristics. It is intuitive that the VAMS-based methods tend to select smaller models in an effort to control the false selection rate. The  $RBIC_{t_3}$  and LSA adaptive LASSO methods are more likely to select slightly larger models in order to have enough informative predictors for accurate predictions. Of the simulation conditions presented in this paper, the case when  $n = 100$  and  $k_T = 50$  best displays the tradeoff between these two types of procedures. Under these conditions it is difficult for the variable selection procedures to make the separation between uninformative and informative predictors. As a result, controlling the false selection rate means failing to include some of the informative predictors. Conversely, attempting to include a large proportion of the informative variables results in a high proportion of uninformative predictors being selected.

Among the four variations of the VAMS procedure the  $\text{VAMS}_{t_3}$  and  $\text{VAMS}_{Bi}$  do the best job of keeping the false selection rate near the target value of .05. These procedures do slightly better than the  $\text{VASMI}_1$  and  $\text{VAMSI}_2$ , the two indicator methods, when either the errors come from heavier-tailed distribution or the number of predictors approaches the number of observations. With similar selection results, the high breakdown point of the  $\text{VAMS}_{Bi}$  procedure gives it an edge over  $\text{VAMS}_{t_3}$ . While the  $\text{VASMI}_1$  and  $\text{VAMSI}_2$  methods do slightly worse in terms of controlling the false selection rate, these procedures have an advantage in terms of simplicity. Each of these methods can be implemented by appending an identity matrix to the original design matrix and using any statistical software that can perform forward selection. The  $\text{VAMS}_{t_3}$  and  $\text{VAMS}_{Bi}$  require software that can both obtain MM-estimates of regression coefficients and optimize complex likelihood functions. The final model obtained by the indicator methods is also easier to interpret. The indicator variables in the final model show exactly which observations have been removed from the data set. With easily interpretable results, the indicator approach is a viable option for the researcher less familiar with statistics.

Among the prediction-oriented methods,  $\text{RBIC}_{t_3}$  distinguishes itself as the best method. The  $\text{RBIC}_H$  and LAD-LASSO methods drastically under-fit when errors came from a Cauchy distribution. Because each of these methods is not robust enough to handle large outliers, the two variations of the LSA adaptive LASSO are really the only competitors for the  $\text{RBIC}_{t_3}$  method. Between these two approaches, the improved selection results for Cauchy errors gives the  $\text{LSA}_{t_3}$  method an edge over the Huber-based approach. The  $\text{RBIC}_{t_3}$  method exhibits a distinct advantage over the  $\text{LSA}_{t_3}$  method in terms of model error with a higher ratio of median model errors for nearly all of the simulations conditions. The advantage of the  $\text{RBIC}_{t_3}$  procedure over the  $\text{LSA}_{t_3}$  is most obvious for Cauchy distributed errors.

Decisions on the appropriate methods to use should be determined by the nature of the data and the goal of the researcher. If it is important to control the number of falsely selected variables, then the  $\text{VAMS}_{Bi}$  or  $\text{VAMSI}_2$  procedures have a distinct

advantage over the  $\text{RBIC}_{t_3}$ . In terms of prediction, there is evidence that when the number of observations is significantly larger than the number of candidate predictors, all three methods are comparable. The ability of the VAMS procedures to control the false selection rate seems to give it the edge.

## 5.2 Breakdown

Throughout this paper the simulation-based breakdown point of the most competitive robust variable selection methods are presented. Despite being generally accepted as a useful measure of robustness, the level of breakdown required for a procedure to handle typical real life data sets is unclear. However, with typical data sets having high-magnitude errors in the range of 1 to 10 percent, the need for procedures to handle some percentage of outlying points is apparent (Hampel, 1974). In Chapter 2 we see that the dual estimation VAMS methods, which only have a breakdown point of .10, give favorable simulation results for all error distributions. The ability to handle Cauchy distributed errors is especially noteworthy. It is not unreasonable to believe that most typical data sets have errors coming from a distribution less heavy-tailed than the Cauchy. So having a breakdown point of .10 will be adequate for many sampling situations.

