Essays on Robust Model Selection and Model Averaging for Linear Models

Le Chang

May 2017

A thesis submitted for the degree of Doctor of Philosophy

The Australian National University

Australian National University

For my family

Declaration

I hereby declare the work in this thesis is my own except where otherwise stated. Any contribution made to the research by others, with whom I have worked at the Australian National University (ANU) or elsewhere, is explicitly acknowledged in the thesis.

Le Chang

Acknowledgements

This thesis emerges from my Ph.D program, funded by the Research School of Finance, Actuarial Studies and Statistics (RSFAS) at the ANU. During my three year Ph.D program, I have been helped by many individuals and institutions. I would like to express my gratitude to those who made this thesis possible.

First of all, I would like to take this opportunity to thank all of my supervisors, Professor Steven Roberts, Professor Alan Welsh, and Dr Yanrong Yang, whose continuous assistance, guidance and encouragement throughout this period have been paramount to the completion of this thesis. I particularly thank Professor Steven Roberts for his continuous guidance throughout all the stages of my Ph.D study. His enthusiastic support has made my Ph.D experience productive and stimulating. I sincerely thank Professor Alan Welsh for providing his expertise in many fields of statistics. His wide knowledge and expert thinking have been of great value to me. I also gratefully acknowledge Dr Yanrong Yang for her timely feedback and kind discussions on the last chapter of my thesis. Her endless patience in supervising me to pursue the excellence has been crucial to the completion of my thesis.

I would like to express special thanks to the RSFAS for providing me with an excellent academic environment and generous funding throughout my Ph.D program. In particular, I would like to thank Associate Professor Timothy Higgins, the Ph.D Convenor in Statistics and the Director in Higher Degree Research of RSFAS, for providing me with considerable assistance in Ph.D-related issues. Moreover, I am grateful to Associate Professor Stephen Sault and Ms Tracy Skin-

ner for their effort to arrange my tutorials.

Doctoral research at the ANU has been a memorable experience for me, and I would like to thank the exceptional RSFAS faculty staff for their academic support and general help. Additionally, thanks to my fellow PhD students of RSFAS, my PhD study has been enjoyable and fruitful.

Finally, I owe a great debt to my mother Yuxiang Li and father Zhaoping Chang. Thank you for your endless love and support. I especially thank Hailun Zhou, who provided me with unconditional companionship, patience and moral support throughout my Ph.D. This work is dedicated to my family.

This research is supported by the Australian Government Research Training Program.

Abstract

Model selection is central to all applied statistical work. Selecting the variables for use in a regression model is one important example of model selection. This thesis is a collection of essays on robust model selection procedures and model averaging for linear regression models.

In the first essay, we propose robust Akaike information criteria (AIC) for MM-estimation and an adjusted robust scale based AIC for M and MM-estimation. Our proposed model selection criteria can maintain their robust properties in the presence of a high proportion of outliers and the outliers in the covariates. We compare our proposed criteria with other robust model selection criteria discussed in previous literature. Our simulation studies demonstrate a significant outperformance of robust AIC based on MM-estimation in the presence of outliers in the covariates. The real data example also shows a better performance of robust AIC based on MM-estimation.

The second essay focuses on robust versions of the "Least Absolute Shrinkage and Selection Operator" (lasso). The adaptive lasso is a method for performing simultaneous parameter estimation and variable selection. The adaptive weights used in its penalty term mean that the adaptive lasso achieves the oracle property. In this essay, we propose an extension of the adaptive lasso named the Tukey-lasso. By using Tukey's biweight criterion, instead of squared loss, the Tukey-lasso is resistant to outliers in both the response and covariates. Importantly, we demonstrate that the Tukey-lasso also enjoys the oracle property. A fast accelerated proximal gradient (APG) algorithm is proposed and implemented for

computing the Tukey-lasso. Our extensive simulations show that the Tukey-lasso, implemented with the APG algorithm, achieves very reliable results, including for high-dimensional data where $p > n$. In the presence of outliers, the Tukey-lasso is shown to offer substantial improvements in performance compared to the adaptive lasso and other robust implementations of the lasso. Real data examples further demonstrate the utility of the Tukey-lasso.

In many statistical analyses, a single model is used for statistical inference, ignoring the process that leads to the model being selected. To account for this model uncertainty, many model averaging procedures have been proposed. In the last essay, we propose an extension of a bootstrap model averaging approach, called bootstrap lasso averaging (BLA). BLA utilizes the lasso for model selection. This is in contrast to other forms of bootstrap model averaging that use AIC or Bayesian information criteria (BIC). The use of the lasso improves the computation speed and allows BLA to be applied even when the number of variables p is larger than the sample size n. Extensive simulations confirm that BLA has outstanding finite sample performance, in terms of both variable and prediction accuracies, compared with traditional model selection and model averaging methods. Several real data examples further demonstrate an improved out-of-sample predictive performance of BLA.

Contents

List of Abbreviations

AIC	Akaike information criteria
APG	Accelerated proximal gradient
BIC	Bayesian information criteria
BLA	Bootstrap lasso averaging
BMA	Bayesian model averaging
FMA	Frequentist model averaging
FR	Forward regression
IRLS	Iteratively reweighted least squares
LAD	Least absolute deviation
LASSO	Least Absolute Shrinkage and Selection Operator
LMS	Least median of squares
LQA	Local quadratic approximation
LTM	Least trimmed sum of squares
MAD	Median absolute deviation
MSPE	Mean squared prediction error
NIR	Near-infrared

OLS	Ordinary least squares
SIS	Sure independence screening
TMSPE	Trimmed mean square prediction error

Chapter 1

Introduction

Model selection is central to all applied statistical work. Selecting the variables for use in a regression model is one important example. Over the past two decades, a number of different model selection approaches have been rapidly developed and there exists a substantial literature that addresses the issue of methods for model selection.

Stepwise procedures (sequential testing), allowing variables to be added or deleted at each step, have often been employed. However, such testing schemes based on p values only compare two nested models and have been widely criticized since hypothesis tests generally form a very poor basis for model selection (Akaike, 1974). Cross-validation and its variations have been suggested and discussed as useful model selection methods (Mosteller and Tukey, 1968; Shao, 1993). However, these methods are quite computer intensive and tend to be impractical if a large number of models need to be evaluated (Burnham and Anderson, 2004). The adjusted coefficient (Draper and Smith, 1981) and Mallow's Cp statistic (Mallows, 1973) are also widely used in least square regression and provide a ranking of all candidate models.

The general approach that is the focus of the first essay is the model selection methods that choose models by minimizing an expression (criterion) that can be written as a loss term in addition to a penalty term. More specifically, these model

selection criteria, such as the Akaike information critera (AIC) (Akaike, 1974), the Bayesian information criteria (BIC) (Hoeting et al., 1999) and their variations will be considered. Most of these prevalent model selection criteria are based on the squared loss, yet it is well known that the commonly used squared loss function is very sensitive to outliers and other violations to the normality assumption for error distribution. A growing body of literature is concerned with the model selection procedures for linear models that are less sensitive to outliers: a robust Cp (Ronchetti and Staudte, 1994), a robust version of cross-validation (Ronchetti et al., 1997), and weighted versions of likelihood estimators (Agostinelli, 2002). Ronchetti (1985) proposes a robust version of AIC for M-estimation by replacing the squared loss with Huber's function. However, Huber's loss can only be robust to outliers in the response values. Outliers in the covariates also appear frequently and they generally have a greater effect on the accuracy of the regression estimates than the outliers in the response. In the first essay, we propose to replace the loss function by Tukey's biweight criterion and develop a robust AIC based on MM-estimation that copes with outliers in both the response and the covariates.

Selecting the best model using these model selection criteria eases the interpretation of the model and generally improves the prediction accuracy. However, results can be extremely variable as evaluation is a discrete process and more importantly, it is computationally expensive and tends to be impractical if a large number of models need to be evaluated. Another technique to improve prediction accuracy is ridge regression. Hoerl and Kennard (1970) proposed ridge regression by adding an $L2$ penalty, not only to sacrifice a little bias of the estimates, but to simultaneously shrink those estimates and reduce their variance. However, with a large number of predictors, data analysts would like to determine a smaller subset of predictors that show the strongest effects. Ridge regression does not set any coefficients to exactly zero and thus does not give an 'easily interpretable' model. Consequently, Tibshirani (1996) proposed a new technique called the LASSO or the 'Least Absolute Shrinkage and Selection Operator', which modifies the $L2$

penalty into an $L1$ penalty. The lasso not only shrinks some coefficients, achieving a better prediction, but also sets others to zero and hence offers parsimonious solutions to ease the interpretation.

Over the last few decades, the lasso has become a very popular technique for simultaneous estimation and variable selection. A significant volume of literature further investigated the properties of the lasso estimates and developed different versions of the lasso. Zou (2006) stated that there exist certain scenarios where the lasso is inconsistent for variable selection and therefore he suggests the adaptive lasso where adaptive weights are used for penalizing different coefficients in the $L1$ penalty. The adaptive lasso enjoys the oracle properties as it performs as well as the true underlying model asymptotically. However, as mentioned previously, datasets with outliers are commonly encountered in statistical analysis. These outliers may appear in the response and/or the predictors. The lasso estimates, which utilize ordinary least squares (OLS), also suffer from the effect of outliers. Some authors have considered robust versions of the lasso, generally utilizing penalized versions of M-estimators, as in Owen (2007), Wang et al. (2007), Li et al. (2011), and in Lambert-Lacroix and Zwald (2011). Wang et al. (2007) proposed to overcome the presence of outliers by combining the least absolute deviation (LAD) loss with the lasso penalty. Unfortunately, it is well known that the LAD loss is not adaptable for small errors because it strongly penalizes small residuals (Owen, 2007; Lambert-Lacroix and Zwald, 2011). In other words, the LAD-Lasso has lower efficiency than OLS estimates when there are no outliers in the response. Owen (2007) and Lambert-Lacroix and Zwald (2011) preferred to replace the squared loss with Huber's loss, a hybrid of the squared error and absolute error loss functions.

Although the robust lasso with the Huber's loss is resistant to outliers in the response and achieves high asymptotic efficiency, it is not robust against high leverage points or outliers in the covariates. In the second paper, we propose replacing the squared loss in the lasso with Tukey's biweight criterion, and name

the method the Tukey-lasso for handling outliers in the response and covariates. In our simulation study, we show that the Tukey-lasso outperforms the adaptive lasso and other robust implementations of the lasso, particularly in the presence of outliers in both the response and the predictors. We further propose an accelerated proximal gradient (APG) algorithm to compute the Tukey-lasso. The APG computes the lasso minimization problem and guarantees a global minimizer for a convex objective function. Although the objective function for the robust lasso with Tukey's biweight is non-convex, the APG algorithm still achieves very reliable results (a local minimizer) when the starting value of the algorithm is carefully selected.

In the first two papers, the model selection approaches we discuss assume that the identity and parameter values of that best model can be estimated, and that, thereafter, inferences will be made from the data only according to the sole and best model. However, for any given data set, the use of a different model selection method may result in a different best model being selected. Conversely, for any given model selection approach, a different best model would likely be chosen if a replicate data set were analyzed. Often, several models fit the data equally well, yet these models may include different explanatory variables and lead to different predictions. This extra component of variation, is often termed 'model uncertainty'. To account for model uncertainty, model averaging, which makes inferences based on weighted support from several models instead of a sole best model, has been proposed and developed. Apart from avoiding the inference drawn from the single best model, model averaging has been shown to improve predictive performance more accurately than reliance on a single model (Raftery et al., 1997). An increasing amount of literature is concerned with the implementation of model averaging including Bayesian model averaging (BMA) (Hoeting et al., 1999; Clyde and George, 2004; Raftery et al., 1997) and frequentist model averaging (FMA) (Rao and Tibshirani, 1997; Hjort and Claeskens, 2003; Burnham and Anderson, 2003; Yuan and Yang, 2012; Claeskens et al., 2008).

Among these papers contributing to model averaging, the most relevant work for our third essay is bootstrap model averaging first proposed by Buckland et al. (1997). Bootstrap model averaging in Buckland et al. (1997) utilizes the bootstrap to generate resamples, applies the model selection criteria independently to each resample and further computes the weights assigned to each model. However, similar to the traditional model selection procedures by AIC or BIC, bootstrap model averaging is computationally intensive with a large number of variables and computationally infeasible with the number of variables greater than the sample size. In this third paper, we modify bootstrap model averaging by utilizing the lasso (Tibshirani, 1996) as a model selection tool, instead of the traditional AIC or BIC, to improve the computation speed and realize the computational feasibility even when the number of variables p is larger than the sample size n. We call this modified version of bootstrap model averaging, 'bootstrap lasso averaging'.

The rest of this thesis is structured as follows. In Chapter 2, we modify the AIC by replacing the loss function with Tukey's biweight criterion and we develop a robust AIC based on MM-estimation that copes with outliers in both the response and the covariates. In Chapter 3, we propose replacing the squared loss in the lasso with Tukey's biweight criterion, and name the method the Tukey-lasso for handling outliers in the response and covariates. Additionally, we further propose an APG algorithm to compute the Tukey-lasso. In Chapter 4, we modify bootstrap model averaging by utilizing the lasso (Tibshirani, 1996) as a model selection tool, instead of the traditional AIC or BIC; we call this 'bootstrap lasso averaging'. Finally, we present brief conclusions and future research directions in Chapter 5.

Chapter 2

A Comparison of Robust Model Selection Criteria Based on M and MM-estimators

2.1 Introduction

It is well known that the commonly used least square estimators in the linear regression setting are very sensitive to outliers and other violations to the normality assumption for error distribution. Various types of robust estimators have been introduced and discussed, such as M-estimator, S-estimator, and MM-estimator (Huber, 2011; Hampel, 1971; Yohai, 1987). However, the presence of outliers not only affects these estimators, but also (and more severely) the model selection procedures, especially these likelihood based criteria (AIC, BIC, and Mallow's Cp) (Ronchetti et al., 1997). An increasing volume of literature is concerned with the model selection procedures for linear models that are less sensitive to outliers: a robust Cp (Ronchetti and Staudte, 1994), a robust version of cross-validation (Ronchetti et al., 1997), and weighted versions of likelihood estimators (Agostinelli, 2002). Ronchetti (1985) proposed a robust version of AIC for M-estimation by replacing the squared loss with a general function ρ (e.g. Huber's

function). However, M-estimation and Huber's loss can only be robust for outliers in the response values. Outliers in the covariates also appear frequently and they generally have a greater effect on the accuracy of the regression estimates than the outliers in the response. To be robust against outliers in both the covariates and the response, the derivative of the loss function needs to be redescending (Rousseeuw, 1984; Yohai, 1987). A commonly used loss function with this property is Tukey's biweight (Tukey, 1960). In this work, we propose to replace the loss function by Tukey's biweight criterion and develop a robust AIC based on MM-estimation that copes with outliers both in the response and the covariates.

Some robust model selection criteria based on MM-estimation have been proposed in previous literature. Tharmaratnam and Claeskens (2013) developed a robust scale based AIC for M and MM-estimation and showed that it performed well compared with the classical AIC in terms of the probability of selecting the correct model. However, the trace term of their model selection criterion behaved quite abnormally, especially when the number of covariates is large. Therefore, we propose to adjust the trace term following Ronchetti (1985) and our simulation study confirms a significant improvement using the adjusted trace term. Müller and Welsh (2005) make use of stratified bootstrap and MM-estimation to combine a robust penalized criterion with a robust conditional expected prediction loss. They also find a consistently better selection probability of their robust model selection criteria in comparison with the traditional criteria based on squared error loss. However, the computation of their robust model selection criteria is intensive as a consequence of the bootstrapping. The emphasis in previous papers has mostly been on the performance of selection probabilities for robust model selection criteria in comparison with classical ones. We will further investigate whether improvements in terms of prediction can be achieved when we select models according to these robust model selection criteria. In this chapter, we only focus on the case of $n > p$ since the computation of these model selection criteria is not feasible when $n < p$.

The purpose of this chapter is to investigate and compare the selection probabilities for several robust model selection criteria, and to ascertain, the prediction ability of the best model selected by these criteria.

2.2 Robust estimation

2.2.1 Linear regression model

Linear regression is the most commonly used approach to model the relationship between a dependent variable and one or more independent variables. We consider the linear regression model,

$$y_i = X_i^T \boldsymbol{\beta} + \epsilon_i, \quad i = 1, 2, \ldots, n, \tag{2.1}$$

where y_i is the response variable on the i^{th} observation, $X_i = (\mathbf{1}, x_{i1}, x_{i2}, \ldots, x_{ip})^T$ are the values of the covariates for the i^{th} observation, p is the number of covariates, $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \ldots, \beta_p)^T$ are coefficient parameters, and ϵ_i, $i = 1, 2, \ldots, n$, is an independently normally distributed random variable with mean zero and variance σ^2.

The estimates of $\boldsymbol{\beta}$ are usually obtained by the method of ordinary least squares (OLS). The OLS estimate is the solution to the problem:

$$\widehat{\boldsymbol{\beta}}_{LS} = \text{argmin} \sum_{i=1}^{n} \left(y_i - X_i^T \boldsymbol{\beta}\right)^2. \tag{2.2}$$

Unfortunately, the use of the OLS method would be inappropriate for use in a problem containing outliers or extreme observations. When there are outliers in the data, the summation part of the above minimization problem is dominated by the residual squares of these extreme observations. In such a situation, the OLS estimators often perform very poorly.

Robust regression methods are designed not to be overly affected by the presence of outliers or the violations of error assumptions. This method is an im-

portant tool for analyzing data that are heavily affected by the outliers and it provides results that are resistant to outliers. Some of the well-known robust regression methods are M-estimation, S-estimation, and MM-estimation.

2.2.2 M-estimation

M-estimation is the most general method of robust regression, introduced by Huber et al. (1964). The letter 'M' indicates that it is an estimator of the maximum likelihood type. If we still consider the linear model as described in (2.1), the M-estimator minimizes the objective function,

$$\widehat{\boldsymbol{\beta}}_M = \text{argmin} \sum_{i=1}^{n} \rho \left(\frac{y_i - X_i^T \boldsymbol{\beta}}{\sigma} \right). \tag{2.3}$$

Compared with the OLS method, the squared residual function is now replaced by another function, which is a symmetric, non-decreasing in $[0, +\infty)$, and with $\rho(0) = 0$. Moreover, to be more robust against outliers that result in large residuals, this ρ function should increase less for large values of residuals than the squared function in the OLS. An optimal choice of ρ is provided by Huber's family with a loss function,

$$\rho_c(u) = \begin{cases} \frac{u^2}{2} & \text{if } |u| \leq c \\ c|u| - \frac{c^2}{2} & \text{otherwise,} \end{cases} \tag{2.4}$$

where c is a tuning constant that controls the level of robustness and a standard choice of c is $c = 1.345$ for 95% asymptotic efficiency in standard normal distribution.

To minimize the objective function as in (2.3), we differentiate it with respect to the coefficients $\boldsymbol{\beta}$, set the partial derivatives to 0, and obtain a system of estimating equations for the coefficients,

$$\sum_{i=1}^{n} \psi_c \left(\frac{y_i - X_i^T \boldsymbol{\beta}}{\sigma} \right) X_i^T = \mathbf{0} \tag{2.5}$$

where ψ_c is the derivative of ρ_c. Further, σ could be estimated by the median absolute deviation (MAD) of the residuals, MAD $= 1.4826 \times \text{Median}_{i=1,\ldots,n}(|r_i|)$, or by Huber's proposal 2 (Huber, 2011).

Solving the estimation equations is a weighted least squares problem and in most cases, the iteratively reweighted least squares (IRLS) algorithm could be performed, which is typically the preferred method. In this work, we estimate the regression parameters using the '*rlm()*' function from the 'MASS' package in R using method of 'M'.

2.2.3 S-estimation

Although M-estimation could be resistant to outliers in response values and achieve high asymptotic efficiency, M-estimators are not robust against high leverage points (outliers in covariates), and more importantly, have a disappointing breakdown property. The breakdown point of an estimator is the proportion of 'bad' data that can be arbitrarily large values without making the estimator arbitrarily bad.

The first estimators with high breakdown points are the least median of squares (LMS) and the least trimmed sum of squares (LTM). A more general class of high breakdown estimators that likewise do not suffer from leverage points is introduced by Rousseeuw and Yohai (1984), and is known as the class of scale-type or S-estimators (Salibian-Barrera and Yohai, 2006).