It is also worth noting that in this paper the breakdown points for the variable selection procedures deal strictly with contamination in the form of outlying points in the response variable. Contamination need not be restricted to the response variable and can also be introduced to the data through extreme points in terms of the independent variables. The basic M-estimation approach used to construct the robust regression variable selection procedures is vulnerable to high leverage points. Adjusting for these types of points requires the use of procedures that bound extreme values in both the  $y$  and  $x$  direction. Some of the more popular methods include: Generalized M-estimation and One-Step Generalized M-estimation. Markatou and Hettmansperger (1990) and Markatou and He (1994) develop the standard tests for

these procedures. This should allow for the incorporation of these methods into the variable selection procedures presented in this thesis.

### 5.3 Summary

The three robust variable selection methods with the best simulation results are the  $\text{VAMSI}_2$ ,  $\text{VAMS}_{Bi}$ , and  $\text{RBIC}_{t_3}$ . With similar average false selection rates to the other indicator-based VAMS methods of Chapter 2, the  $\text{VAMSI}_2$  method distinguishes itself with slightly higher model error ratios. The  $\text{VAMS}_{Bi}$  is developed in Chapter 3 using a robust forward addition sequence. While slightly more computationally intensive than the indicator-based VAMS methods, the  $\text{VAMS}_{Bi}$  method does the best job of obtaining the target false selection rate of .05. In contrast to these first two methods the  $\text{RBIC}_{t_3}$  is a prediction-oriented method not designed to control the false selection rate. The  $\text{RBIC}_{t_3}$  method has particularly high model error ratios when the errors follow a Cauchy distribution. The “optimal” variable selection method is dependent on the nature of the data and the goal of the researcher. When the number of observations is significantly larger than the number of candidate predictors the  $\text{VAMSI}_2$  and  $\text{VAMS}_{Bi}$  methods are recommended. These methods are competitive with the  $\text{RBIC}_{t_3}$  procedure in terms of model error but benefit from the ability to control the false selection rate. As the number of candidate predictors approaches the number of observations the appropriate method depends more on the goal of the researcher. If controlling the false selection rate is important, then again the  $\text{VAMSI}_2$  and  $\text{VAMS}_{Bi}$  methods are the only reasonable choice. If the goal is strictly to produce a model for future prediction, the  $\text{RBIC}_{t_3}$  method has an advantage.

Table 5.1: Simulations Results for  $n=100$ ,  $k_T=20$ , and  $k_I=4$ 

Normal Errors							
Method	Ratio of Med. ME	Ratio of Mean ME	CS	FS	Outliers	Size	FSR
VASMI <sub>1</sub>	0.72	0.47	3.68	0.33	0.13	4.01	0.05
VAMSI <sub>2</sub>	0.72	0.47	3.68	0.33	0.09	4.01	0.05
VAMS <sub><math>t_3</math></sub>	0.68	0.45	3.68	0.20	—	3.88	0.04
VAMS <sub><math>Bi</math></sub>	0.77	0.50	3.74	0.23	—	3.97	0.04
RBIC <sub><math>t_3</math></sub>	0.49	0.43	3.85	0.81	—	4.66	0.12
LSA <sub><math>t_3</math></sub>	0.44	0.42	3.87	1.14	—	5.01	0.15
Max SE	0.10	0.04	0.07	0.16			0.02
$t_3$ Errors							
Method	Ratio of Med. ME	Ratio of Mean ME	CS	FS	Outliers	Size	FSR
VASMI <sub>1</sub>	0.47	0.33	3.61	0.45	2.46	4.06	0.08
VAMSI <sub>2</sub>	0.50	0.33	3.61	0.45	2.14	4.06	0.08
VAMS <sub><math>t_3</math></sub>	0.64	0.40	3.68	0.33	—	4.01	0.05
VAMS <sub><math>Bi</math></sub>	0.68	0.36	3.61	0.31	—	3.92	0.05
RBIC <sub><math>t_3</math></sub>	0.47	0.42	3.85	0.77	—	4.62	0.12
LSA <sub><math>t_3</math></sub>	0.40	0.42	3.91	1.16	—	5.07	0.16
Max SE	0.11	0.04	0.10	0.14			0.02
Cauchy							
Method	Ratio of Med. ME	Ratio of Mean ME	CS	FS	Outliers	Size	FSR
VASMI <sub>1</sub>	0.46	0.26	3.78	0.45	18.61	4.23	0.07
VAMSI <sub>2</sub>	0.49	0.27	3.78	0.45	19.04	4.23	0.07
VAMS <sub><math>t_3</math></sub>	0.54	0.27	3.78	0.34	—	4.12	0.06
VAMS <sub><math>Bi</math></sub>	0.46	0.30	3.84	0.34	—	4.18	0.06
RBIC <sub><math>t_3</math></sub>	0.57	0.42	3.93	0.33	—	4.25	0.05
LSA <sub><math>t_3</math></sub>	0.28	0.23	3.90	0.89	—	4.79	0.12
Max SE	0.09	0.05	0.08	0.12			0.02

CS=average # correctly selected, FS=average # falsely selected, Size=CS+FS

Note: Entries are based on 100 replicates

Max SE=maximum standard error across the methods



Table 5.2: Simulations Results for  $n=100$ ,  $k_T=50$ , and  $k_I=10$ 

Normal Errors							
Method	Ratio of Med. ME	Ratio of Mean ME	CS	FS	Outliers	Size	FSR
VASMI <sub>1</sub>	0.21	0.24	4.14	0.96	0.16	5.10	0.14
VAMSI <sub>2</sub>	0.21	0.24	4.14	0.96	0.04	5.10	0.14
VAMS <sub><math>t_3</math></sub>	0.18	0.20	3.12	0.55	—	3.67	0.10
VAMS <sub><math>B_i</math></sub>	0.18	0.20	3.12	0.62	—	3.74	0.11
RBIC <sub><math>t_3</math></sub>	0.23	0.26	6.69	4.05	—	10.74	0.33
LSA <sub><math>t_3</math></sub>	0.30	0.32	6.86	4.41	—	11.27	0.32
Max SE	0.02	0.01	0.19	0.33			0.02
$t_3$ Errors							
Method	Ratio of Med. ME	Ratio of Mean ME	CS	FS	Outliers	Size	FSR
VASMI <sub>1</sub>	0.20	0.21	3.69	0.71	2.08	4.40	0.12
VAMSI <sub>2</sub>	0.20	0.21	3.69	0.71	1.61	4.40	0.12
VAMS <sub><math>t_3</math></sub>	0.19	0.20	3.08	0.46	—	3.54	0.10
VAMS <sub><math>B_i</math></sub>	0.17	0.19	2.87	0.51	—	3.38	0.12
RBIC <sub><math>t_3</math></sub>	0.26	0.28	6.25	2.95	—	9.20	0.27
LSA <sub><math>t_3</math></sub>	0.26	0.27	5.86	3.75	—	9.61	0.32
Max SE	0.02	0.01	0.20	0.29			0.02
Cauchy							
Method	Ratio of Med. ME	Ratio of Mean ME	CS	FS	Outliers	Size	FSR
VASMI <sub>1</sub>	0.10	0.11	3.69	0.95	21.1	4.61	0.15
VAMSI <sub>2</sub>	0.10	0.11	3.69	0.95	11.41	4.61	0.15
VAMS <sub><math>t_3</math></sub>	0.09	0.10	2.81	0.57	—	3.38	0.13
VAMS <sub><math>B_i</math></sub>	0.10	0.11	3.45	0.55	—	4.00	0.10
RBIC <sub><math>t_3</math></sub>	0.14	0.15	5.43	2.02	—	7.45	0.22
LSA <sub><math>t_3</math></sub>	0.11	0.10	4.63	3.23	—	7.86	0.28
Max SE	0.01	0.01	0.23	0.17			0.02