The S-estimator is defined by $\widehat{\boldsymbol{\beta}}_S = \text{argmin}\ \widehat{\sigma}(\boldsymbol{\beta})$ with a determining minimum robust scale estimator and the scale function satisfies the equation,

$$\frac{1}{n}\sum_{i=1}^{n}\rho\left(\frac{y_i - X_i^T\boldsymbol{\beta}}{\widehat{\sigma}(\boldsymbol{\beta})}\right) = K, \quad \text{where } K = \int \rho(z)\phi(z)dz. \tag{2.6}$$

A commonly used family of loss functions ρ is given by the Tukey's biweight,

$$\rho_d(u) = \begin{cases} \frac{d^2}{6}\left\{1 - \left[1 - \left(\frac{u}{d}\right)^2\right]^3\right\} & \text{if } |u| \leq d \\ \frac{d^2}{6} & \text{if } |u| \geq d \end{cases} \tag{2.7}$$

The solution to the S-estimator could also be found by using an IRLS method. The choice of $d = 1.548$ yields $K = 0.5$, indicating that the S-estimator achieves a 50% breakdown point. The S-estimator has much higher breakdown point than the M-estimator when its ψ function redescends, where ψ is the derivative of ρ. However, a high breakdown point generally results in a low efficiency. When the S-estimator described above obtains a 50% breakdown point, it only achieves 28.7% efficiency at the core model with a standard normal distribution.

2.2.4 MM-estimation

It is possible to have both high breakdown point and high efficiency. This is achieved by MM-estimation, a three-stage procedure introduced by Yohai (1987). In the first stage, an initial regression estimate (e.g. LMS) is computed with high breakdown point but is not necessarily efficient. In the second stage, an M-estimation of the errors scale is calculated based on the initial estimate (the first M). In the third stage, an M-estimate of regression coefficients is computed based on the scale estimate obtained in the second stage (the second M). Briefly speaking, the MM-estimator is computed as an M-estimator starting at the coefficients provided by a high breakdown S-estimator and using the fixed scale afforded by the S-estimator. As mentioned by one of the referees, another robust estimator, τ-estimator (Yohai and Zamar, 1988), also combines good robustness and high efficiency. However, we only focus on a more widely used MM-estimator in this chapter.

If we define ρ_{d0} to be the ρ function as in the S-estimation procedure and ρ_{d1} to be the ρ function as in the M-estimation of the third stage, ρ_{d1} should satisfy $\rho_{d1}(u) \leq \rho_{d0}(u)$ for all $u \in R$ and $\sup \rho_{d1}(u) = \sup \rho_{d0}(u)$. MM-estimator is then the solution of

$$\widehat{\boldsymbol{\beta}}_{MM} = \text{argmin} \sum_{i=1}^{n} \rho_{d1} \left(\frac{y_i - X_i^T \boldsymbol{\beta}}{\widehat{\sigma}_s} \right), \tag{2.8}$$

where ρ_{d1} is still the Tukey's biweight function mentioned in S-estimation. However, a larger value of d is chosen and $d = 4.685$ is a typical choice for d. Further, $\widehat{\sigma}_s$ is the S-scale estimator, which was derived in the S-estimation procedure by using ρ_{d0} as the ρ function.

Again, to solve the above equation, we could differentiate the objective function with respect to the coefficients $\boldsymbol{\beta}$ and set the partial derivatives to $\mathbf{0}$, obtaining a system of estimating equations for the coefficients,

$$\sum_{i=1}^{n} \psi_{d1} \left(\frac{y_i - X_i^T \boldsymbol{\beta}}{\widehat{\sigma}_s} \right) X_i^T = \mathbf{0} \tag{2.9}$$

where ψ_{d1} is the derivative of ρ_{d1} and an IRLS algorithm could be applied to find the solution to this equation. With $d = 4.685$, the MM-estimator achieves 95% efficiency at standard normal distribution and the 50% breakdown point. Therefore, it is a superior estimator to either the M-estimator or the S-estimator, in terms of breakdown point and efficiency. Here, we estimate the MM-regression parameters using the '*rlm()*' function from the 'MASS' package in R using method of 'MM'.

2.3 Robust model selection criteria

Model selection is a key component of all statistical work with data. Selecting the variables for use in a regression model is one important example. Using all variables in the model suffers from high variability in parameter estimation, and thus, results in a very poor prediction accuracy. Over last few decades, various model selection criteria have been rapidly developed. Among these model selection criteria, AIC (Akaike, 1974) is increasingly used for all statistical analysis.

However, likelihood based criteria, such as AIC, are highly sensitive to out-

liers or to other departures from normality assumptions in the error distribution. Therefore, model selection procedures require special care in the presence of outliers. A growing number of papers are concerned with model selection procedures for linear models that are less sensitive to outliers: Ronchetti (1985); Ronchetti and Staudte (1994); Ronchetti et al. (1997); Agostinelli (2002); Tharmaratnam and Claeskens (2013); Müller and Welsh (2005). In this section, we will introduce the six different types of model selection criteria considered in our analysis and comparison.

2.3.1 Classical AIC

AIC (Akaike, 1974), a variant of the Kullback-Leibler divergence between the true model and the approximating candidate model, has been widely used as a model selection tool over the past decades. According to Ronchetti (1985), if we consider a linear regression model in Section 2.2 and assume that the errors follow some distribution with density g, a generalized AIC proposed by Bhansali and Downham (1977) for a given fixed α is

$$AIC(p;\alpha) = -2\sum_{i=1}^{n} \log g\left(\frac{y_i - X_i^T\widehat{\boldsymbol{\beta}}}{\widehat{\sigma}}\right) + \alpha p + K(n,\widehat{\sigma}), \tag{2.10}$$

where $K(n,\widehat{\sigma})$ is a function of n and $\widehat{\sigma}$, $\widehat{\sigma}$ is an estimate of σ, and p is the number of parameters in the linear regression model. When g is a standard normal distribution and the choice of α is $\alpha = 2$, $AIC(p;2)$ is reduced to the well-known criterion, Mallows'C_p, proposed by Mallows (1973),

$$C_p = \sum_{i=1}^{n}\left(\frac{y_i - X_i^T\widehat{\boldsymbol{\beta}}}{\widehat{\sigma}}\right)^2 - n + 2p. \tag{2.11}$$

Further, when $\widehat{\sigma}$ is the maximum likelihood estimate of σ, this criterion reduces to the traditional AIC for linear regression,

$$AIC = 2n\log\widehat{\sigma}_{ML} + 2p \tag{2.12}$$

where $\widehat{\sigma}_{ML}$ is the maximum likelihood estimate of σ. Further, it is widely known that C_p is asymptotically equivalent to the traditional AIC. However, it could be noted that AIC and its variant, C_p are based on the normality assumption of error distribution and they are sensitive to outlying observations. In the presence of outliers, $\sum_{i=1}^{n}\left(\frac{y_i - X_i^T\widehat{\boldsymbol{\beta}}}{\widehat{\sigma}}\right)^2$ is dominated by the residuals of these extreme observations. In this situation, AIC or C_p performs poorly. Therefore, we search for more robust alternatives.

2.3.2 Robust AIC for M-estimation

Ronchetti (1985) proposed a robust version of AIC for M-estimation by replacing $\log g$ with a more general function ρ,

$$RAIC.M = 2\sum_{i=1}^{n}\rho\left(\frac{y_i - X_i^T\widehat{\boldsymbol{\beta}}_M}{\widehat{\sigma}}\right) + \alpha p, \tag{2.13}$$

where in this case, $\widehat{\boldsymbol{\beta}}_M$ is the M-estimator as in Section 2.2, $\widehat{\sigma}$ is a robust estimate of σ and does not change from model to model. The model with the smallest value of $RAIC.M$ is selected as the best model. As discussed in Ronchetti (1985), the extension of AIC to a more robust version of itself is the exact counterpart of the maximum likelihood estimation for M-estimation.

Another issue we address here is the choice of parameter α. Following the results of Stone (1977), the penalty term $\alpha p = 2\ \text{trace}(M^{-1}Q)$, where

$$M = -E\left[\frac{\partial\psi}{\partial\boldsymbol{\beta}}\right] = E\left[\psi' xx^T\right], \qquad Q = E\left[\psi\psi^2\right] = E\left[\psi^2 xx^T\right].$$

Here, ψ is the derivative of ρ with respect to coefficient parameters. Ronchetti (1985) suggests that $M = E\left[\psi' xx^T\right] = E[\psi'] \cdot E[xx^T]$ and $Q = E\left[\psi^2 xx^T\right] = E[\psi^2] \cdot E[xx^T]$. Thus, the penalty term is

$$\alpha p = 2\ \text{trace}(M^{-1}Q) = 2\frac{E[\psi^2]}{E[\psi']}p. \tag{2.14}$$

Here, the choice of α is $2\frac{E[\psi^2]}{E[\psi']}$. This α is then fixed across all the possible models. When the ρ function is simply a square function (as in the case of OLS), αp is exactly equal to $2p$ and the selection criterion $RAIC.M$ simply reduces to the traditional AIC (or its variant, C_p).

Therefore, in this work, we compute robust version of AIC for M-estimation suggested by Ronchetti (1985) as,

$$RAIC.M = 2\sum_{i=1}^{n} \rho_c \left(\frac{y_i - X_i^T \widehat{\boldsymbol{\beta}}_M}{\widehat{\sigma}} \right) + 2\frac{E[\psi_c^2]}{E[\psi_c']} p \tag{2.15}$$

where ρ_c is the Huber's function, $\widehat{\boldsymbol{\beta}}_M$ is the M-estimator, $\widehat{\sigma}$ is the median absolute deviation (MAD) of the residuals from the full model, and $E[\psi_c^2]$ and $E[\psi_c']$ are estimated respectively by the average of the empirical values of ψ_c^2 and ψ_c' from the full model.

2.3.3 Robust AIC for MM-estimation

It is widely known that M-estimators have a low breakdown point. Since the robust version of AIC based on M-estimation applies the same ρ function (the Huber function) as the M-estimators, it loses its robust property in the presence of a high proportion of outliers. More importantly, in regression analysis, M-estimation is only robust to the outliers in the response but not the outliers in the covariates.

Yohai (1987) proposed an MM-estimation, that utilizes the S-scale estimator in an M-estimation equation and is robust to outliers in both the response and the covariates. A corresponding robust version of AIC could also be further derived based on MM-estimation. Therefore, we propose a new robust version of AIC based on MM-estimator, which obtains a higher breakdown point than the M-estimator and copes with outliers in both the response and the covariates,

$$RAIC.MM = 2\sum_{i=1}^{n} \rho_d \left(\frac{y_i - X_i^T \widehat{\boldsymbol{\beta}}_{MM}}{\widehat{\sigma}} \right) + 2\frac{E[\psi_d^2]}{E[\psi_d']} p, \tag{2.16}$$

where ρ_d is the Tukey's biweight function, and $\widehat{\sigma}$ is the MAD of the residuals of the full model by MM-estimation. The structure of $RAIC.MM$ is similar to $RAIC.M$, as suggested by Ronchetti (1985). However, the loss function is replaced by the Tukey's biweight function (ρ_d) and high breakdown estimators, MM-estimators ($\widehat{\boldsymbol{\beta}}_{MM}$) are used instead of M-estimators. In our simulation study, we find that $RAIC.MM$ performs significantly better than $RAIC.M$ when the proportion of outliers in the response reaches 20% or when outliers are present in the covariates.

2.3.4 Robust AIC with a prediction loss part

Robust versions of AIC for both M-estimation and MM-estimation involve specifying estimators and computing the required model selection criteria based on these estimators. Additionally, the ρ functions in robust versions of model selection criteria are in the same class as the ρ functions in the estimation procedure.

Müller and Welsh (2005) broaden the usual approach to robust model selection by separating the ρ function during the estimation and the model selection. They further state that a useful linear regression model should also be able to predict independent new observations. Therefore, they propose to add a conditional (given the sample) expected prediction loss part to the penalized loss function (the traditional structure of a model selection criterion). The traditional way to estimate the prediction loss part is to utilize the bootstrap method. To ensure that outliers or observations in the extreme tails were present in each bootstrap sample, Müller and Welsh (2005) used an m out of n stratified bootstrap (see Müller and Welsh (2005)) when estimating the conditional expected prediction loss. Hence, Müller and Welsh (2005) constructed the robust model selection criteria as,

$$Mn = \sum_{i=1}^{n} \rho \left(\frac{y_i - X_i^T \widehat{\boldsymbol{\beta}}_c}{\widehat{\sigma}_c} \right) + 2p + E_* \sum_{i=1}^{n} \rho \left(\frac{y_i - X_i^T \widehat{\boldsymbol{\beta}}_c^{m*}}{\widehat{\sigma}^c} \right), \tag{2.17}$$

where $\widehat{\boldsymbol{\beta}}_c$ denotes an estimator of type c of $\boldsymbol{\beta}$ (e.g. OLS-type, M-type, MM-type), $\widehat{\sigma}_c$ is the MAD of residuals from a full model by type c estimation, E_* defines expectation with respect to the bootstrap distribution and $\widehat{\boldsymbol{\beta}}_c^{m*}$ are estimators of $\boldsymbol{\beta}$ for m out of n stratified bootstrap samples, and ρ is a bounded function in the form of

$$\rho(u) = min(u^2, b^2).$$

A reasonable choice of the constant b mentioned in Muller and Welsh (2005) is 2.

The penalized loss term in the above criterion is similar in conception to the robust versions of AIC for M-estimation and MM-estimation. However, as discussed in Müller and Welsh (2005), the choice of the ρ function intentionally does not correspond to any commonly used estimators, and thus, such a criterion could be used to compare different estimators. Linking the criterion to any estimators (using the same ρ function in the selection criterion and estimation) may excessively favor the selected estimator (Müller and Welsh, 2005). However, such a separation yields a different penalty term αp to those revealed in the traditional information criteria (e.g. AIC or robust version of AIC). Müller and Welsh (2005) simply chose a penalty term that does not depend on the choice of a ρ function (e.g. $log(n)p$). To be more comparable with the robust version of AIC, we consider $2p$ (the penalty term in AIC) as a penalty term for the selection criterion here.

In our simulation study, we choose the bounded function ρ as in the form of $\rho(u) = min(u^2, b^2)$ mentioned in Müller and Welsh (2005), and where $b = 2$. We consider M_n for both M-type and MM-type estimators, denoting them by $Mn.M$ and $Mn.MM$ respectively.

2.3.5 Robust scale based AIC for M and MM-estimation

Tharmaratnam and Claeskens (2013) stated that in line with the application of the AIC for use with maximum likelihood estimation, parameters are re-estimated for each possible model, implying that both regression and scale estimators change

from model to model. Therefore, in the same spirit as using AIC for regression, they consider $\widehat{\sigma}$ not fixed and define the scale-based robust version of AIC for M-estimation as,

$$RAIC_S.M = 2n \log \widehat{\sigma}_M + 2 \text{ trace}(M^{-1}Q), \tag{2.18}$$

where $\widehat{\sigma}_M$ is the scale estimator from the M-estimation procedure and it changes from model to model. Moreover, Tharmaratnam and Claeskens (2013) suggest a different way to estimate the trace term (the penalty term). The whole empirical information matrices are considered as the estimates of M and Q,

$$M = E[\psi_c' \cdot xx^T] \approx \frac{1}{n} \sum_{i=1}^{n} \rho_c'' \left(\frac{y_i - X_i^T \widehat{\boldsymbol{\beta}}_M}{\widehat{\sigma}_M} \right) \frac{x_i x_i^T}{\widehat{\sigma}_M^2} \tag{2.19}$$

$$Q = E[\psi_c^2 \cdot xx^T] \approx \frac{1}{n} \sum_{i=1}^{n} \rho_c'^2 \left(\frac{y_i - X_i^T \widehat{\boldsymbol{\beta}}_M}{\widehat{\sigma}_M} \right) \frac{x_i x_i^T}{\widehat{\sigma}_M^2}, \tag{2.20}$$

where ρ_c' and ρ_c'' are the first and second derivatives of Huber's ρ function respectively. Then, the estimate of the trace term, which also varies from model to model, is computed by 2 trace$(M^{-1}Q)$.

Robust scale based AIC for MM-estimation could be computed in a similar way,

$$RAIC_S.MM = 2n \log \widehat{\sigma}_{MM} + 2 \text{ trace}(M^{-1}Q) \tag{2.21}$$

where $\widehat{\sigma}_{MM}$ is the scale estimator from MM-estimation. Moreover, when calculating the trace term for $RAIC_S.MM$, M-estimators should be replaced by MM-estimators, and also, the Tukey's biweight loss function should be used.

While previously mentioned robust model selection criteria are all based on some types of ρ functions, $RAIC_S.M$ and $RAIC_S.MM$ are scale-based and arranged in the same spirit as the traditional AIC for regression. However, the trace term trace$(M^{-1}Q)$ estimated in the equations (2.19) and (2.20) behaves quite abnormally under our simulation settings, especially when the number of

covariates is large. More of the simulation results and explanations are given in Section 4. Therefore, we will now further consider the criteria in equation (2.18) and (2.21) using an adjusted penalty term.

2.3.6 Robust scale based AIC for M and MM-estimation with the trace term adjusted

We calculate the penalty term for the robust scale based AIC in a similar way as the one suggested by Ronchetti (1985) and further denote these adjusted criteria using $RAIC'_S.M$ and $RAIC'_S.MM$ for M and MM-estimation accordingly,

$$RAIC'_S.M = 2n \log \widehat{\sigma}_M + 2\frac{E[\psi_c^2]}{E[\psi'_c]}p \tag{2.22}$$

$$RAIC'_S.MM = 2n \log \widehat{\sigma}_{MM} + 2\frac{E[\psi_d^2]}{E[\psi'_d]}p, \tag{2.23}$$

where ψ_c and ψ_d are Huber's function and Tukey's biweight respectively. In our simulation study, we have clearly found that these adjusted criteria outperform the original ones suggested by Tharmaratnam and Claeskens (2013), in terms of selection probabilities at various contamination levels.

2.4 Simulation results

In this section, we first introduce our simulation settings and then carry out a simulation study to compare the performances of different model selection criteria with respect to model selection and prediction accuracies.

2.4.1 Simulation settings

We recall the following linear regression model,

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \epsilon_i. \tag{2.24}$$

Therefore, we have p variables in total. In our simulation study, we investigate two cases of p: $p = 6$ and $p = 10$, with $n = 50$.

In the case of $p = 6$, the first three variables (X_1, X_2, X_3) are generated independently from standard normal distribution $N(0, 1)$. However, the other three variables (X_4, X_5, X_6) are considered correlated and we generate them through the relationship displayed below,

$$X_4 \sim N(0,1), \quad X_5 \sim 0.7X_4 + N(0,1), \quad X_6 \sim 0.7X_5 + N(0,1)$$

It is easy to show that the theoretical correlation between X_4 and X_5 is $\frac{0.7}{\sqrt{(1+0.7^2)}} = 0.5735$. Similarly, we have the theoretical correlation matrix for X_4, X_5, X_6 as follows,

$$\begin{pmatrix} 1 & 0.5735 & 0.3725 \\ 0.5735 & 1 & 0.6496 \\ 0.3725 & 0.6496 & 1 \end{pmatrix}$$

In addition, we define the true model by only using variables X_2, X_3, X_4, and X_5,

$$y_i = 1 + x_{2i} + x_{3i} + x_{4i} + x_{5i} + \epsilon_i \tag{2.25}$$

Hence, the true model is in the form of (2.24) but where $\boldsymbol{\beta} = (1, 0, 1, 1, 1, 1, 0)^T$. To investigate the performance of robust model selection criteria against outliers, we consider different percentages of outliers $(0\%, 10\%, 20\%, 30\%, 40\%)$ from $N(10, 1)$ on the response value $\mathbf{y}$. Therefore, we define the error ϵ_i distribution to be a mixture of normal distribution,

$$(1 - \varepsilon)N(0, 1) + \varepsilon N(10, 1), \tag{2.26}$$

where ε denotes the percentage of outliers, which varies from 0% to 40%. Additionally, the design matrix $\mathbf{X}$ is fixed over all simulation samples to reduce the simulation variability. Finally, 100 simulated samples are generated from the equation (2.25).

In the case of $p = 10$, we define the first six variables to be independent and the last four variables to be correlated. X_1, X_2, and X_3 are generated from uniform distribution and X_4, X_5, X_6, and X_7 are generated from standard normal distribution. Therefore, we have

$$X_1, X_2, X_3 \sim U(-1, 1), \quad X_4, X_5, X_6, X_7 \sim N(0, 1),$$

The other three variables are generated as follows,

$$X_8 \sim 0.7X_7 + N(0, 1), \quad X_9 \sim 0.7X_8 + N(0, 1), \quad X_{10} \sim 0.7X_9 + N(0, 1).$$

Similar to the case of $p = 6$, it is easy to find the theoretical correlation matrix for X_7, X_8, X_9 and X_{10}, which is as follows,

$$\begin{pmatrix} 1 & 0.5735 & 0.3725 & 0.2523 \\ 0.5735 & 1 & 0.6496 & 0.4400 \\ 0.3725 & 0.6496 & 1 & 0.6773 \\ 0.2523 & 0.4400 & 0.6773 & 1 \end{pmatrix}$$

For $p = 10$, we define the true model by using only variables X_2, X_3, X_5, X_6, X_7 and X_8,

$$y_i = 1 + x_{2i} + x_{3i} + x_{5i} + x_{6i} + x_{7i} + x_{8i} + \epsilon_i. \tag{2.27}$$

Hence, the regression coefficients for $p = 10$ are $\boldsymbol{\beta} = (1, 0, 1, 1, 0, 1, 1, 1, 1, 0, 0)^T$. Again, the design matrix for $p = 10$ is fixed and the error distribution is as in (2.26).