CS=average # correctly selected, FS=average # falsely selected, Size=CS+FS

Note: Entries are based on 100 replicates

Max SE=maximum standard error across the methods

Table 5.3: Simulations Results for  $n=500$ ,  $k_T=50$ , and  $k_I=10$ 

Normal Errors							
Method	Ratio of Med. ME	Ratio of Mean ME	CS	FS	Outliers	Size	FSR
VASMI <sub>1</sub>	0.71	0.72	9.98	0.61	0.03	10.59	0.05
VAMSI <sub>2</sub>	0.71	0.72	9.98	0.61	0.03	10.59	0.05
VAMS <sub><math>t_3</math></sub>	0.60	0.61	9.97	0.60	—	10.57	0.05
VAMS <sub><math>Bi</math></sub>	0.68	0.66	10.00	0.66	—	10.66	0.05
RBIC <sub><math>t_3</math></sub>	0.53	0.53	9.96	1.00	—	10.96	0.08
LSA <sub><math>t_3</math></sub>	0.46	0.49	10.0	1.12	—	11.12	0.09
Max SE	0.05	0.03	0.04	0.12			0.01
$t_3$ Errors							
Method	Ratio of Med. ME	Ratio of Mean ME	CS	FS	Outliers	Size	FSR
VASMI <sub>1</sub>	0.59	0.58	9.94	0.63	14.28	10.57	0.05
VAMSI <sub>2</sub>	0.62	0.57	9.94	0.63	10.84	10.57	0.05
VAMS <sub><math>t_3</math></sub>	0.75	0.71	9.97	0.59	—	10.56	0.05
VAMS <sub><math>Bi</math></sub>	0.75	0.70	9.98	0.49	—	10.47	0.05
RBIC <sub><math>t_3</math></sub>	0.82	0.73	9.97	0.47	—	10.44	0.04
LSA <sub><math>t_3</math></sub>	0.52	0.48	9.98	0.99	—	10.97	0.07
Max SE	0.06	0.04	0.02	0.13			0.01
Cauchy							
Method	Ratio of Med. ME	Ratio of Mean ME	CS	FS	Outliers	Size	FSR
VASMI <sub>1</sub>	0.46	0.46	9.99	0.99	100.14	10.98	0.08
VAMSI <sub>2</sub>	0.49	0.50	9.99	0.99	111.73	10.98	0.08
VAMS <sub><math>t_3</math></sub>	0.52	0.55	10.00	0.66	—	10.66	0.05
VAMS <sub><math>Bi</math></sub>	0.52	0.52	10.00	0.66	—	10.66	0.05
RBIC <sub><math>t_3</math></sub>	0.75	0.71	10.0	0.12	—	10.12	0.01
LSA <sub><math>t_3</math></sub>	0.43	0.43	10.0	0.74	—	10.74	0.06
Max SE	0.06	0.03	0.01	0.12			0.01

CS=average # correctly selected, FS=average # falsely selected, Size=CS+FS

Note: Entries are based on 100 replicates

Max SE=maximum standard error across the methods

# Bibliography

Andrews, D. F., Bickel, P. J., Hampel, F. R., Huber, P. J., Rogers, W. H., and Tukey, J. W. (1972), *Robust Estimates of Location Survey and Advances*, Princeton University Press.

Boos, D. D. (1992), “On Generalized Score Tests,” *The American Statistician*, 46, 327–333.

Boos, D. D., Stefanski, L. A., and Wu, Y. (2008), “Fast FSR Variable Selection with Applications to Clinical Trials,” *Biometrics*.

Dennis, J. and Schnable, R. B. (1983), *Numerical methods for unconstrained optimization and nonlinear equations*, Englewoods Cliffs, N.J.: Prentice-Hall.

Donoho, D. L. and Huber, P. J. (1983), *A Festschrift for Erich L. Lehmann*, Wadsworth.

Fan, J. and Li, R. (2001), “Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties,” *Journal of the American Statistical Association*, 96, 1348–1360.

Hampel, F. R. (1971), “A General Qualitative Definition of Robustness,” *The Annals of Mathematical Statistics*, 42, 1887–1896.