For both cases of $p = 6$ and $p = 10$, we further consider scenarios with outliers in covariates. We artificially add 10% of outliers to covariates from the $N(10, 1)$ distribution and denote the new design matrix by $\mathbf{X}^*$. In this situation, we still use the original design matrix $\mathbf{X}$ to generate the response y according to (2.26), considering that we require 'bad' leverage points that have a larger effect on the

regression estimation.

To compute the robust model selection criteria discussed in the above section, we find M and MM-estimators by using the function '*rlm*' from the R package 'MASS'. While calculating $Mn.M$ and $Mn.MM$, due to high computation cost, 20 bootstraps are used to find the expected conditional loss. In addition, we find that increasing the number of bootstraps does not improve the performance substantially.

The comparison is conducted by measuring model selection and prediction accuracies. Model selection accuracy is measured by the selection probability. The selection probability is the proportion of times that the selected best model includes all significant variables and excludes all noise variables over 100 simulations. Prediction accuracy is measured by the mean squared prediction error (MSPE) $n^{-1}\sum_{i=1}^{n}(\widehat{y}_i - y_i)^2$, computed over a set of independent test samples using the same sample size n as the training sample.

2.4.2 Simulation results

We now present the simulation results of various robust model selection criteria in terms of both model selection and prediction accuracy for each simulation scenario.

Table 2.1 displays detailed simulation results for the simulation settings of $p = 6$ and $n = 50$ without any outliers in the covariates. We first investigated the performance of the criteria based on loss function (using the first five methods). As expected, classical AIC works well when there are no outliers presented in the data. However, as the proportion of outliers in the response increases in the dataset, classical AIC selects the true model less frequently. When the contamination level reaches 20%, it rarely selects the true model. This indicates that it is worth investigating the robust model selection criteria under the presence of outliers in the dataset. The third and fifth columns confirm that the model selection criteria based on M-estimation ($Mn.M$ and $RAIC.M$) select a higher

Table 2.1: Selection probabilities for $p = 6$

ε	Based on loss function					Based on scale estimator			
%	AIC	$Mn.M$	$Mn.MM$	$RAIC.M$	$RAIC.MM$	$RAIC_S.M$	$RAIC_S.MM$	$RAIC'_S.M$	$RAIC'_S.MM$
0	0.72	0.57	0.71	0.63	0.63	0.22	0.17	0.30	0.41
10	0.19	0.76	0.72	0.60	0.71	0.12	0.33	0.19	0.45
20	0.01	0.51	0.81	0.47	0.70	0.19	0.38	0.24	0.73
30	0.03	0.12	0.51	0.10	0.41	0.09	0.33	0.09	0.73
40	0.00	0.00	0.12	0.01	0.06	0.01	0.24	0.00	0.44

proportion of correct models than the classical AIC in the presence of outliers in the response. However, then these two criteria break down when the contamination level reaches 30%, as a consequence of the low breakdown point of the M-estimators. This fact is also shown in Figure 2.1. Conversely, the fourth and sixth columns suggest that when the contamination level is high, the robust model selection criteria based on MM-estimation ($Mn.MM$ and $RAIC.MM$) generally outperform both the classical AIC and their counterparts based on M-estimation.

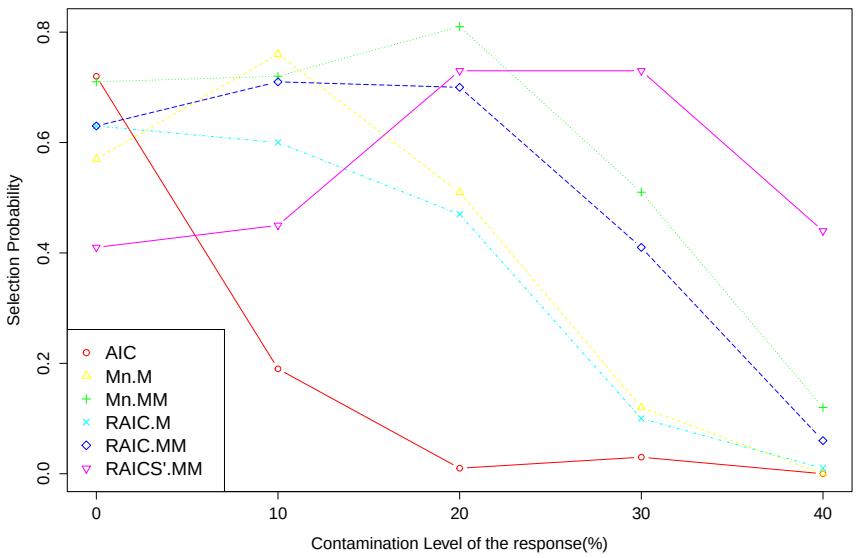


Figure 2.1: Selection probabilities of various model selection criteria for $p = 6$

However, the criteria based on MM-estimation become less effective after the contamination level reaches 30%, though the theoretical breakdown point of the

Table 2.2: Bias of the robust scale estimators for the full model when $p = 6$

ε	n=50				n=1000			
%	M	$MAD.M$	MM	$MAD.MM$	M	$MAD.M$	MM	$MAD.MM$
0	-0.08	-0.10	-0.07	-0.10	0.00	0.00	0.00	0.00
5	-0.00	-0.01	0.03	-0.04	0.06	0.06	0.06	0.06
10	0.11	0.09	0.15	0.06	0.16	0.14	0.15	0.14
20	0.79	0.54	0.40	0.28	0.52	0.40	0.38	0.39
30	3.35	2.04	1.15	1.38	3.88	0.97	0.73	0.84
40	5.26	3.64	2.83	3.67	5.95	2.10	1.54	2.30

MM-estimator is at 50%. This could be explained by the bias of the robust estimation of scale as discussed in Martin et al. (1993), who argued that the maximal asymptotic bias is substantial for large ε even for the MAD, the scale estimator that we proposed for our robust model selection criteria based on loss function. We also generated 100 samples to find the average bias of the following scale estimators under our simulation setting: scale estimator from M-estimation, MAD of the residuals from M-regression, scale estimator from MM-estimation (S-estimation), and MAD of the residuals from MM-regression. Table 2.2 suggests that when the contamination level goes up, the bias of MAD increases and becomes quite significant when ε reaches 30% for $MAD.M$ and 40% for $MAD.MM$. Further, compared with the large sample case (n=1000), the bias of these scale estimators is more notable in the small sample case (n=50) that we used in our simulation setting. The loss function in the above criteria may behave less effectively when the MAD is highly biased. Therefore, these robust model selection criteria do not perform as well as we expect when the contamination level increases. Additionally, Müller and Welsh (2005) also recommend using their model selection criterion only if σ is estimated to have a small expected bias.

We now concentrate on the last four columns in Table 2.1, which represent the selection probabilities for those scale based robust model selection criteria. Obviously, MM-type scale-based robust selection criteria outperform the M-type criteria over all ranges of contamination levels, especially for higher ε. Impor-

tantly, it is worth noting that for both of M-type and MM-type scale-based criteria the one with trace term adjusted significantly performs better than the original one suggested by Tharmaratnam and Claeskens (2013) considering various proportions of outliers.

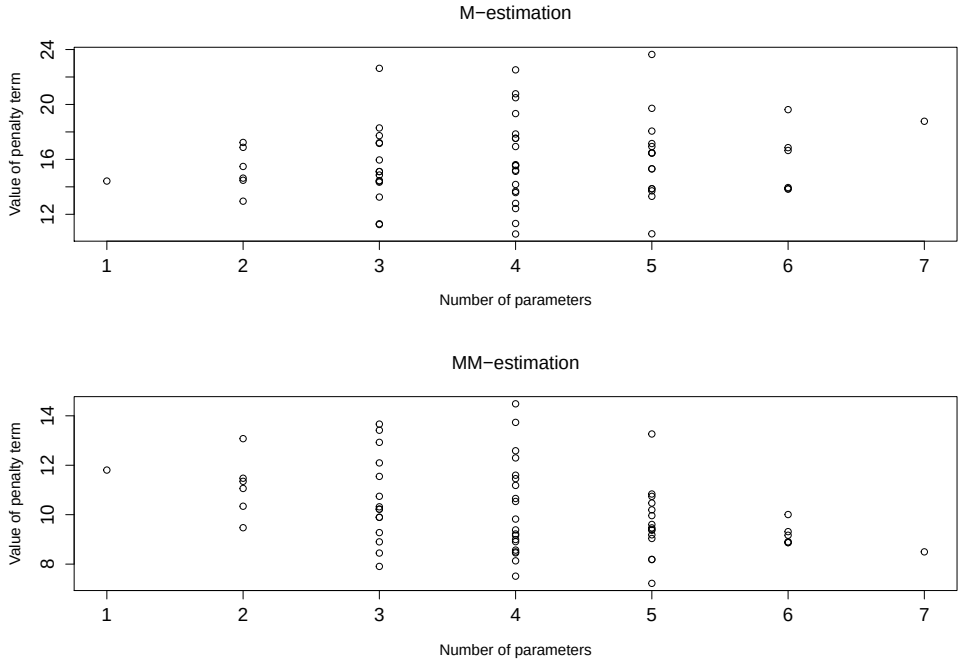


Figure 2.2: Penalty terms in $RAIC_S.M$ and $RAIC_S.MM$ for each possible model when $p = 6$, $n = 50$ and $\varepsilon = 20\%$

Ideally, the trace term (the penalty term) should increase as the number of variables included in the model goes up. However, in our simulation study, we further find that this property of the penalty term could not be achieved by the estimation procedure in Tharmaratnam and Claeskens (2013). Figure 2.2 represents the penalty terms in $RAIC_S.M$ and $RAIC_S.MM$ for each possible model in one of the simulation samples where $\varepsilon = 20\%$. As shown in Figure 2.2, when the number of regression parameters included in the model rises, there are no clear increasing trends for the penalty terms in either $RAIC_S.M$ or $RAIC_S.MM$. The penalty term for $RAIC_S.MM$ even exhibits a slightly decreasing pattern, which violates its usefulness. We consider this is due to the fact that Tharmaratnam and Claeskens (2013) estimated the trace term by using the entire empirical informa-

tion matrices, incorporating the covariance term of the covariates $x_i x_i^T$. However, in Ronchetti's computation method, the penalty term is evaluated by $2\frac{E[\psi^2]}{E[\psi']}p$, in which the expected covariance term of the covariates has already been cancelled out in the estimation procedure and the penalty term is linearly related to the number of variables in the model. Therefore, the adjusted penalty term turns out to be much more stable and it leads to a better performance of $RAIC'_S.MM$ than $RAIC_S.MM$. Though Tharmaratnam and Claeskens (2013) stated that the effect of the outliers on the penalty element of robust selection criteria based on scale estimators was seen to be non-influential, the finding under our simulation setting shows a significant improvement in the adjusted penalty term for selection probabilities. Because of the poor performance of $RAIC_S.M$ and $RAIC_S.MM$, we exclude them from comparison for the rest of the simulation studies and only account for $RAIC'_S.MM$ as a model selection criterion based on a scale estimator.

Another remarkable finding is that the robust model selection criterion based on the MM-type scale-based estimator selects a higher proportion of true models as the contamination level increases up to 40 %, compared with those criteria based on the loss function. This is explained in Claeskens and Tharmaratnam (2011). When the contamination level is low (0% or 10%), the overfit model obtains a smaller scale estimate than true models on average. In such cases, the model selection criteria based on scale estimate will often select an overfit model, as a small scale estimate is preferable since we are minimizing the criteria values. However, when the contamination level goes up to 20% or 30%, the true model results in a smaller scale estimate on average than the overfit or wrong fit models. Hence, the model selection criteria based on scale estimates will more often tend to select the true model.

Moreover, as displayed in Figure 2.1, the pink line lies above each of the blue and green dashed lines when ε reaches 30 %, indicating that the selection probability of the MM-type scale-based criterion with an adjusted penalty term ($RAIC'_S.M$) is outstanding by comparison with all other criteria, includ-

Table 2.3: Selection probabilities for $p = 6$ with x-outliers

ε %	AIC	$Mn.M$	$Mn.MM$	$RAIC.M$	$RAIC.MM$	$RAIC'_S.MM$
0	0.05	0.39	0.63	0.46	0.43	0.44
10	0.00	0.35	0.53	0.11	0.31	0.41
20	0.00	0.15	0.44	0.07	0.22	0.44
30	0.02	0.10	0.32	0.05	0.15	0.46
40	0.00	0.02	0.10	0.01	0.04	0.26

ing $Mn.MM$ and $RAIC.MM$. Although the criteria based on loss function tend to perform quite well when the proportion of outliers is below 20%, the scale-based model selection criteria are preferable when the contamination level is high. Therefore, we are further inspired to choose the appropriate robust model selection criteria depending on the proportion of outliers.

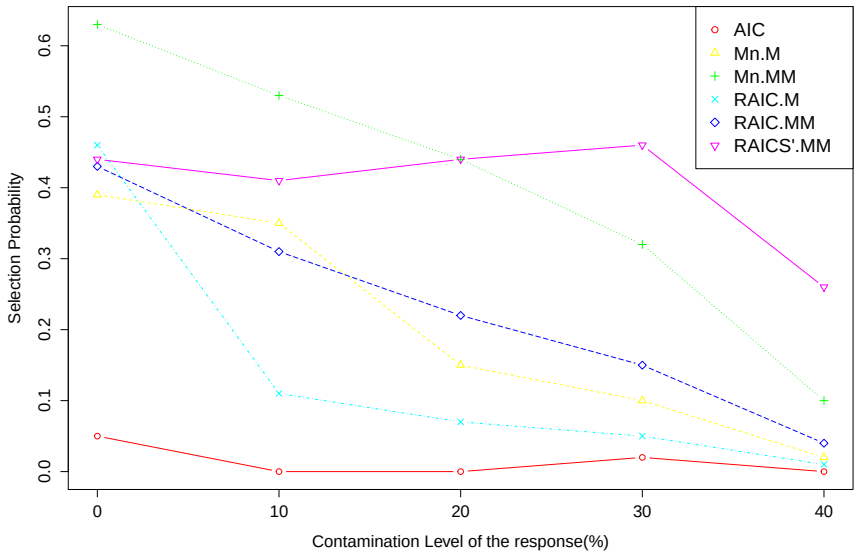


Figure 2.3: Selection probabilities of various model selection criteria for $p = 6$ with x-outliers

Table 2.3 and Figure 2.3 demonstrate the simulation results $p = 6$ and $n = 50$, but with 10% of outliers in the covariates (x-outliers) as discussed in the simulation setting. It is quite obvious that AIC exhibits a poor performance with any level of outliers in the response (y-outliers). It is also worth noting that

Table 2.4: Selection probabilities for $p = 10$

ε %	*AIC*	*Mn.M*	*Mn.MM*	*RAIC.M*	*RAIC.MM*	$RAIC_S.MM$	$RAIC'_S.MM$
0	0.31	0.25	0.46	0.30	0.27	0.02	0.09
10	0.04	0.38	0.42	0.35	0.28	0.06	0.10
20	0.00	0.30	0.47	0.13	0.31	0.09	0.32
30	0.00	0.04	0.18	0.01	0.12	0.01	0.36
40	0.00	0.00	0.06	0.00	0.00	0.06	0.20

Table 2.5: Selection probabilities for $p = 10$ with x-outliers

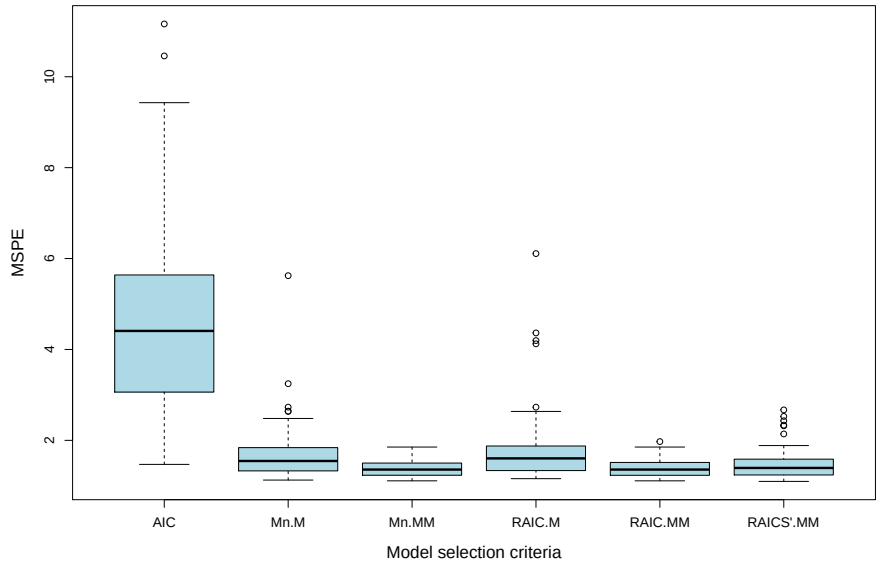
ε %	*AIC*	*Mn.M*	*Mn.MM*	*RAIC.M*	*RAIC.MM*	$RAIC'_S.MM$
0	0.00	0.28	0.40	0.11	0.22	0.10
10	0.00	0.16	0.37	0.09	0.24	0.12
20	0.00	0.05	0.27	0.01	0.09	0.23
30	0.00	0.03	0.07	0.00	0.03	0.15
40	0.00	0.00	0.01	0.00	0.00	0.14

in the presence of x-outliers, the model selection criteria based on MM-estimation generally outperform those based on M-estimation even when the contamination level of ϵ is low. This is also shown in Figure 2.3 as the selection probabilities of the MM-type criteria reside above those of the M-type criteria. This strongly indicates the usefulness of the model selection criteria based on MM-estimation in the presence of both x- and y-outliers.

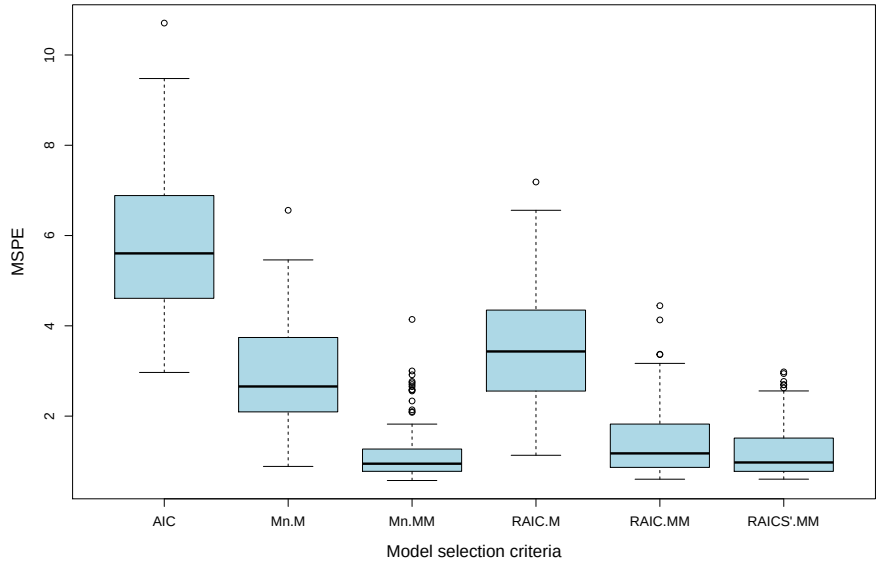
Figure 2.4 presents the mean squared predication errors for $p = 6$ and $\epsilon = 10\%$ with and without x-outliers. From the prediction point of view, we still see very strong evidence that the robust model selection criteria outperform non-robust traditional AIC in the presence of y-outliers. In the presence of both x- and y-outliers, those based on MM-estimation achieve substantially lower MSPEs and significantly outperform those based on M-estimation.

We further investigate the case of $p = 10$. The detailed simulation results of $p = 10$ and $n = 50$ with and without outliers in the covariates are shown in Tables 2.4 and 2.5, respectively.

Overall, the selection probabilities for each of the criteria we consider are



(a)



(b)

Figure 2.4: Mean squared predication errors for $p = 6$ and $\epsilon = 10\%$, (a) without x-outliers, (b) with x-outliers

.

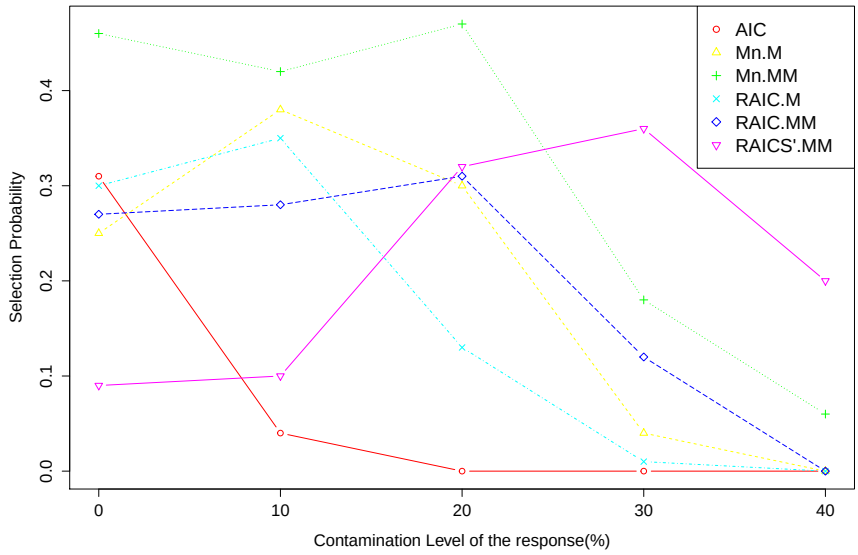


Figure 2.5: Selection probabilities of various model selection criteria for $p = 10$

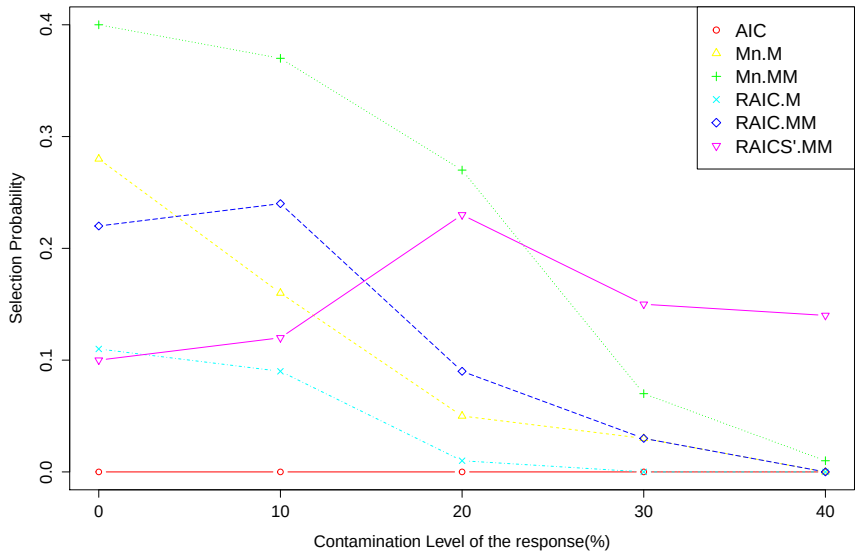
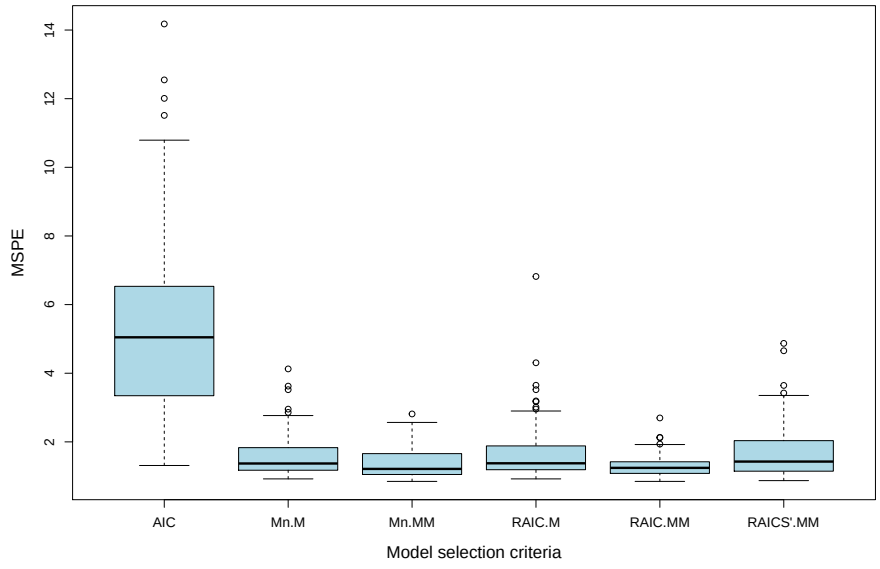
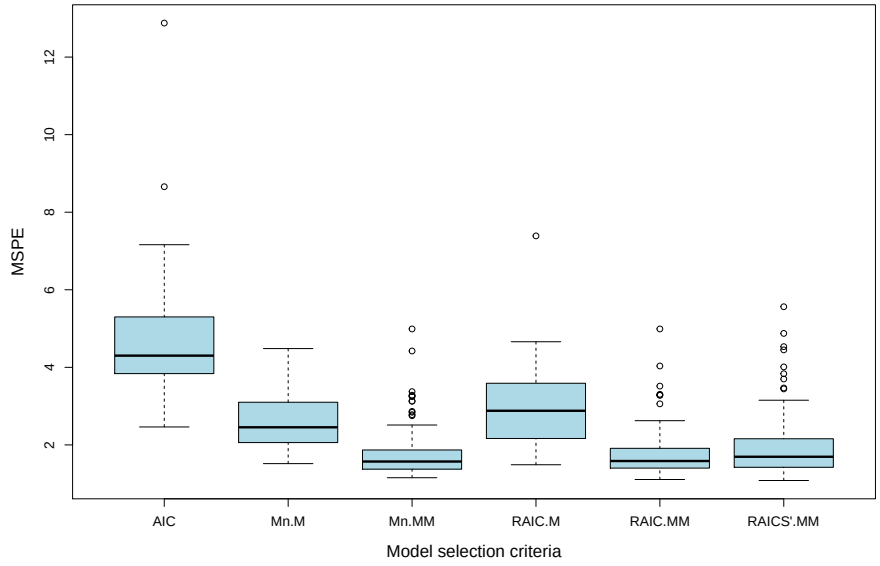


Figure 2.6: Selection probabilities of various model selection criteria for $p = 10$ with x-outliers



(a)



(b)

Figure 2.7: Mean squared predication errors for $p = 10$, (a) without x-outliers, (b) with x-outliers

.

smaller in the case of $p = 10$, compared with the case of $p = 6$. This could be expected as the selection difficulty increases when the number of variables included in the model goes up. Figures 2.5 and 2.6 suggest that similar to the case of $p = 6$, the criteria based on MM-estimation outperform those based on M-estimation in the presence of x-outliers or a high proportion of y-outliers. Moreover, as shown in Table 2.4 it is worth noting that the relative discrepancy of selection probabilities between $RAIC'_S.MM$ and $RAIC_S.MM$ is much larger in the case of $p = 10$ than in the case of $p = 6$. This further indicates that when the number of variables in the model increases the original criteria as suggested by Tharmaratnam and Claeskens (2013) perform more poorly as a result of the increasing variability in the covariance term of the covariates $x_i x_i^T$. Therefore, we could further conclude that the adjustment to the penalty term of the scale-based criteria improves its performance on selection probability.

From the prediction point of view, it is quite obvious that those based on MM-estimation still substantially outperform those based on M-estimation, as illustrated in Figure 2.7.

2.5 Real data example

We now apply these robust model selection criteria to analyze the well-known Boston housing data (available at `http://lib.stat.cmu.edu/datasets/boston`). The data contain the following 14 variables: crim (per capita crime rate by town), zn (proportion of residential land zoned for lots over 25,000 sq.ft), indus (proportion of nonretail business acres per town), chas (Charles River dummy variable), nox (nitrogen oxides concentration: parts per 10 million), rm (average number of rooms per dwelling), age (proportion of owner-occupied units built prior to 1940), dis (weighted mean of distances to five Boston employment centres), rad (index of accessibility to radial highways), tax (full-value property-tax rate per \$10,000), ptratio (pupil-teacher ratio by town), black ($1000(B_k - 0.63)^2$), where B_k is the proportion of African-American residents by town), lstat (lower status

Table 2.6: Trimmed mean square prediction error (TMSPE) for Boston housing data

Method	Average TMSPE	SD TMSPE
AIC	9.23	0.59
$Mn.M$	7.43	0.34
$Mn.MM$	7.38	0.73
$RAIC.M$	7.43	0.34
$RAIC.MM$	7.28	0.31
$RAICS'.MM$	7.30	0.33

of the population in percentages), and medv (median value of owner-occupied homes in thousand dollars). There are 506 observations in the dataset. The response variable is medv. Following Müller and Welsh (2005), we utilized m out of n stratified bootstrap (see Müller and Welsh (2005)) to generate 100 bootstrap resamples as testing samples to ensure that outliers are present in each. The comparison was then measured by the average prediction loss of these 100 testing samples (bootstrap resamples), namely, the conditional expected prediction loss. To be robust, a good model should capture the pattern of the majority of data. Therefore, we used the trimmed mean square prediction error (TMSPE), as a more appropriate measure of the prediction loss for this dataset. We truncated the largest 10 % of squared residuals and computed the TMSPE using the remaining 90% of the squared residuals. TMSPE is a measure of prediction accuracy for the majority of the data and is no longer dominated by extreme prediction errors. Table 2.6 and Figure 2.8 present the TMSPE for Boston housing data using different model selection criteria.

From Table 2.6, we can clearly see that the best model selected by AIC obtains the highest TMSPE of 9.23, showing a very poor performance compared with these robust model selection criteria and indicating a certain proportion of outliers in the Boston housing data. These robust model selection criteria demonstrate a comparable performance. Among them, $RAIC.MM$ achieves the lowest average TMSPE with the smallest standard error. This is also shown in

Figure 2.8.

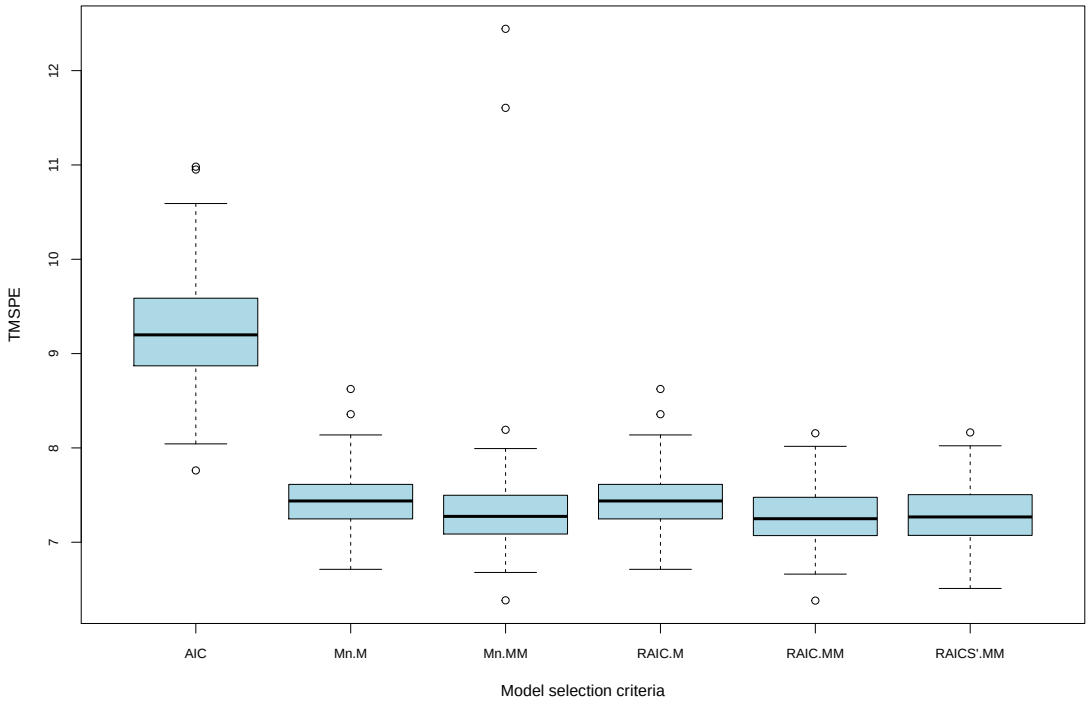


Figure 2.8: Trimmed mean square prediction error (TMSPE) for Boston housing data

2.6 Conclusion

In this thesis, we propose a robust AIC for MM-estimation and an adjusted robust scale based AIC for M and MM-estimation. Our proposed model selection criteria can maintain their robust properties in the presence of a high proportion of outliers and the outliers in the covariates. We compare our proposed criteria with other robust model selection criteria discussed in previous literature. Our simulation studies demonstrate a significant outperformance of robust AIC based on MM-estimation with outliers in the covariates and show a better performance of the adjusted robust scale based AIC for MM-estimation when the proportion of outliers in the response is relatively high. Real data examples also show a

better performance of robust AIC based on MM-estimation. More robust model selection criteria with different penalty terms (e.g. robust BIC) will require future research. Overall, the traditional AIC is very sensitive to outliers, while the robust versions of model selection criteria are resistant to outliers and should receive more attention in model selection studies.

Chapter 3

Robust Lasso Regression Using Tukey's Biweight Criterion *

3.1 Introduction

In multiple linear regression models, the ordinary least squares (OLS) estimates can give inaccurate predictions when there are a large number of predictors or multicollinearity is present. One way to improve the predictions is to reduce the number of variables in the model by model selection. Tibshirani (1996) proposed a new technique for model selection termed the "LASSO" for "Least Absolute Shrinkage and Selection Operator", which incorporates an $L1$ penalty into the OLS loss function. The lasso shrinks some coefficients to exactly zero. This property of the lasso means that it provides parsimonious solutions that are easy to interpret.

Over the past decade, the lasso has become a very popular technique for simultaneous estimation and variable selection. Many authors (Zou, 2006; Knight and Fu, 2000; Zou et al., 2007; Zhao and Yu, 2006) have investigated the properties of the lasso and developed different variants of the lasso. Zou (2006) demonstrated

*The core contribution of Chapter 3 was submitted to *Technometrics* in April 2016 and has been accepted for publication in February 2017. This chapter was presented at the CM-Statistics Conference held in London, the UK in December 2015.

that there exist certain scenarios where the lasso is inconsistent for variable selection. Thus, he suggested the adaptive lasso, where adaptive weights are used for penalizing coefficients differently in the $L1$ penalty. The adaptive lasso enjoys the oracle property, that is, asymptotically it performs as well as if the true underlying model were known. Moreover, the adaptive lasso can be computed using the same efficient algorithms that are used to compute the lasso, for example, least angle regression (LARS) of Efron et al. (2004). Zou and Hastie (2005) emphasized the inappropriateness of using the lasso as a variable selection method if a group of variables are very highly correlated. In such a situation, the lasso tends to select only one variable from the group and does not care which variable is selected. To overcome this drawback of the lasso, Zou and Hastie (2005) developed the elastic net, a new regularization and variable selection method that combines an $L1$ and $L2$ penalty. The elastic net performs variable selection, continuous shrinkage, and more importantly, it selects groups of strongly correlated variables. Fan and Li (2001) argued that a good penalty function should have the properties of sparsity and unbiasedness. They proposed a special non-concave penalty function named the Smoothing Clipped Absolute Deviation (SCAD) that can produce sparse solutions and unbiased estimates for large parameters.

Datasets with outliers are commonly encountered in statistical analysis. These outliers may appear in the response and/or the predictors. It is well known that OLS estimates in linear regression are very sensitive to outliers. The lasso estimates, which utilize OLS, also suffer from the effect of outliers. Some authors have considered robust versions of the lasso, generally utilizing penalized versions of M-estimators, as in Owen (2007), Wang et al. (2007), Li et al. (2011), and Lambert-Lacroix and Zwald (2011). Wang et al. (2007) proposed to overcome the presence of outliers by combining the least absolute deviation (LAD) loss and the lasso penalty. Unfortunately, it is well known that the LAD loss is not adaptable for small errors because it penalizes strongly for small residuals (Owen, 2007; Lambert-Lacroix and Zwald, 2011). That is, the LAD-Lasso has

lower efficiency than OLS estimates when there are no outliers in the response. Owen (2007) and Lambert-Lacroix and Zwald (2011) preferred to replace the squared loss with Huber's loss, a hybrid of the squared error and absolute error loss functions. Extensive simulation studies in Lambert-Lacroix and Zwald (2011) have demonstrated the superior performance of the lasso with Huber's criterion over the traditional lasso and the LAD-Lasso.

Although the robust lasso with the Huber's loss is resistant to outliers in the response and achieves high asymptotic efficiency, it is not robust against high leverage points or outliers in the covariates. As discussed above, the literature on the robust lasso (Owen, 2007; Wang et al., 2007; Li et al., 2011; Lambert-Lacroix and Zwald, 2011) considers only outliers in the response. However, outliers in the covariates also appear frequently and they generally have a greater effect on the accuracy of the regression estimates than outliers in the response. Only a few papers have considered robust methods with respect to contamination in the covariates. Maronna (2011) proposed S-Ridge and MM-Ridge regressions that add an $L2$ penalty to traditional unpenalized S-regression and MM-regression. These methods are shown to be robust against outliers in the covariates. However, similar to ridge regression, S-Ridge and MM-Ridge do not perform variable selection. Khan et al. (2007) introduced a robust version of LARS (Efron et al., 2004), named Rlars, by replacing the classical correlation in the LARS algorithm with robust correlation estimates. Rlars orders the importance of variables robustly. However, the Rlars procedure does not optimize a clearly defined objective function, which means its asymptotic properties cannot be explored theoretically. Alfons et al. (2013) proposed another approach that is robust with respect to high leverage points, by adding the $L1$ penalty to the well-known least trimmed squares (LTS) criterion, naming this approach the Sparse-LTS. They derived the breakdown point of this Sparse-LTS estimator and demonstrated that it can be robust to outliers in both the covariates and the response. However, simulation studies in Alfons et al. (2013) and in our work, find that Sparse-LTS loses effi-

ciency when there are no outliers. Moreover, Alfons et al. (2013) do not provide any asymptotic theory for their Sparse-LTS estimator.

To be robust against outliers in both covariates and the response, the derivative of the loss function needs to be redescending (Rousseeuw and Yohai, 1984; Yohai, 1987). A commonly used loss function with this property is Tukey's biweight (Tukey, 1960). In this paper, we propose replacing the squared loss in the lasso with Tukey's biweight criterion, and name the method the Tukey-lasso, to handle outliers in the response and covariates. Contemporary works based on a similar idea can be found in Smucler and Yohai (2015, 2017). In our simulation study, we show that the Tukey-lasso outperforms the adaptive lasso and other robust implementations of the lasso, particularly in the presence of outliers in both the response and the predictors.

We further propose an accelerated proximal gradient (APG) algorithm to compute the Tukey-lasso. The APG computes the lasso minimization problem and guarantees a global minimizer for a convex objective function. Although the objective function for the robust lasso with Tukey's biweight is non-convex, the APG algorithm still achieves very reliable results (a local minimizer) when the starting value of the algorithm is carefully selected. We further demonstrate that the computation time for the Tukey-lasso through the APG algorithm is substantially lower than that of its competitors, including Rlars and Sparse-LTS.

In Section 3.2, we describe the traditional lasso, other robust versions of the lasso and introduce the robust lasso with Tukey's biweight loss. In Section 3.3, we introduce the accelerated proximal gradient algorithm and its implementations to lasso problems. In Section 3.4, we discuss the method of selecting the tuning parameter. In Section 3.5, we present our simulation settings, show our simulation results and compare the prediction and variable selection accuracy of various forms of the lasso. Computation times for these methods are also reported and compared. We analyse three real examples in Section 3.6. Finally, we present brief conclusions in Section 3.7.

3.2 The lasso-type estimate

3.2.1 The traditional lasso

Consider a standard linear regression model,

$$y_i = X_i^T \boldsymbol{\beta} + \epsilon_i, \qquad i = 1, 2, \ldots, n, \tag{3.1}$$

where y_i is the response variable on the i-th observation, $X_i = (1, x_{i1}, x_{i2}, \ldots, x_{ip})^T$ is a $p+1$ vector of covariates on the i-th observation, $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \ldots, \beta_p)^T$ is a $p+1$ vector of regression parameters, and ϵ_i are independently distributed random variables with expected value 0 and variance σ^2.

Although OLS estimates are unbiased, they can result in highly variable predictions when no variable selection is performed. To improve prediction by shrinking unnecessary coefficients to 0, Tibshirani (1996) proposed to add the $L1$ norm of the estimates to the squared loss, leading to the lasso estimator. However, the lasso is inconsistent for variable selection, so Zou (2006) suggested the adaptive lasso in which adaptive weights are used to penalize the coefficients differently. The adaptive lasso considers the following modified lasso criterion,

$$\text{argmin} \sum_{i=1}^{n} \left(y_i - X_i^T \boldsymbol{\beta}\right)^2 + \lambda \sum_{j=1}^{p} \widehat{w}_j |\beta_j|, \tag{3.2}$$

where λ is a tuning constant and $\widehat{w}_j = 1/|\widehat{\beta}_j|$, where $\widehat{\beta}_j$ is the OLS estimate for the j^{th} coefficient. When the adaptive weights $\widehat{w}_j = 1$ for each j, the adaptive lasso reduces to the lasso, as proposed in Tibshirani (1996). The adaptive lasso can be implemented using the same algorithm that is used for the lasso. Zou (2006) showed that the adaptive lasso enjoys the oracle property.