— (1974), “The Influence Curve and Its Role in Robust Estimation,” *Journal of the American Statistical Association*, 69, 383–393.

- (1985), “The breakdown points of the mean combined with some rejection rules,” *Technometrics*, 27, 95–107.
- He, X., Simpson, D. G., and Wang, G. (2000), “Breakdown Points of t-Type Regression Estimators,” *Biometrika*, 87, 675–687.
- Holland, P. and Welsch, R. (1977), “Robust Regression Using Iteratively Reweighted Least-squares,” *Communications in Statistics-Theory and Methods*, 6, 813–827.
- Huber, P. J. (1964), “Robust Estimation of a Location Parameter,” *The Annals of Mathematical Statistics*, 35, 73–101.
- (1973), “Robust Regression: Asymptotics, Conjectures and Monte Carlo,” *The Annals of Statistics*, 1, 799–821.
- (1981), *Robust Statistics*, Wiley.
- Lange, K. L., Little, R. J. A., and Taylor, J. M. G. (1989), “Robust Statistical Modeling Using the t Distribution,” *Journal of the American Statistical Association*, 84, 881–895.
- Luo, X., Stefanski, L. A., and Boos, D. D. (2006), “Tuning Variable Selection Procedures by Adding Noise,” *Technometrics*, 48, 165–175.
- Mallows, C. (1973), “Some Comments on Cp,” *Technometrics*, 15, 661–671.
- Markatou, M. and He, X. (1994), “Bounded-Influence and High Breakdown Point Testing Procedures in Linear Models,” *Journal of the American Statistical Association*, 89, 543–549.
- Markatou, M. and Hettmansperger, T. P. (1990), “Robust Bounded-Influence Tests in Linear Models,” *Journal of the American Statistical Association*, 85, 187–190.
- Maronna, R. A. (2006), *Robust Statistics Theory and Methods*, John Wiley and Sons.

- Mickey, M. R., Dunn, O. J., and Clark, V. (1967), “Note on the use of stepwise regression in detecting outliers,” *Computers and Biomedical Research*, 1, 105–111.
- Rousseeuw, P. J. (1984), “Least Median of Squares Regression,” *Journal of the American Statistical Association*, 79, 871–8810.
- Schwarz, G. (1978), “Estimating the Dimension of a Model,” *The Annals of Statistics*, 6, 461–464.
- Shao, J. (1993), “Linear Model Selection by Cross-Validation,” *Journal of the American Statistical Association*, 88, 486–494.
- Simpson, D. G., Ruppert, D., and Carroll, R. J. (1992), “On One-Step GM Estimates and Stability of Inferences in Linear Regression,” *Journal of the American Statistical Association*, 87, 439–450.
- Street, J. O., Carroll, R. J., and Ruppert, D. (1988), “A Note on Computing Robust Regression Estimates Via Iteratively Reweighted Least Squares,” *The American Statistician*, 42, 152–154.
- Tibshirani, R. (1996), “Regression Shrinkage and Selection via the Lasso,” *Journal of the Royal Statistical Society*, 58, 267–288.
- Wang, H. and Leng, C. (2007), “Unified LASSO Estimation by Least Squares Approximation,” *Journal of the American Statistical Association*, 102, 1039–1048.
- Wang, H., Li, G., and Jiang, G. (2007), “Robust Regression Shrinkage and Consistent Variable Selection Through the LAD-Lasso,” *Journal of Business and Economic Statistics*, 25, 347–355.
- Wu, Y., Boos, D. D., and Stefanski, L. A. (2007), “Controlling Variable Selection by the Addition of Pseudovariables,” *Journal of the American Statistical Association*, 102, 235–243.

- Yohai, V. J. (1987), “High Breakdown-Point and High Efficiency Robust Estimates for Regression,” *The Annals of Statistics*, 15, 642–656.
- Zhang, P. (1993), “Model Selection Via Multifold Cross Validation,” *The Annals of Statistics*, 21, 299–313.
- Zou, H. (2006), “The Adaptive Lasso and Its Oracle Properties,” *Journal of the American Statistical Association*, 101, 1418–1429.