3.2.2 Robust lasso

It is well known that squared loss used in the traditional lasso is very sensitive to outliers. The goal of the robust lasso is to offer a more stable alternative that is not sensitive to outliers. We propose combining robust M-estimation and the adaptive lasso penalty, to obtain the generalized adaptive lasso,

$$\operatorname{argmin} 2\sum_{i=1}^{n}\rho\left(\frac{y_i - X_i^T\boldsymbol{\beta}}{\sigma}\right) + \lambda\sum_{j=1}^{p}\widehat{w}_j|\beta_j|. \tag{3.3}$$

When ρ is the squared loss and $\sigma = 1$, (3.3) is simply the adaptive lasso (3.2). When ρ is the LAD loss, $\sigma = 1$ and the weights $\widehat{w}_j = 1/|\widehat{\beta}_j|$, where $\widehat{\beta}_j$ is the unpenalized LAD estimate of the j^{th} coefficient, (3.3) leads to the LAD-Lasso proposed by Wang et al. (2007). Since the squared loss in (3.2) has been replaced by the LAD criterion ($L1$ loss) in (3.3), the resulting estimator is expected to be more robust to outliers. The LAD-Lasso estimator produces consistent variable selection and extensive simulation studies in Wang et al. (2007) demonstrate the satisfactory finite-sample performance of the LAD-Lasso.

It is worth noting that when there are no outliers in the response, the LAD-Lasso achieves lower efficiency than the adaptive lasso. Another choice of ρ that is robust against heavy-tailed errors or outliers is Huber's loss function. Huber et al. (1964) describes Huber's loss function as

$$\rho_c(u) = \begin{cases} \frac{u^2}{2} & \text{if } |u| \le c \\ c|u| - \frac{c^2}{2} & \text{if } |u| > c, \end{cases} \tag{3.4}$$

where c is a tuning constant that determines where the transition from quadratic to linear occurs. For large values of c, Huber's loss function acts like a least squares function, while for small values of c, it is similar to the LAD loss ($L1$ norm), making it more robust against outliers but less efficient for normal errors. That is, the tuning constant c controls the trade-off between robustness and efficiency. A standard choice of c is $c = 1.345$ for 95% asymptotic efficiency for

the standard normal distribution. Therefore, when ρ is Huber's loss function and $\widehat{w}_j = 1/|\widehat{\beta}_j^M|$, where $\widehat{\beta}_j^M$ denotes the unpenalized Huber estimate for the j^{th} coefficient, (3.3) leads to the robust adaptive lasso with Huber's criterion, discussed in Owen (2007) and Lambert-Lacroix and Zwald (2011). Lambert-Lacroix and Zwald (2011) demonstrate that the lasso with Huber's loss achieves the oracle property. More details of the estimation of $\boldsymbol{\beta}$ and σ are given in Lambert-Lacroix and Zwald (2011).

Although the estimates obtained from the LAD-Lasso and the lasso with Huber's loss are resistant to outliers in the response, they are not robust against outliers in the covariates. Alfons et al. (2013) proposed Sparse-LTS by adding the $L1$ penalty to the least trimmed squares (LTS) criterion. The Sparse-LTS is robust with respect to high leverage points. Denote $r_i = y_i - X_i^T\boldsymbol{\beta}$, and $r_{1n}^2 \leq \ldots \leq r_{nn}^2$ the order statistics of the squared residuals. Further define $I\left(r_i^2 \leq r_{hn}^2\right)$ the indicator function that equals 1 when the i^{th} squared residual $r_i^2 \leq r_{hn}^2$. Then, when $\rho = \frac{1}{2}r_i^2 I\left(r_i^2 \leq r_{hn}^2\right)$, $\sigma = 1$ and the weights $\widehat{w}_j = 1$, (3.3) reduces to the Sparse-LTS as introduced in Alfons et al. (2013). A standard choice of h is $h = \lfloor 0.75(n+1) \rfloor$. However, this truncation of the data may result in a loss of statistical efficiency. To overcome this loss of efficiency, Alfons et al. (2013) proposed a reweighting step to increase efficiency, Alfons et al. (2013) and our simulation study show that the Sparse-LTS performs unsatisfactorily when the data have no outliers.

3.2.3 Robust lasso with Tukey's biweight criterion

To be robust against outliers in both covariates and responses, the derivative of the loss function needs to be redescending (Rousseeuw and Yohai, 1984; Yohai, 1987). A commonly used family of such loss functions is Tukey's biweight,

$$\rho_d(u) = \begin{cases} \frac{d^2}{6}\left\{1 - \left[1 - \left(\frac{u}{d}\right)^2\right]^3\right\} & \text{if } |u| \leq d \\ \frac{d^2}{6} & \text{if } |u| \geq d, \end{cases} \tag{3.5}$$

where d is a tuning constant that, similar to c in Huber's function, controls the level of robustness. Tukey's biweight function truncates the residuals that are larger than d to the constant $d^2/6$. Therefore, small values of d imply higher robustness while large values of d provide higher efficiency. To achieve 95% asymptotic efficiency at the standard normal distribution, the suggested choice of d is 4.685. We propose to replace ρ in (3.3) by Tukey's biweight loss to deal with the outliers in both of the covariates and the response, the Tukey-lasso solves the following,

$$\text{argmin}\, 2\sum_{i=1}^{n} \rho_d\left(\frac{y_i - X_i^T\boldsymbol{\beta}}{\widehat{\sigma}}\right) + \lambda\sum_{j=1}^{p} \widehat{w}_j|\beta_j|, \tag{3.6}$$

where ρ_d is Tukey's biweight function and $\widehat{\sigma}$ is a robust estimate of σ. In our study, σ is estimated by the median absolute deviation (MAD) (Rousseeuw and Croux, 1993) of the residuals from the full model fitted by S-estimation (Rousseeuw and Yohai, 1984). Other robust scale estimates, such as S-scale estimates, are also acceptable here. Overall, we find that the performance of the Tukey-lasso is not sensitive to the choice of MAD or S-scale for estimating σ. The $\widehat{w}_j = 1/|\widehat{\beta}_j^{MM}|$ are weights based on the MM-estimates $\widehat{\beta}_j^{MM}$ (Yohai, 1987).

Without loss of generality, we assume the true model contains the first p_0 variables such that $\mathcal{A} = \{1, 2, \cdots, p_0\}$. Write

$$\mathbf{C} = \begin{bmatrix} \mathbf{C_{11}} & \mathbf{C_{12}} \\ \mathbf{C_{21}} & \mathbf{C_{22}} \end{bmatrix},$$

where $\mathbf{C_{11}}$ is a $p_0 \times p_0$ positive definite matrix. To prove that the Tukey-lasso achieves the oracle property, we make the following assumptions,

- A1: $\frac{1}{n}\mathbf{X}^T\mathbf{X} \to \mathbf{C}$;
- A2: $\max_{1\le i\le n}\|X_i\|/\sqrt{n} \to 0$ as $n \to \infty$
- A3: the ϵ_i have a symmetric distribution with mean 0 and variance σ^2;

- A4: $\sqrt{n}(\widehat{\sigma}_n - \sigma)$ is bounded in probability.

Theorem 1. Assume conditions A1 to A4, and further suppose that $\lambda_n \to \infty$ such that $\lambda_n/\sqrt{n} \to 0$. Then the adaptive robust lasso estimator with Tukey's biweight loss and preliminary scale $\widehat{\sigma}_n$ (i.e. the Tukey-lasso) satisfies the following:

1. Asymptotic normality: $\sqrt{n}(\widehat{\boldsymbol{\beta}}_{\mathcal{A}}^{(n)} - \boldsymbol{\beta}_{\mathcal{A}}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \sigma^2 \frac{E\psi_d^2}{(E\psi_d')^2}\mathbf{C}_{11}^{-1})$, where $E\psi_d^2$ and $E\psi_d'$ are the expected values of ψ_d^2 and ψ_d' respectively.

2. Consistency in variable selection: $\lim_n P(\mathcal{A}_n = \mathcal{A}) = 1$

Proof: The proof of Theorem 1 is provided in the Appendix A.

We use the MM-estimator to calculate the predetermined weights because the MM-estimator achieves high efficiency and is robust against both response and covariate outliers. MM-estimation is a three-stage procedure introduced by Yohai (1987). In the first stage, we compute an initial regression estimate (e.g. LMS, Rousseeuw (1984)) with a high breakdown point that is not necessarily efficient. In the second stage, an M-estimate of scale is calculated based on the initial estimator (first M). In the third stage, we compute an M-estimator of regression coefficients based on the scale estimator obtained in the second stage (second M). Briefly, the MM-estimator is computed as an M-estimator starting at a high breakdown S-estimator and with a fixed scale given by the S-estimator. For further details of MM-estimation, see Yohai (1987).

3.2.4 Robust lasso with Tukey's biweight criterion when $p > n$

In a high dimensional setting, traditional robust estimation methods, such as MM-estimation, become computationally infeasible. One major advantage of the Tukey-lasso, compared with unpenalized robust estimation, is that it is computationally efficient and produces both robust and sparse results in high dimensional settings where $p > n$.

The adaptive weights for the Tukey-lasso as in (3.6) are determined by unpenalized MM-estimates $\widehat{\beta}_j^{MM}$. To compute the adaptive weights when $p > n$, we replace the MM-estimate with MM-Ridge, as introduced in Maronna (2011). MM-Ridge is confirmed to be robust to both outliers in the covariates and the response. Further, $\widehat{\sigma}$, the robust estimate of σ, is estimated by the MAD of the residuals from the model fitted by MM-Ridge.

Simulations and real examples demonstrate that the Tukey-lasso with weights computed by MM-Ridge estimates produces reliable results when $p > n$. Algorithms for computing the Tukey-lasso are proposed in the following section. In our simulation study for both $p < n$ and $p > n$, we show that the Tukey-lasso outperforms its competitors in prediction and variable selection accuracy.

3.3 Algorithms for numerical optimization

We apply an accelerated proximal gradient (APG) algorithm to compute the Tukey-lasso estimators. When the starting values of the algorithm are carefully chosen, the APG algorithm achieves very reliable results for the Tukey-lasso. Generally, the APG algorithm is very fast and also suitable for solving lasso-type problems with differentiable loss functions.

3.3.1 The traditional lasso

Consider the minimization problem,

$$\underset{x}{\operatorname{argmin}}\{F(x) := f(x) + g(x)\}, \tag{3.7}$$

where $f : R^n \to R$ is a differentiable convex function and $g : R^n \to (-\infty, \infty]$ is a proper, lower semi-continuous convex function. The lasso (Tibshirani, 1996) is

exactly the minimization problem (3.7), with

$$f(\boldsymbol{\beta}) = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 \quad \text{and} \quad g(\boldsymbol{\beta}) = \lambda\|\boldsymbol{\beta}\|_1.$$

Further note that $\nabla f(\boldsymbol{\beta}) = X^T(X\boldsymbol{\beta} - y)$ and the proximal is given by

$$\text{prox}_{tg}(\boldsymbol{\beta}) = \arg\min_z \left\{ \|z\|_1 + \frac{1}{2\lambda t}\|z - \boldsymbol{\beta}\|^2 \right\} = S_{\lambda t}(\boldsymbol{\beta}), \tag{3.8}$$

where S_λ denotes the soft thresholding function used by Donoho and Johnstone (1994),

$$S_\lambda(v) = \begin{cases} v - \lambda & \text{if } v > \lambda \\ 0 & \text{if } |v| \leq \lambda \\ v + \lambda & \text{if } v < \lambda. \end{cases} \tag{3.9}$$

Therefore, as discussed in Daubechies et al. (2004), the proximal gradient method becomes

$$\boldsymbol{\beta}_{k+1} = S_{\lambda t_k}\left(\boldsymbol{\beta}_k - t_k \boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{\beta}_k - \boldsymbol{y})\right),$$

with a suitably chosen step size $0 < t_k < 1/L$, where L is the Lipschitz constant of ∇f. This method is also known as the iterative soft thresholding algorithm (ISTA). Daubechies et al. (2004) show that the rate of convergence for this method is the same as for the classical gradient method, which is no worse than $O(1/k)$. For details of ISTA see Daubechies et al. (2004). To improve the rate of convergence, Beck and Teboulle (2009) proposed the APG method which preserves the computational simplicity of the proximal gradient method but achieves a global rate of convergence $O(1/k^2)$. Beck and Teboulle (2009) further applied the APG method to solve the lasso using the iterative method,

$$\begin{aligned} \widetilde{\boldsymbol{\beta}}_k &= \boldsymbol{\beta}_k + \frac{k-1}{k+2}(\boldsymbol{\beta}_k - \boldsymbol{\beta}_{k-1}) \\ \boldsymbol{\beta}_{k+1} &= S_{\lambda t_k}\left(\widetilde{\boldsymbol{\beta}}_k - t_k \boldsymbol{X}^T(\boldsymbol{X}\widetilde{\boldsymbol{\beta}}_k - \boldsymbol{y})\right), \end{aligned} \tag{3.10}$$

with suitably chosen step size $0 < t_k < 1/L$, which is also called the fast iterative soft thresholding algorithm (FISTA). According to Beck and Teboulle (2009), let $\boldsymbol{\beta}^*$ be any minimizer of $F(\boldsymbol{\beta})$ over R^n, then,

$$0 \leq F(\boldsymbol{\beta}_k) - F(\boldsymbol{\beta}^*) \leq \frac{\|\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*\|^2}{t(k+1)^2} \quad for \quad k \geq 1. \tag{3.11}$$

That is, the rate of convergence of the APG method is shown to be $O(1/k^2)$. Details of the proof of the convergence rate can be found in Beck and Teboulle (2009). Due to its fast convergence, we adopt this APG method to solve the Tukey-lasso.

3.3.2 Robust lasso with Tukey's biweight criterion

To solve the robust lasso with Tukey's biweight loss, note that $f(\boldsymbol{\beta})$ and its gradient are,

$$f(\boldsymbol{\beta}) = \rho_d(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \quad \text{and} \quad \nabla f(\boldsymbol{\beta}) = -\frac{1}{\sigma}\boldsymbol{X}^T\psi_d\left(\frac{\boldsymbol{y} - \boldsymbol{X}^T\boldsymbol{\beta}}{\sigma}\right),$$

where ρ_d is Tukey's biweight and ψ_d is its derivative,

$$\psi_d(u) = \begin{cases} u\left[1 - \left(\frac{u}{d}\right)^2\right]^2 & \text{if } |u| \leq d \\ 0 & \text{if } |u| > d. \end{cases} \tag{3.12}$$

Hence, the accelerated proximal gradient method leads to the following iterative method,

$$\begin{aligned} \widetilde{\boldsymbol{\beta}}_k &= \boldsymbol{\beta}_k + \frac{k-1}{k+2}(\boldsymbol{\beta}_k - \boldsymbol{\beta}_{k-1}) \\ \boldsymbol{\beta}_{k+1} &= S_{\lambda t_k}\left(\widetilde{\boldsymbol{\beta}}_k + \frac{t_k}{\sigma}\boldsymbol{X}^T\psi_d\left(\frac{\boldsymbol{y} - \boldsymbol{X}^T\boldsymbol{\beta}}{\sigma}\right)\right). \end{aligned} \tag{3.13}$$

Again, we estimate σ by the MAD of the residuals from the full model fitted by S-estimation. By simply replacing ρ_d with ρ_c, the algorithm (3.13) can be

used to solve the robust lasso with Huber's loss. However, as previously noted, Tukey's biweight loss function leads to a non-convex objective function and only a local minimizer can be achieved. In contrast to the traditional lasso and the lasso with Huber's loss (the convex cases), an appropriate choice of the initial value is essential. We consider the MM-estimator when $p < n$ and MM-Ridge when $p > n$, as initial values. These methods are considered because they are both efficient and robust. As such, the robust lasso with Tukey's biweight can be treated as a shrinkage operator for MM-estimation. Details of the computation of MM-estimators and MM-Ridge are discussed in Yohai (1987) and Maronna (2011), respectively.

3.3.3 Lasso-type problems with adaptive penalties

Now we consider algorithms for solving the adaptive lasso problems. Consider the adaptive lasso,

$$\min_{\boldsymbol{\beta}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \lambda\|\widehat{w} * \boldsymbol{\beta}\|_1, \tag{3.14}$$

where $\widehat{\boldsymbol{w}}$ is the weight vector with j^{th} component $\widehat{w}_j = 1/|\widehat{\beta}_j|$, with the choice of $\widehat{\beta}$ discussed in Section 3.2. The computational details for minimizing (3.14) are as follows,

- define a new design matrix $\widetilde{\boldsymbol{X}}$, where the j^{th} column $\widetilde{x}_j = x_j/\widehat{w}_j, j = 1, 2, \cdots, p$, for x_j, the j^{th} column of the original matrix $\boldsymbol{X}$.
- solve the lasso problem for a given λ,

$$\boldsymbol{\beta}^* = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \|\boldsymbol{y} - \widetilde{\boldsymbol{X}}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|_1. \tag{3.15}$$

- back-transform $\beta_j = \beta_j^*/\widehat{w}_j, j = 1, 2, \cdots, p$.

The quantity $\boldsymbol{\beta}^*$ in (3.15) can be solved exactly by the APG method discussed in Section 3.1. Further, to solve the adaptive robust lasso with either Huber's or Tukey's biweight criterion (i.e. the Tukey-lasso), similar algorithms can be

applied. The only adjustment that should be made is to the loss functions in (3.15).

We find that the APG algorithm for solving the Tukey-lasso is significantly faster than the CVX algorithm, a MATLAB package developed by Grant et al. (2008). CVX is implemented in both Owen (2007) and Lambert-Lacroix and Zwald (2011) for solving the robust lasso with Huber's loss. We also find that the APG algorithm is significantly faster than the *sparseLTS()* function and *rlars()* function from the R package *robustHD* (Alfons, 2014) for solving Sparse-LTS and Rlars respectively.

3.4 Choice of tuning parameters

We now consider how to select the optimal value of the tuning parameter in lasso-type problems. In lasso problems, a typical method is to compute the entire solution path for a sequence of values of λ and then select the λ value that provides the smallest cross-validation (CV) error. As stated in Friedman et al. (2010), the sequence begins with $\lambda_{max} = \max_j |\langle x_j, y\rangle|/N$, the smallest value of λ for which the entire vector $\widehat{\boldsymbol{\beta}} = \mathbf{0}$. In our study, we follow the idea of using bivariate winsorization as in Khan et al. (2007) and replace $|\langle x_j, y\rangle|$ by its truncated counterpart. We then apply a similar method to Friedman et al. (2010) to select a grid of values for λ : select the maximum value of λ, λ_{max}, set a minimum value $\lambda_{min} = \epsilon\lambda_{max}$, and finally construct a sequence of K values of λ increasing from λ_{min} to λ_{max} on the log scale.

Typical values are $\epsilon = 0.001$ and $K = 100$. After a grid of values of λ is chosen, the APG method is used to produce the entire solution path of the lasso. However, instead of using the conventional CV criteria, we use the classical BIC (Schwarz et al., 1978) and robust BIC criteria (Machado, 1993; Konishi and Kitagawa, 1996) to determine the optimal λ. Given that we consider datasets containing outliers, the traditional CV criteria are not recommended because they are sensitive to outliers. Moreover, BIC criteria require less computation

than the traditional CV. Hence, for the traditional lasso and the adaptive lasso, we select the value of λ that minimizes the classical BIC,

$$BIC(\lambda_k) = \log\left[\sum_{i=1}^{n}\left\{y_i - X_i^T\widehat{\boldsymbol{\beta}}(\lambda_k)\right\}^2\right] + \log(n)p_{\lambda_k}, \tag{3.16}$$

where λ_k is the k^{th} value of the λ in the sequence, $\widehat{\boldsymbol{\beta}}(\lambda_k)$ is the lasso estimates solved by algorithm (3.10) when $\lambda = \lambda_k$, and p_{λ_k} is the number of non-zero coefficients in $\widehat{\boldsymbol{\beta}}(\lambda_k)$.

For the robust versions of the lasso, we select the value of λ that minimizes a generalized robust BIC with the following structure,

$$RBIC(\lambda_k) = L(\widehat{\boldsymbol{\beta}}(\lambda_k), \widehat{\sigma}) + \log(n)p_{\lambda_k}, \tag{3.17}$$

where $L(\widehat{\boldsymbol{\beta}}(\lambda_k), \widehat{\sigma})$ is a robust version of the loss. For the LAD-Lasso, we set $\widehat{\sigma}$ equal to 1 and consider $L(\widehat{\boldsymbol{\beta}}(\lambda_k), \widehat{\sigma}) = 2n\log\left(\sum_{i=1}^{n}\left|y_i - X_i^T\widehat{\boldsymbol{\beta}}(\lambda_k)\right|\right)$, where $\widehat{\boldsymbol{\beta}}(\lambda_k)$ is the LAD-Lasso estimate at $\lambda = \lambda_k$. Following Ronchetti (1985), for the robust lasso with Huber's loss, we use $L(\widehat{\boldsymbol{\beta}}(\lambda_k), \widehat{\sigma}) = 2\sum_{i=1}^{n}\rho_c\left(\frac{y_i - X_i^T\widehat{\boldsymbol{\beta}}(\lambda_k)}{\widehat{\sigma}}\right)$, where $\widehat{\sigma}$ is a robust estimate of σ and $\widehat{\boldsymbol{\beta}}(\lambda_k)$ here is the lasso estimator with Huber's loss at $\lambda = \lambda_k$. Similarly, for the robust lasso with Tukey's biweight, we consider $L(\widehat{\boldsymbol{\beta}}(\lambda_k), \widehat{\sigma}) = 2\sum_{i=1}^{n}\rho_d\left(\frac{y_i - X_i^T\widehat{\boldsymbol{\beta}}(\lambda_k)}{\widehat{\sigma}}\right)$, where $\widehat{\boldsymbol{\beta}}(\lambda_k)$ is the Tukey-lasso estimator when $\lambda = \lambda_k$.

3.5 Simulation results

3.5.1 Simulations for $p < n$

Recall the linear regression model (3.1),

$$y_i = X_i^T\boldsymbol{\beta} + \epsilon_i, \quad i = 1, 2, \ldots, n.$$

In our simulation study for $p < n$, we set $p = 10$, the first column of the design matrix is a $\mathbf{1}_n$ vector and the next 10 columns are filled by covariates generated from the multivariate normal distribution with mean 0 and

$$Cov(x_{ij}, x_{il}) = \rho^{|j-l|}, \quad 1 \leq j, l \leq 10, \tag{3.18}$$

where $\rho = 0.5$. In addition, we define the true model using the first five variables,

$$y_i = x_{i1} + x_{i2} + x_{i3} + x_{i4} + x_{i5} + \epsilon_i. \tag{3.19}$$

Hence, the true model is in the form of equation (3.1) with

$$\boldsymbol{\beta} = (0, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0),$$

and $\epsilon_i \sim N(0, 1)$. That is, the first five regression covariates are significant variables and the rest are noise.

The purpose of our simulation study is to compare the finite sample performance of the lasso estimators considered in Section 3.2 and to determine the relative performance of the Tukey-lasso compared to the other lasso-type estimators, namely, the lasso in Tibshirani (1996), the adaptive lasso in Zou (2006), the LAD-Lasso in Wang et al. (2007), the robust lasso with Huber's loss in Owen (2007) and Lambert-Lacroix and Zwald (2011), the Sparse-LTS in Alfons et al. (2013) and the Rlars in Khan et al. (2007). Therefore, datasets containing outliers in the covariates (x-outliers) and/or the responses (y-outliers) need to be considered. In our simulations, we investigate the following three types of simulated data:

- Scenario 1: Data without any outliers: the design matrix $\mathbf{X}$ is generated as above, and the response y was generated from the true model (3.19).

- Scenario 2: Data with y-outliers only: again, the design matrix $\mathbf{X}$ is generated as above. However, when we generate the response y according to

(3.19), the error (ϵ_i) distribution is taken to be a mixture of normal distribution,

$$(1-\varepsilon)N(0,1)+\varepsilon N(10,1), \tag{3.20}$$

where ε denotes the percentage of outliers and we set $\varepsilon = 10\%$.

- Scenario 3: Data with both x-outliers and y-outliers: the original design matrix $\mathbf{X}$ is generated as described above. We artificially add 10% x-outliers from the $N(10,1)$ distribution and denote the new design matrix by $\mathbf{X}^*$. We use the original design matrix $\mathbf{X}$ to generate the response $\boldsymbol{y}$ according to (3.19), given that we require "bad" leverage points that have a larger effect on the regression estimation. Further, the error (ϵ_i) distribution here is the mixture of normal distributions as in (3.20).

For each type of the data, the design matrix is fixed over all simulations. Then, 100 samples of size n=100 and 200 are generated for each of the three scenarios. Additionally, test samples of size n are simulated independently from the true model (3.20), using the original design matrices for each training sample. That is, the test samples do not contain any x-outliers or y-outliers, given that we focus on the prediction errors for the majority of the data. For each of the training samples, we compute the entire lasso solution path and choose the optimal λ as discussed in Section 3.4. The performances in terms of prediction error and variable selection accuracy are measured for the lasso estimators at their optimal λ values. For Rlars, following Alfons et al. (2013), we fit robust MM-regression (Yohai, 1987) for the sequenced variables. The optimal model is chosen according to BIC implemented with a robust scale estimate as discussed in Alfons et al. (2013).

Variable selection accuracy is measured by the number of correctly identified significant variables (No. correct), the number of included noise variables (No. incorrect) and the percentage of correctly fitted models (Correctly fitted). The percentage of correctly fitted models is the proportion of times that the selected

model includes all significant variables and excludes all noise variables over 100 simulations. Prediction accuracy is measured by the mean squared prediction error (MSPE) $n^{-1}\sum_{i=1}^{n}(\widehat{y}_i - y_i)^2$, computed over a set of independent test samples with the same sample size n as the training sample. We conduct 100 repeated simulations and compute the average, the median and the standard error of the MSPEs.

Table 3.1 presents the detailed simulation results for the data generated by Scenarios 1, 2, and 3, using $p = 10$, and $n = 100, 200$. In Scenario 1, from the perspective of variable selection, all methods tend to identify the five significant variables correctly; ad-Lasso (adaptive lasso), H-Lasso (robust lasso with Huber's criterion and lasso penalty) and T-Lasso (robust lasso with Tukey's biweight and lasso penalty, an abbreviation of Tukey-lasso) generally select fewer noise variables and achieve higher selection probabilities than the other methods. When the sample size is large ($n = 200$), as implied by the oracle property, the adaptive penalty term works well, the methods with the adaptive penalty terms (ad-Lasso, LAD-Lasso, H-Lasso and T-Lasso) include fewer insignificant variables, while S-LTS (the Sparse-LTS) and Rlars are more likely to over-fit. From the prediction point of view, as can be seen in the right-hand panel of Table 3.1 or in Figure 3.1, all methods maintain satisfactory MSPE, while Lasso and S-LTS obtain slightly higher average MSPEs due to the lack of adaptive weights.

From Scenario 2 in Table 3.1, it is clear that Lasso and ad-Lasso exhibit very poor performance in terms of both variable selection and prediction accuracy. This is due to the failure of quadratic loss in the presence of y-outliers. In contrast, the remainder of the robust methods are confirmed to be robust to y-outliers and, among them, Rlars and T-Lasso always achieve lower MSPEs with smaller standard deviations. This is also shown in the middle panel of Figure 3.1. Again, when the sample size increases, the adaptive penalty term in T-Lasso becomes more effective in variable selection and prediction accuracy.

Finally, the bottom panel in Table 3.1 reports simulation results for Scenario

Table 3.1: Simulation results for the data generated by Scenarios 1, 2, and 3, using $p = 10, n = 100, 200$

Scenario	Method	No. correct	No. incorrect	Correctly fitted	Average MSPE	Median MSPE	SD MSPE
1, n=100	Lasso	5.00	0.69	0.49	1.37	1.37	0.25
	ad-Lasso	5.00	0.19	0.84	1.17	1.18	0.21
	LAD-Lasso	5.00	0.54	0.63	1.19	1.21	0.24
	H-Lasso	5.00	0.39	0.71	1.16	1.17	0.19
	S-LTS	5.00	2.96	0.25	1.44	1.42	0.32
	Rlars	5.00	0.87	0.59	1.13	1.13	0.19
	T-Lasso	5.00	0.34	0.75	1.18	1.17	0.20
1, n=200	Lasso	5.00	0.58	0.61	1.10	1.09	0.15
	ad-Lasso	5.00	0.11	0.91	0.97	0.97	0.13
	LAD-Lasso	5.00	0.32	0.77	0.99	0.98	0.17
	H-Lasso	5.00	0.13	0.90	0.96	0.97	0.13
	S-LTS	5.00	2.31	0.49	1.19	1.22	0.24
	Rlars	5.00	0.73	0.66	0.95	0.94	0.13
	T-Lasso	5.00	0.09	0.94	0.96	0.95	0.14
2, n=100	Lasso	4.91	1.04	0.38	2.88	2.79	1.29
	ad-Lasso	4.39	0.49	0.34	3.11	2.82	1.53
	LAD-Lasso	5.00	0.16	0.85	1.26	1.24	0.32
	H-Lasso	5.00	0.50	0.66	1.22	1.19	0.31
	S-LTS	5.00	2.60	0.23	1.40	1.36	0.35
	Rlars	4.98	0.62	0.67	1.12	1.06	0.25
	T-Lasso	5.00	0.18	0.83	1.16	1.12	0.25
2, n=200	Lasso	5.00	1.59	0.14	2.60	2.53	0.71
	ad-Lasso	4.79	0.46	0.49	2.72	2.71	0.87
	LAD-Lasso	5.00	0.06	0.94	1.05	1.04	0.21
	H-Lasso	5.00	0.19	0.84	1.05	1.05	0.18
	S-LTS	5.00	1.45	0.49	1.19	1.17	0.22
	Rlars	5.00	0.15	0.87	0.94	0.92	0.15
	T-Lasso	5.00	0.08	0.92	0.97	0.95	0.15
3, n=100	Lasso	4.84	1.92	0.10	4.30	4.20	1.11
	ad-Lasso	3.85	0.47	0.08	4.81	4.72	1.45
	LAD-Lasso	4.80	0.35	0.62	2.23	2.02	0.88
	H-Lasso	4.96	0.92	0.38	2.23	2.14	0.70
	S-LTS	5.00	2.38	0.22	1.36	1.30	0.33
	Rlars	5.00	0.45	0.75	1.12	1.09	0.23
	T-Lasso	5.00	0.12	0.89	1.18	1.15	0.25
3, n=200	Lasso	4.98	2.37	0.07	4.73	4.65	0.73
	ad-Lasso	4.51	0.44	0.34	5.00	4.85	0.84
	LAD-Lasso	4.67	0.12	0.60	1.96	1.89	0.44
	H-Lasso	4.72	0.28	0.56	2.19	2.16	0.40
	S-LTS	5.00	1.24	0.44	1.19	1.17	0.19
	Rlars	5.00	0.13	0.91	0.98	0.96	0.16
	T-Lasso	4.95	0.04	0.91	1.06	1.00	0.24

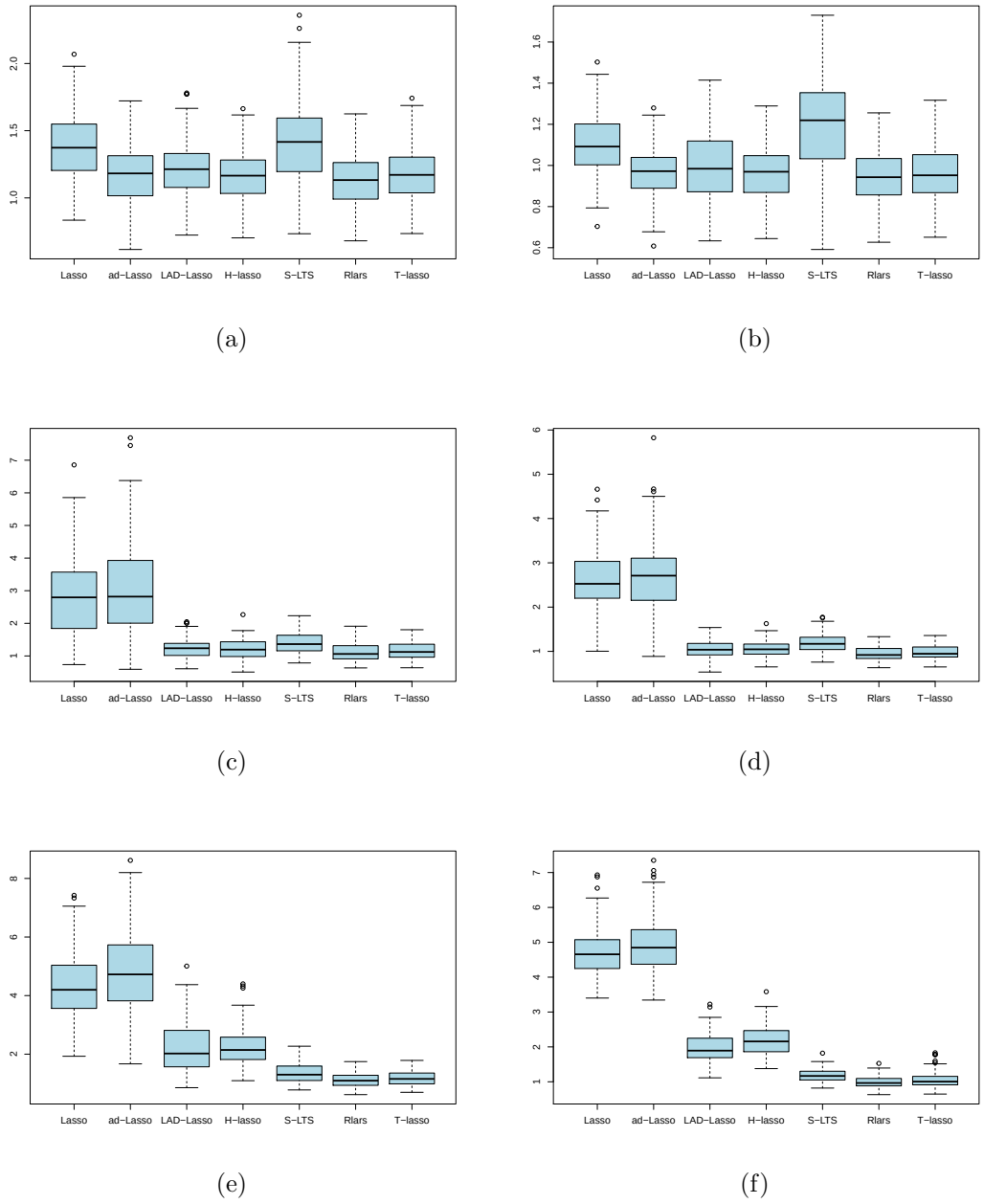


Figure 3.1: Mean squared predication errors for $p = 10$. (a) Scenario 1 with $n = 100$, (b) Scenario 1 with $n = 200$, (c) Scenario 2 with $n = 100$, (d) Scenario 2 with $n = 200$, (e) Scenario 3 with $n = 100$, and (f) Scenario 3 with $n = 200$.

3 (the Scenario with both x- and the y-outliers). Again, the performance of Lasso and ad-Lasso suffers greatly from the outliers. When $n = 100$, T-Lasso significantly outperforms all the other methods in variable selection accuracy and achieves a desirable selection probability of 89 %. From the perspective of prediction, it is worth noting that in the presence of x-outliers, LAD-Lasso and H-Lasso have a far inferior performance than the other three robust methods. In contrast, S-LTS, Rlars and T-Lasso maintain desirable MSPEs and are shown to be resistant to both x-outliers and y-outliers, as expected. As evidenced in the bottom panel of Figure 3.1, these three methods consistently dominate the others by achieving much lower MSPEs. When $n = 100$, the average MSPE for T-Lasso or Rlars is nearly half of that of LAD-Lasso and H-Lasso. This strongly illustrates the usefulness of T-Lasso in the presence of both x-outliers and y-outliers.

In Scenario 3, we set the contamination level for both the covariates and the response to 10%. To illustrate the behavior of these robust methods at other contamination levels, we present Figure 3.2, which shows the averaged MSPEs for all methods at various contamination levels for both the covariates and the response, when $n = 100$ and $p = 10$. As stated, Lasso and ad-Lasso are not resistant to any percentage of outliers; the average MSPE increases dramatically when the contamination level increases. The MSPEs for LAD-Lasso and H-Lasso gradually increase after 5% contamination. Further, Rlars and T-Lasso demonstrate consistent outperformance over all methods, their averaged MSPEs are almost equal and maintain a satisfactory level even when the contamination level for both the covariates and the response reaches 25%.

Overall, S-LTS, Rlars and T-Lasso generally outperform the other lasso-type methods in terms of prediction accuracy, particularly in the presence of both x-outliers and y outliers. However, from the perspective of variable selection, T-Lasso is strongly preferred because it achieves very satisfactory selection probabilities in all scenarios, as suggested by its oracle property. S-LTS in Alfons et al. (2013) and Rlars in Khan et al. (2007) are not shown to enjoy the oracle

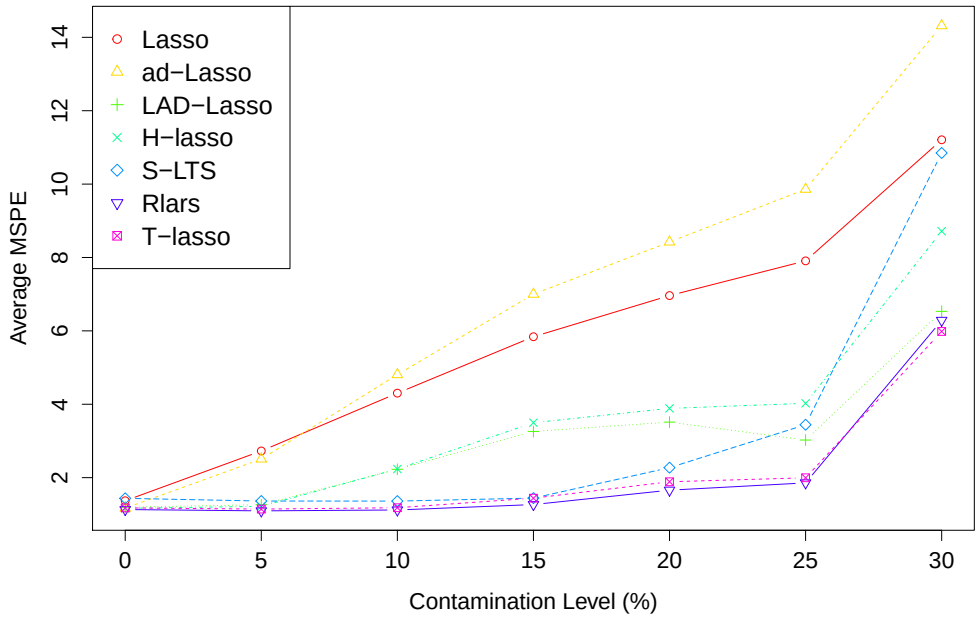


Figure 3.2: Average mean squared predication errors at various contamination levels for simulation Scenario 3 with $n = 100$ and $p = 10$.

property.

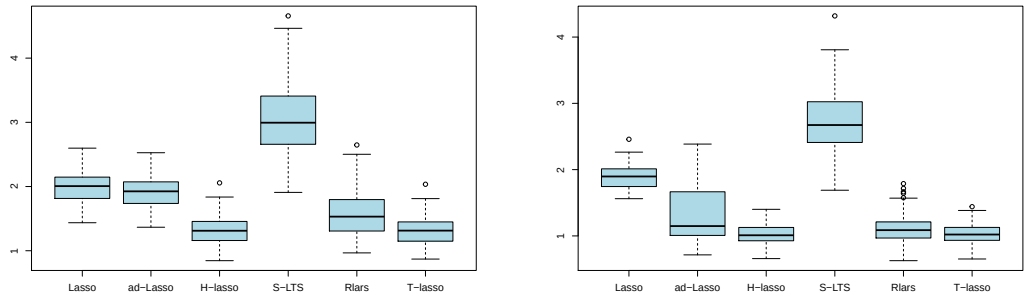
3.5.2 Simulations for $p > n$

This section reports simulations for a high dimensional case with $p > n$. The simulation settings are almost identical to those of $p < n$. The only difference is we now consider the covariate dimension $p = 300$ instead of $p = 10$. Accordingly, the true model is still in the form of equation (3.1), but now with $\boldsymbol{\beta} = (0, 1, 1, 1, 1, 1, 0, \ldots, 0)$. The first five are important variables and the rest are noise. Note, similar to Alfons et al. (2013), we no longer include LAD-Lasso in the simulations due to its expensive computation and poor performance. In this case of $p > n$, we adopt non-robust ridge estimates to compute adaptive weights for ad-Lasso. For T-Lasso and H-Lasso, we compute the adaptive weights by MM-Ridge as noted in Section 3.2.

Table 3.2 contains the simulation results for the data generated under Scenario

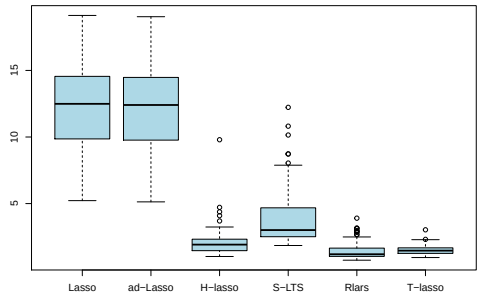
Table 3.2: Simulation results for the data generated by Scenarios 1, 2, and 3, using $p = 300, n = 100, 200$

Scenario	Method	No. correct	No. incorrect	Correctly fitted	Average MSPE	Median MSPE	SD MSPE
1, n=100	Lasso	5.00	109.34	0.00	2.01	2.01	0.22
	ad-Lasso	5.00	87.15	0.01	1.93	1.93	0.22
	H-Lasso	5.00	0.04	0.96	1.32	1.31	0.23
	S-LTS	5.00	68.59	0.00	3.03	2.99	0.56
	Rlars	5.00	14.52	0.02	1.57	1.53	0.35
	T-Lasso	5.00	0.02	0.98	1.32	1.31	0.23
1, n=200	Lasso	5.00	210.09	0.00	1.90	1.90	0.16
	ad-Lasso	5.00	50.94	0.58	1.30	1.15	0.41
	H-Lasso	5.00	0.31	0.80	1.03	1.01	0.15
	S-LTS	5.00	143.35	0.00	2.72	2.67	0.47
	Rlars	5.00	8.39	0.10	1.13	1.09	0.23
	T-Lasso	5.00	0.23	0.83	1.03	1.02	0.15
2, n=100	Lasso	4.67	115.96	0.00	12.39	12.49	3.13
	ad-Lasso	4.89	98.61	0.00	12.30	12.40	3.13
	H-Lasso	4.93	1.13	0.61	2.06	1.92	1.03
	S-LTS	4.87	68.79	0.00	3.88	3.01	2.10
	Rlars	4.69	2.15	0.36	1.45	1.20	0.66
	T-Lasso	4.98	0.04	0.94	1.50	1.47	0.34
2, n=200	Lasso	4.94	232.83	0.00	12.42	12.75	2.48
	ad-Lasso	4.95	200.67	0.00	12.36	12.69	2.48
	H-Lasso	5.00	0.35	0.75	1.39	1.36	0.30
	S-LTS	4.99	143.41	0.00	3.42	2.95	1.22
	Rlars	4.99	1.63	0.54	0.99	0.95	0.24
	T-Lasso	5.00	0.04	0.96	1.18	1.16	0.19
3, n=100	Lasso	4.74	116.03	0.00	12.27	12.13	3.26
	ad-Lasso	4.81	99.37	0.00	12.11	11.82	3.26
	H-Lasso	4.82	1.82	0.31	2.51	2.40	0.78
	S-LTS	4.63	69.09	0.00	4.97	3.83	3.03
	Rlars	4.81	2.28	0.37	1.49	1.16	1.77
	T-Lasso	4.90	0.10	0.81	1.78	1.65	0.58
3, n=200	Lasso	4.93	232.51	0.00	13.50	13.71	2.71
	ad-Lasso	4.95	202.29	0.00	13.24	13.45	2.66
	H-Lasso	5.00	0.84	0.54	1.63	1.54	0.42
	S-LTS	4.98	143.47	0.00	4.08	3.61	1.57
	Rlars	4.99	0.93	0.67	0.98	0.94	0.24
	T-Lasso	5.00	0.03	0.97	1.29	1.25	0.27

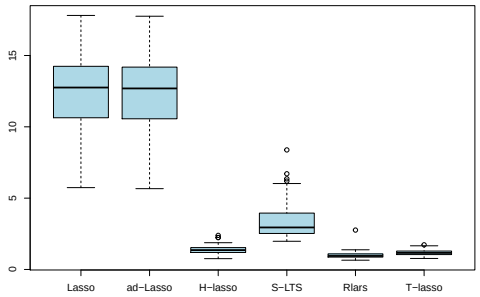


(a)

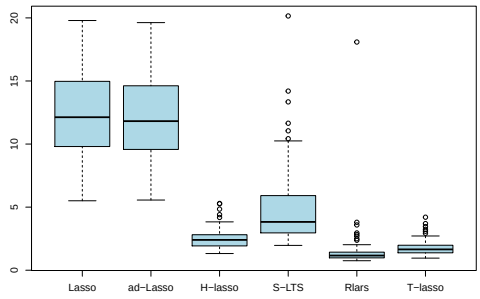
(b)



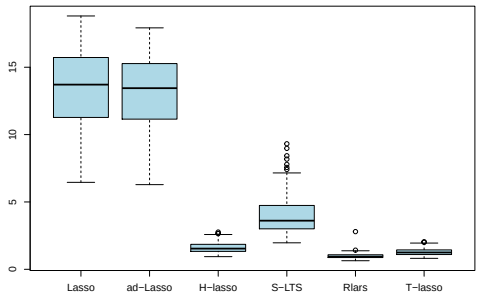
(c)



(d)



(e)



(f)

Figure 3.3: Mean squared predication errors for $p = 300$. (a) Scenario 1 with $n = 100$, (b) Scenario 1 with $n = 200$, (c) Scenario 2 with $n = 100$, (d) Scenario 2 with $n = 200$, (e) Scenario 3 with $n = 100$, and (f) Scenario 3 with $n = 200$.

1, 2, and 3, using $p = 300, n = 100, 200$. In Scenario 1, S-LTS obtains the highest average MSPE because it suffers from efficiency problems when the data contain no outliers. It is clear that H-Lasso and T-Lasso achieve outstanding performance in variable selection and prediction accuracy, while all other methods have an over fitting problem. In addition, Rlars, which we identified as the main competitor of T-Lasso when $p = 10$, shows very poor performance in variable selection in this scenario but obtains a slightly lower MSPE than T-Lasso. With vertical outliers added in Scenario 2, it is clear that the prediction power of Lasso and ad-Lasso breaks down. Similar to the case of $p = 10$, Rlars and T-Lasso in the high dimensional case still maintain substantially lower MSPEs even when outliers in the covariates are introduced in Scenario 3. This is also shown in Figure 3.3

To summarize, for both cases of $p = 10$ and $p = 300$, T-Lasso demonstrates superior performance to the adaptive lasso and other robust methods in terms of both variable selection and prediction accuracy.

3.5.3 Computation time

Another important advantage of the Tukey-lasso is its computational efficiency via the APG method. Table 3.3 presents the computation times of Lasso, H-Lasso, S-LTS, Rlars, and T-Lasso based on a simulated data with $n = 200$ and varying p. The computations are performed on an Intel Xeon W3540. The lasso is computed by the *glmnet()* function in the R package *glmnet* (Friedman et al., 2009) and H-Lasso by the CVX package in MATLAB (Grant et al., 2008). Computations for S-LTS and Rlars are performed using the *sparseLTS()* function and *rlars()* function from the R package *robustHD* (Alfons, 2014), respectively. T-Lasso is solved by the APG algorithm as noted in Section 3.3, implemented in MATLAB. For these lasso type methods (e.g. Lasso, H-Lasso, S-LTS, and T-Lasso), the reported times are averaged over a grid value of five values of λ. For Rlars, we set the number of covariates to be sequenced as $n/2$ (the default setting in *rlars*), which is 100 in this case. We feel that sequencing a larger number of covariates is

not necessary because sparsity is usually assumed. All methods apply a tolerance level for convergence at 10^{-6}. This is also the tolerance level we used in the simulation study. For T-Lasso, we set the maximum number of iterations to 10^4 to ensure convergence. According to (3.11), on average, using a tolerance of 10^{-6} requires 10^3 iterations to reach convergence, so 10^4 is considered to be a very conservative choice.

Table 3.3: Computation times (in seconds) for simulated data with $n = 200$ and varying covariate dimension p

p	ad-Lasso	H-Lasso	S-LTS	Rlars	T-Lasso
100	0.002	1.915	1.594	80.060	0.122
500	0.002	3.193	32.068	81.090	0.219
1000	0.004	4.954	130.208	80.950	0.182
2000	0.006	8.622	519.780	85.180	1.026
3000	0.008	14.489	1180.594	85.470	2.604

Table 3.3 demonstrates that S-LTS is reasonably fast until p reaches 1000. The computation time for H-Lasso slightly increases as p increases. For Rlars, the computation time is relatively large, but fixed, for all p. However, if a full sequence of covariates is required, instead of a fixed number $n/2$, the computation time for Rlars will dramatically increase with p. In contrast, it is worth noting that T-Lasso remains fast for increasing p, which is even comparable with the non-robust Lasso fitted by *glmnet*.

To conclude, T-Lasso not only achieves outstanding performance in terms of both variable selection and prediction accuracy, but it is also faster to compute.

3.6 Real data examples

3.6.1 Example 1: Earnings forecasting in Chinese stock market

In this section we investigate an earnings forecast study using Chinese stock market data from Wang et al. (2007) and Lambert-Lacroix and Zwald (2011).

The dataset is derived from China Center for Economic Research (CCER) China stock, which was developed by the CCER at Peking University. The dataset contains a total of 2247 records, with each record corresponding to a one-yearly observation of a company. Among these records, 1092 are from the year 2002 and we consider these records as the training data as in Wang et al. (2007) and Lambert-Lacroix and Zwald (2011). The remaining 1155 observations come from 2003 and serve as the test data. The response variable is the return on equity (ROE) (i.e. earnings divided by total equity) of the following year (denoted by ROE_{t+1}). Eight explanatory variables are all measured at year t; they include the ROE of the current year (ROE_t), asset turnover ratio (ATO), profit margin (PM), debt-to-asset ratio or leverage (LEV), sales growth rate ($GROWTH$), price-to-book ratio (PB), account receivables/revenues (ARR), inventory/asset (INV), and the logarithm of total assets ($ASSET$).

Various lasso-type methods, were used to select the best model from the training dataset from 2002. The prediction accuracy of these methods was evaluated on the test data. Wang et al. (2007) and Lambert-Lacroix and Zwald (2011) considered only the mean absolute prediction error (MAPE) as a measure of the prediction accuracy. However, we observe that for all the methods considered, there always exists several extremely large prediction errors in the test data after the models are fitted. These residuals dominate the measure of the prediction accuracy, even if MAPE is used. To be robust, a good model should capture the pattern of the majority of data. Therefore, we used the trimmed mean square prediction error (TMSPE), as a more appropriate measure of the prediction accuracy for this dataset. We truncated the largest 10 % of squared residuals and computed the TMSPE using the remaining 90% of the squared residuals. TMSPE is a measure of prediction accuracy for the majority of the data and is no longer dominated by extreme prediction errors. The estimation results are summarized in Table 3.4. Note that for all the real data examples, we set the tolerance level for convergence as 10^{-10} and correspondingly set the maximum number of iteration

to 10^6 to ensure convergence.

Our results are similar to those in Wang et al. (2007) and Lambert-Lacroix and Zwald (2011). As Table 3.4 demonstrates, the MAPE and TMSPE for ad-Lasso are substantially worse than they are for the robust methods, which indicates the superiority of the robust methods for this dataset. Among the three robust methods, H-Lasso obtains the lowest MAPE of 0.12081. T-Lasso achieves the best predictive performance with the lowest TMSPE of 0.00196. This means that T-Lasso tends to produce the most accurate prediction for the majority of data.

For comparison purposes, in Table 3.5, we also present the estimation results for the model selected by traditional robust model selection using BIC. For example, RBIC.M indicates the model that minimizes robust BIC in (3.17) with ρ function given by Huber's loss. Table 3.5 demonstrates that the best models selected by various versions of BIC perform equally well with their counterparts in Table 3.4. It is also worth noting that the best model selected by RBIC.MM coincides with the model selected by Rlars. However, traditional model selection is computationally intensive when there are a large number of variables and computationally infeasible when the number of variables is greater than the sample size. Therefore, we strongly prefer robust lasso methods such as T-Lasso due to their computational efficiency. Given that the traditional model selection procedure and these lasso-type methods produce similar estimation results, we do not present the results for the model selected by traditional robust model selection in the following two examples.

3.6.2 Example 2: Boston housing data

We then applied robust lasso methods to analyze the well-known Boston housing data, which is available on `http://lib.stat.cmu.edu/datasets/boston`. The data have been described in Section 2.5. In our study, the first 300 observations were treated as the training data and the remaining 206 observations were treated as the test data. Given that we focused on prediction for the majority of the data,

Table 3.4: Estimation results of the earnings forecast study

Variable	ad-Lasso	LAD-Lasso	H-Lasso	S-LTS	Rlars	T-Lasso
Intercept	-2.05930	-0.28613	-0.42034	-0.13987	-0.18958	-0.17040
ROE	-0.17733	0.20518	0.14324	0.75410	0.49672	0.50207
ATO	0.15991	0.05896	0.06589	0.01767	0.03141	0.03087
PM	0.18899	0.13878	0.15947	0.05893	0.07946	0.07730
LEV	-0.24504	-0.01950	-0.02892	0.00120		
GROWTH	0.03275	0.01624	0.01773	0.01896	0.01247	0.01210
PB	0.01837		0.00167	0.00121	0.00199	0.00174
ARR		-0.00023	-0.00041	-0.00455	-0.00490	-0.00460
INV	0.34380		0.03065	0.00083		
ASSET	0.09992	0.01302	0.01906	0.00533	0.00796	0.00712
MAPE	0.23242	0.12088	0.12081	0.14288	0.12783	0.12816
TMSPE	0.02432	0.00234	0.00256	0.00232	0.00199	0.00196

Table 3.5: Estimation results for traditional robust model selection using BIC

Variable	BIC	RBIC.LAD	RBIC.M	RBIC.MM
Intercept	-2.08855	-0.31729	-0.43980	-0.18958
ROE	-0.18088	0.20795	0.11701	0.49672
ATO	0.16127	0.05875	0.07471	0.03141
PM	0.19480	0.14038	0.17750	0.07946
LEV	-0.24533	-0.02086	-0.03633	
GROWTH	0.03312	0.01877	0.01972	0.01247
PB	0.01845			0.00199
ARR				-0.00490
INV	0.35080		0.05531	
ASSET	0.10116	0.01454	0.01997	0.00796
MAPE	0.23283	0.12143	0.12410	0.12783
TMSPE	0.02458	0.00239	0.00298	0.00199

we measured the prediction accuracy by the TMSPE. We included the MAPE for comparison with the previous example.

Table 3.6: Estimation results of study of Boston housing data

Variable	ad-Lasso	LAD-Lasso	H-Lasso	S-LTS	Rlars	T-Lasso
Intercept	-12.598	-12.481	-15.543	-0.476	-16.067	-14.719
crim	0.749			0.199		
zn				0.022		
indus				-0.014		
chas				0.546		
nox	-6.315			-3.530		
rm	9.286	8.815	9.262	6.443	9.145	9.414
age	-0.048	-0.052	-0.053	-0.044	-0.054	-0.060
dis	-0.874	-0.704	-0.660	-0.847	-0.675	-0.560
rad				0.254		
tax	-0.012	-0.009	-0.011	-0.012	-0.012	-0.011
ptratio	-0.699	-0.666	-0.652	-0.481	-0.636	-0.624
black	0.014	0.008	0.008	0.012	0.012	
lstat	-0.094	-0.135	-0.069	-0.207	-0.077	
MAPE	7.743	4.893	5.055	5.693	4.972	5.308
TMSPE	45.377	15.783	16.626	28.494	16.650	19.181

Table 3.6 compares the regression coefficients and prediction accuracy for all the lasso methods and Rlars. Ad-Lasso has the largest MAPE and TMSPE, clearly indicating the usefulness of the robust methods in the presence of outliers. Among the robust adaptive lasso methods, T-Lasso selected the simplest model with five variables (rm, age, dis, tax, ptratio) and LAD-Lasso achieved the lowest MAPE and TMSPE. Additionally, S-LTS did not perform a great deal of shrinkage in this dataset. All other robust methods obtained satisfactory prediction accuracy.

To investigate the influence of x-outliers on the robust lasso methods, we contaminated the data by adding 5 % outliers to the three significant variables (rm, age, dis). Figure 3.4 presents the pairs plots for these three variables (rm, age dis) before and after the contamination. Table 3.7 summarizes the results for the Boston housing data with contaminated covariates. Clearly, the coefficients esti-

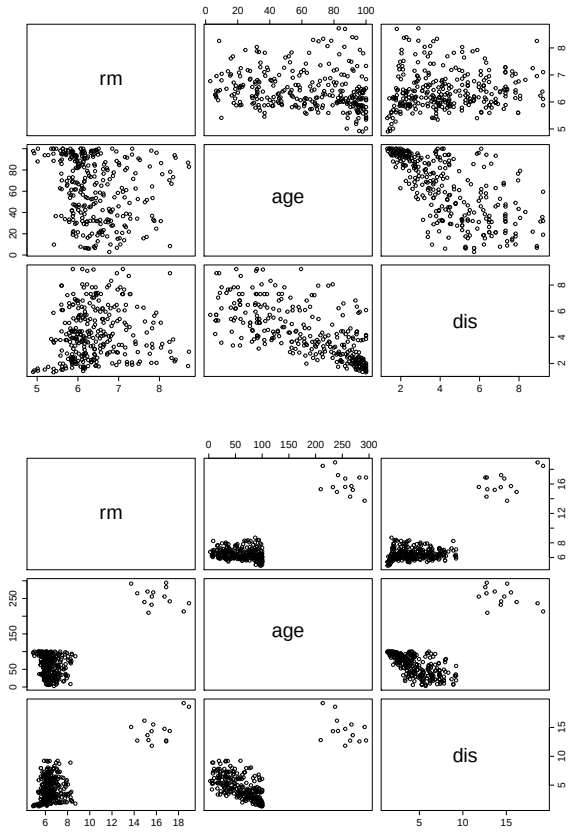


Figure 3.4: Pairs plots for the variables (rm, age dis) before (left) and after (right) contamination

Table 3.7: Estimation results of study of Boston housing data with contamination on variables (rm, age, dis)

Variable	ad-Lasso	LAD-Lasso	H-Lasso	S-LTS	Rlars	T-Lasso
Intercept	37.673	30.082	33.263	-3.476	-14.955	-14.446
crim	1.387	0.627	0.645	0.214		
zn	0.056	0.081	0.068	0.022		
indus		-0.079	-0.073	-0.019		
chas				0.375		
nox	-17.024	-10.735	-11.883	-2.565		
rm	4.030	3.764	3.823	6.867	8.951	8.987
age	-0.061	-0.050	-0.058	-0.046	-0.046	-0.048
dis	-2.305	-2.127	-2.192	-0.885	-0.652	-0.579
rad	0.547	0.462	0.536	0.281		
tax	-0.023	-0.020	-0.022	-0.012	-0.012	-0.012
ptratio	-0.822	-0.524	-0.653	-0.498	-0.647	-0.640
black	0.017	0.012	0.013	0.013	0.013	0.009
lstat	-0.427	-0.383	-0.343	-0.177	-0.121	-0.089
MAPE	14.680	8.823	9.490	6.058	4.929	4.973
TMSPE	161.420	66.135	76.884	33.347	16.479	16.112

mated using the LAD loss (LAD-Lasso) and Huber's loss (H-Lasso) were heavily affected by the covariate outliers and the prediction properties of these methods break down. In contrast, the Rlars and T-Lasso estimators barely changed and T-Lasso still achieved the lowest TMSPE with the simplest model. Compared to all the other robust lasso methods, T-Lasso is highly competitive.

3.6.3 Example 3: Glioblastoma gene expression data

For a further illustration of T-Lasso, we investigated a real dataset where $p > n$. We analyze the glioblastoma gene expression data, which was originally studied by Horvath et al. (2006), and further investigated in Wang et al. (2011). The glioblastoma data are from two independent sets of clinical tumor samples of $n = 55$ and $n = 65$ with expression values for $p = 3600$ genes, available from `https://labs.genetics.ucla.edu/horvath/CoexpressionNetwork/ASPMgene/`. Details of these data can be found in Horvath et al. (2006). Similar to Wang et al.

(2011), before analyzing these data we exclude nine censored subjects, five from the first set of 55 patients and four from the second set of 65 patients, and use the logarithm of time to death as the response. Then, the first data set serves as the training set with $n = 50, p = 3600$ and the second set as the test set with $n = 61, p = 3600$.

Table 3.8 presents the glioblastoma gene expression data analysis and compares the performance of T-Lasso with the other lasso methods. We still measure the prediction accuracy by the TMSPE and the MAPE. The number of variables selected by each method is also reported. The lasso selects 50 variables and achieves the highest MAPE and TMSPE. It seems that robust methods are more favorable, excepting Rlars. Among the robust lasso methods, T-Lasso significantly outperforms the other methods by achieving the lowest MAPE and TMSPE, while identifying only a small number of variables in the model. Overall, T-Lasso has obvious advantages over other methods in the analysis of the glioblastoma data.

Table 3.8: Glioblastoma gene expression data analysis

Method	MAPE	TMSPE	Model Size
Lasso	1.1225	1.2823	51
ad-Lasso	0.7319	0.5246	2
H-Lasso	0.6567	0.3864	10
S-LTS	0.6728	0.3896	37
Rlars	0.9612	0.8047	12
T-Lasso	0.6470	0.3794	6

3.7 Conclusion

In this paper, we propose the Tukey-lasso method, which combines Tukey's biweight loss and the adaptive lasso penalty. The Tukey-lasso is resistant to outliers in both the response and the covariates. Importantly, it also enjoys the oracle property as does the adaptive lasso. Using the APG method, the Tukey-lasso

can be computed very efficiently and rapidly. Our simulation studies demonstrate that the Tukey-lasso compares favorably with the adaptive lasso and other robust implementations of the lasso. Real data examples also support the use of the Tukey-lasso in variable selection and prediction problems.

Chapter 4

Bootstrap Lasso Averaging

4.1 Introduction

Model selection is central to all applied statistical work. Selecting the variables for use in a regression model is one important example of model selection. Over the past two decades, a number of different model selection approaches have been rapidly developed, ranging from the widely recognized AIC and BIC to more recent methods such as the lasso, elastic net, and SCAD (Tibshirani, 1996; Zou and Hastie, 2005; Fan and Li, 2001). These traditional approaches assume that the parameter values of the best model can be estimated, and that, thereafter we will make inferences from the data only according to that sole best model. However, for any given data set, the use of a different model selection method may result in a different best model being selected. Conversely, for any given model selection approach, a different best model will likely be chosen if a new data set is analyzed. Often several models fit the data equally well, yet these models may include different explanatory variables and lead to different predictions. This extra component of variation, is often termed as 'model uncertainty'.

To account for model uncertainty, model averaging, which makes inferences based on weighted support from several models instead of a sole best model, has been proposed and developed. Apart from avoiding the inference drawn from the

single best model, model averaging exhibits more accurate predictive performance than reliance on a single model (Raftery et al., 1997).

Inspired by Bayesian considerations, significant work incorporating model uncertainty has been conducted using a Bayesian Model Averaging (BMA) framework. BMA considers that quantities of interest, such as predicted values, can often be expressed as a weighted average of model specific quantities, where the weights depending on how much the data support each model can be measured by the posterior probabilities on the models. Numerous papers have discussed the implementation of BMA including the choice of prior distribution and computational issues (Hoeting et al., 1999; Clyde and George, 2004; Raftery et al., 1997). Most of this literature shows that BMA procedures provide significantly better predictive performances than any single model that might reasonably have been chosen.

Although the development of BMA has been substantial over the past decade, it remains problematic for setting up prior probabilities. As argued in Hjort and Claeskens (2003), the typical application of BMA also involves mixing together many conflicting prior opinions regarding interest parameters. An alternative to BMA is non-Bayesian model averaging, or Frequentist Model Averaging (FMA) (Buckland et al., 1997; Rao and Tibshirani, 1997; Hjort and Claeskens, 2003; Burnham and Anderson, 2003; Yuan and Yang, 2012; Claeskens et al., 2008). The fundamental difference between BMA and FMA is that the weights assigned to each candidate model in FMA are completely determined by the data, while the posterior probabilities in BMA depend on the prior probabilities set up by the user.

Among these papers contributing to model averaging, the most relevant work for our purposes is bootstrap model averaging first proposed by Buckland et al. (1997). Bootstrap model averaging in Buckland et al. (1997) utilizes the bootstrap to generate resamples, applies the model selection criteria independently to each resample and further computes the weights assigned to each model. More

specifically, Buckland et al. (1997) proposed a bootstrap weighting scheme based on either AIC or BIC that involves four steps: (1) apply model selection to the original data to identify the 'best' model; (2) use the parametric bootstrap based on the model selected in (1) to generate resamples of the original data; (3) for each resample repeat the model selection process of (1); and (4) assign weight to each model equal to the proportion of times that it was selected in (3). Another similar machine learning ensemble technique is called bagging, short for 'bootstrap aggregating' as proposed by Breiman (1996). Under a linear regression framework, Breiman (1996) utilized forward variable selection to identify the best model for each bootstrap resample. However, similar to the traditional model selection procedures by AIC or BIC, bootstrap model averaging (or bagging) is computationally intensive with a large number of variables and computationally infeasible with the number of variables greater than the sample size. For example, when a model is constructed from 20 variables, 2^{20} model combinations are taken into consideration and bootstrapping further exaggerates the computation cost. Augustin et al. (2005) and Buchholz et al. (2008) proposed to include a variable screening step prior to bootstrap model averaging to eliminate variables with negligible effect on the response. This variable screening step results in a much smaller set of candidate models to be considered in the bootstrap step. More variable screening methods can be found in Fan and Lv (2008) and Wang (2009).

In this work, we modify bootstrap model averaging by utilizing the lasso (Tibshirani, 1996) as a model selection tool, instead of the traditional AIC or BIC, to improve the computation speed and realize computational feasibility even when the number of variables p is larger than the sample size n. The lasso for 'Least Absolute Shrinkage and Selection Operator', first proposed by Tibshirani (1996), incorporates a $L1$ penalty into the OLS loss function. The lasso shrinks some coefficients to exactly zero and hence gives parsimonious solutions that are easy to interpret. Many authors (Zou, 2006; Knight and Fu, 2000; Zou

et al., 2007; Zhao and Yu, 2006) have investigated the properties of the lasso and developed different variants of it. Zou (2006) showed that there exist certain scenarios where the lasso is inconsistent for variable selection. He suggested the adaptive lasso, where adaptive weights are used for penalizing coefficients differently in the $L1$ penalty. The adaptive lasso enjoys the oracle property, that is, asymptotically it performs as well as if we knew the true underlying model. Moreover, both the lasso and adaptive lasso can be solved efficiently and fast by the coordinate descent algorithm proposed in Friedman et al. (2010). A number of studies considered the implementation of bootstrap resamples for improving the lasso estimates. Bach (2008) proposed the Bolasso, which intersects with the supports of the lasso bootstrap estimates, to achieve consistent model selection. Hall et al. (2009) suggested using the m-out-of-n bootstrap, to choose the optimal regularization parameter for the adaptive lasso setup. Wang et al. (2011) used a set of randomly selected variables in each bootstrap resample and applied the lasso to remove highly correlated variables altogether or to select them all and called it 'random lasso'. The asymptotic properties of the residual bootstrap for lasso estimators are also discussed in Knight and Fu (2000), Chatterjee and Lahiri (2010), and Chatterjee and Lahiri (2011). However, none of them attempted the model averaging to account for model uncertainty.

Therefore, we propose to apply the lasso to identify the 'best' model in step (1) of bootstrap model averaging as in Buckland et al. (1997). Replacing the model selection criteria of AIC or BIC with the lasso not only improves the computation speed, but also screens the number of variables up to the sample size. This modified Step (1) can be viewed as a variable screening step and further allows for the implementation of the adaptive lasso in Step (3). Although the adaptive lasso is selection consistent under general design conditions, it requires the sample size to be larger than the predictor dimension since the adaptive weights generally rely on a consistent unpenalized estimator, which is generally infeasible for computing in a high dimensional case. Thereafter, we utilize the adaptive lasso to repeat the

model selection process in Step (3) to further achieve the selection consistency. We call this modified bootstrap model averaging, 'bootstrap lasso averaging' (BLA). More details of the algorithms will be provided in Section 4.2. In this work, we mainly focus on the comparison of the predictive performances of the method we propose and the other traditional model averaging methods. Outstanding finite sample predictive performance by BLA is confirmed by extensive simulation studies.

The rest of this paper is organized as follows. In Section 4.2, we introduce three different algorithms to perform BLA. In Section 4.3, we present our simulation settings, show our simulation results and compare the prediction and variable selection accuracy of various procedures of model averaging. We analyse several real examples in Section 4.4. Finally, we present brief conclusions in Section 4.5.

4.2 Bootstrap lasso averaging

Consider a standard linear regression model,

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}^* + \boldsymbol{\epsilon}, \tag{4.1}$$

where $\boldsymbol{X}$ is an $n \times (p+1)$ design matrix with the first column of $\mathbf{1}$, $\boldsymbol{\beta}^* = (\beta_0^*, \beta_1^*, \ldots, \beta_p^*)^T$ is a $p+1$ vector of regression parameters, some of which are zero, and $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \ldots, \epsilon_n)^T$ is an $n \times 1$ vector with n independently distributed random variables whose expected value is zero and whose variance is σ^2. Define the i^{th} row of $\boldsymbol{X}$ by $\boldsymbol{x}_i$, $i = 1, \ldots, n$. Further define the response vector by $\mathbf{y} = (y_1, \ldots, y_n)$. If the intercept (i.e. β_0) is assumed to be included in each candidate model, there are $K = 2^p$ candidate models (denoted by $M_1, \ldots, M_K$) that can be constructed by using subsets of the full set of p variables. The sub design matrix $\mathbf{X}_{(k)}$ and sub vector $\boldsymbol{\beta}^*_{(k)}$ denote the partitions of $\mathbf{X}$ and $\boldsymbol{\beta}^*$ that consist of the variables included in the candidate model $M_k, k = 1, \ldots, K$, respectively. Define $\boldsymbol{u}_{(k)} = \boldsymbol{X}_{(k)}\boldsymbol{\beta}^*_{(k)}$, $\mathcal{K} = (M_1, \ldots, M_K)$ as the whole set of

candidate models; $p_{(k)}$ as the number of variables included in model M_k; and $\widehat{\boldsymbol{\beta}}_{(k)}$ and $\widehat{\sigma}_{(k)}$ as the OLS coefficients and standard error obtained from fitting the model M_k to the data under consideration, respectively.

Drawing inference based on a single best model ignores model uncertainty. Model averaging has been proposed to account for model uncertainty. Using model averaging, quantities of interest are obtained as a weighted average over a set of models $(M_1, \ldots, M_K)$, rather than from one selected single model. The choice of weights $w_{(k)}$ associated with each candidate model M_k, with $\sum w_{(k)} = 1$, is essential for performing model averaging. BMA uses the posterior model probabilities as weights, while bootstrap model averaging proposed by Buckland et al. (1997) utilizes the frequencies to choose the weights. Then, we are interested in the model averaging estimator of predicted value,

$$\widetilde{\boldsymbol{u}} = \sum_{k=1}^{K} \widehat{\boldsymbol{u}}_{(k)} \widehat{w}_{(k)}, \tag{4.2}$$

where $\widehat{\boldsymbol{u}}_{(k)}$ is the OLS estimator of $\boldsymbol{u}_{(k)}$ for model M_k using data $(\boldsymbol{y}, \boldsymbol{X}_{(k)})$ and the model weight $\widehat{w}_{(k)}$ is an estimate of $w_{(k)}$, the probability that model M_k is the true model. $\widehat{w}_{(k)}$ is subject to the constraint $\sum_{k=1}^{K} \widehat{w}_{(k)} = 1$. In this work, we treat the lasso as a model selection tool and consider a modified bootstrap model averaging, called 'bootstrap lasso averaging' (BLA). BLA is described in the following algorithm,

Algorithm 1. [BLA1]

Step 1: Implement the lasso on the original data $(\mathbf{y}, \mathbf{X})$ and denote the best model selected by the lasso M_l.

Step 2: Fit a least square regression using the data $(\mathbf{y}, \mathbf{X}_{(l)})$ and simulate B response vectors $\{\mathbf{y}_1, \ldots, \mathbf{y}_B\}$ from this regression model; that is: $\mathbf{y}_b = \mathbf{X}_{(l)}\widehat{\boldsymbol{\beta}}_{(l)} + \mathcal{N}\left(0, \mathbf{I}\widehat{\sigma}^2_{(l)}\right)$.

Step 3: For $b = 1, \ldots, B$, fit the adaptive lasso to each bootstrap sample $(\mathbf{y}_b, \mathbf{X}_{(l)})$ and record the best model selected by the adaptive lasso.

Step 4: Define the weight vector $(\widehat{w}_1, \ldots, \widehat{w}_K)$ using the observed frequency under which M_k, $k = 1, \ldots, K$, was selected as the best model over the B simulated resamples in Step 3 and assign the weights to each model.

Step 5: Define $\widehat{\boldsymbol{u}}_{(k)}$ as the estimate of quantity (i.e. prediction) derived from the least square regression using the data $(\mathbf{y}, \mathbf{X}_{(k)})$ and compute the model averaging estimator according to formula (4.2).

More specifically, in Step 1, we implement the lasso using the original data and define l as the indices of nonzero elements in the lasso estimator, and $\boldsymbol{X}_{(l)}$ (an $n \times p'$ matrix) and $\boldsymbol{\beta}^*_{(l)}$ (a $p' \times 1$ vector) as the partitions of $\boldsymbol{X}$ and $\boldsymbol{\beta}^*$ that consist of the variables selected by the lasso. In Step 2, bootstrap resamples are then generated from the model,

$$\mathbf{y}_b = \mathbf{X}_{(l)}\widehat{\boldsymbol{\beta}}_{(l)} + \mathcal{N}\left(0, \mathbf{I}\hat{\sigma}^2_{(l)}\right), \quad b = 1, \ldots, B,$$

where $\widehat{\boldsymbol{\beta}}_{(l)}$ and $\widehat{\sigma}_{(l)}$ are the OLS coefficients and standard error obtained by fitting model M_l to the original data, respectively. In Step 3, we consider the adaptive lasso estimator $\widehat{\boldsymbol{\beta}}^{(b)}_{AL}$ for the b^{th} bootstrap sample $(\mathbf{y}_b, \mathbf{X}_{(l)})$ as the solution to the following minimization problem,

$$\widehat{\boldsymbol{\beta}}^{(b)}_{AL} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\, Q_n(\boldsymbol{\beta}) = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ \|\boldsymbol{y}_b - \boldsymbol{X}_{(l)}\boldsymbol{\beta}\|^2 + \lambda_n \sum_{j=1}^{p'} \widehat{v}_j |\beta_j| \right\},$$

where λ_n is the tuning parameter, and $\widehat{v}_j$ is the adaptive weight for j^{th} variable. A standard choice of $\widehat{v}_j$ is $|\widehat{\beta}_j|^{-1}$, where $\widehat{\beta}_j$ is the OLS estimate for the j^{th} variable. The R package '*glmnet*' (Friedman et al., 2009) is used for fitting the lasso and the adaptive lasso and the five folds cross-validation is performed to find the optimal value of tuning parameter λ in this paper. We further define $M_{k'}, k' = 1, \ldots, M'$ as the candidate models that include the variables selected by the lasso, and $\mathcal{K}' = (M_1, \ldots, M_{k'})$ as the set of candidate models in the adaptive lasso step. Since the variables excluded from the lasso step (Step 1) will no longer be considered

in the adaptive lasso step (Step 3). Therefore, in Step 4, we define the estimated model weights by

$$\widehat{w}_{(k)} = \begin{cases} \frac{1}{B}\sum_{l=1}^{B} I\left(\widehat{\boldsymbol{\beta}}^{(b)}_{AL(k)} \neq \mathbf{0}, \widehat{\boldsymbol{\beta}}^{(b)}_{AL(k)^c} = \mathbf{0}\right) & \text{if } M_k \in \mathcal{K}' \\ 0 & \text{if } M_k \notin \mathcal{K}', \end{cases}$$

where $\widehat{\boldsymbol{\beta}}^{(b)}_{AL(k)}$ defines the partition of $\widehat{\boldsymbol{\beta}}^{(b)}_{AL}$ that consists of coefficients for variables included in the model M_k. Finally, we compute our BLA estimator $\widetilde{\boldsymbol{u}}$ according to (4.2).

Compared with bootstrap model averaging in Buckland et al. (1997), BLA as described in Algorithm 1 reduces the computation cost and allows model averaging in high dimensional cases. Step 1 in Algorithm 1 not only provides a model for generating bootstrap resamples, but also reduces the dimension of the variables to be taken into account in the adaptive lasso step (Step 3). Without the 'irrepresentable condition' (Zhao and Yu, 2006; Zou, 2006), the lasso is inconsistent in variable selection. However, according to Propositions 1 and 2 in Bach (2008), under some mild conditions, the lasso selects all significant variables with probability tending to one exponentially fast, although it also includes insignificant variables with certain probabilities. Therefore, we consider this modified Step 1 as a valid variable screening step. It also reduces the dimension of covariates to smaller than the sample size, which further allows the implementation of the adaptive lasso in Step 3, since the adaptive weights generally rely on a consistent unpenalized estimator, which is generally infeasible for computing in a high dimensional case where $p > n$. In the last step, we derive the model averaging estimator based on the OLS post-model selection estimator computed by using the original sample. That is, we are not trying to combine the adaptive lasso estimators from each bootstrap resample, but only to utilize the bootstrap resamples to estimate the model weights. There are several reasons why we separate the model selection from the parameter estimation. Conceptually, the model averaging approach targets the discovery of an appropriate weight for each candidate

model. Therefore, we prefer to perform the estimation using the original data instead of the bootstrap resamples. In addition, Belloni et al. (2013) showed that the OLS post-lasso estimator performs at least as well as the lasso in terms of the convergence rate and has the advantage of a smaller bias. More details of OLS post various forms of lasso estimator can be found in Belloni et al. (2013).

As stated in Efron (1992), considering the fixed design matrix $\mathbf{X}$, resampling for regression problems should be from the residuals so that analysis remains conditional on the covariate values. Alternatively, for a random design matrix, we could consider another bootstrap scheme that generates the resamples from the sampling units; that is, bootstrapping the response along with the associated covariates. In a similar way to Algorithm 1, we implement the lasso to construct a model averaging procedure for random design as follows,

Algorithm 2. [**BLA2**]

Step 1: Draw B bootstrap samples from the sampling units with replacement from the original $(\mathbf{y}, \mathbf{X})$. Implement the lasso on each B bootstrap sample and denote the best model selected by the lasso $M_{lb}, b = 1, \ldots, B$.

Step 2: Fit a least square regression using the data $(\mathbf{y}, \mathbf{X}_{(lb)})$ and simulate a response vector $\mathbf{y}_b$ from this regression model; that is: $\mathbf{y}_b = \mathbf{X}_{(lb)}\widehat{\boldsymbol{\beta}}_{(lb)} + \mathcal{N}\left(0, \mathbf{I}\widehat{\sigma}^2_{(lb)}\right)$. Repeat this for $b = 1, \ldots, B$.

Step 3: For $b = 1, \ldots, B$, fit the adaptive lasso to each bootstrap sample $(\mathbf{y}_b, \mathbf{X}_{(lb)})$ and record the best model selected by the adaptive lasso.

Step 4 - Step 5: These are identical to Steps 4 and 5 in Algorithm 1.

The major difference between Algorithm 1 and Algorithm 2 is that Algorithm 2 involves an extra layer of bootstrapping from the sampling units. Hence, different variables may be screened by the lasso in each of the B resamples. In many well-studied variable screening methods, such as sure independence screening (SIS) (Fan and Lv, 2008) and forward regression (FR) screening (Wang, 2009),

once the variables have been excluded from the screening step, they will not be reconsidered for further variable selection methods. Although both Fan and Lv (2008) and Wang (2009) have shown the asymptotic screening consistency of their methods: that is, all relevant variables are discovered as sample size $n \to \infty$, it is still highly likely that important variables are eliminated by variable screening in a small sample case, where the model uncertainty is substantial. Algorithm 1 suffers the same problem, while in Algorithm 2, variables excluded by the lasso in one bootstrap sample can still be selected in the other. Therefore, Algorithm 2 tends to be more conservative in variable screening than Algorithm 1. This is numerically confirmed in our simulation studies and real example analysis, where we show the significant outperformance of Algorithm 2.

Knight and Fu (2000) and Chatterjee and Lahiri (2010) considered the non-parametric residual bootstrap for the lasso estimator and stated that the bootstrap lasso estimator is inconsistent whenever there are one or more zero components of the parameter vector $\boldsymbol{\beta}$. Thereafter, Chatterjee and Lahiri (2011) proposed a non-parametric residual bootstrap method for the adaptive lasso and further asserted its strong consistency in variable selection under some mild regularity assumptions. Both Algorithm 1 and Algorithm 2 we propose utilize a parametric residual bootstrap to generate resamples from the presumed parametric model (e.g. normal distribution), while the parametric model can be sometimes mis-specified. Therefore, for comparison purposes, we further propose a nonparametric BLA based on Chatterjee and Lahiri (2011) and describe it in the following algorithm,

Algorithm 3. [BLA3]

Step 1: Implement the lasso on the original data $(\mathbf{y}, \mathbf{X})$ and denote the best model selected by the lasso M_l. Let $\widehat{\boldsymbol{\beta}}_{las}$ denote the lasso estimator of $\boldsymbol{\beta}$ and define the residuals

$$e_i = y_i - X_i^T \widehat{\boldsymbol{\beta}}_{las}, \quad i = 1, \ldots, n.$$

Step 2: Define the set of centered residuals $\mathbf{e}$ using $\mathbf{e}^* = (e_1^*, \ldots, e_n^*)$. Draw B bootstrap resamples of $\mathbf{e}^*$ with replacement, define them by $\{\mathbf{e}_1^*, \ldots, \mathbf{e}_B^*\}$ and formulate B response vectors $\{\mathbf{y}_1, \ldots, \mathbf{y}_B\}$; that is: $\mathbf{y}_b = \mathbf{X}\widehat{\boldsymbol{\beta}}_{las} + \mathbf{e}_b^*$.

Step 3: For $b = 1, \ldots, B$, fit the adaptive lasso to each bootstrap sample $(\mathbf{y}_b, \mathbf{X}_{(l)})$ and record the best model selected by the adaptive lasso.

Step 4 - Step 5: These are identical to Steps 4 and 5 in Algorithm 1.

The only difference between Algorithm 1 and Algorithm 3 is that Algorithm 3 applies the nonparametric bootstrap. Chatterjee and Lahiri (2011) consider the simple residual bootstrap, which is shown to consistently estimate the distribution and provide variance estimates for the adaptive lasso. The adaptive weights in their residual bootstrap estimator is based on the OLS estimator, which cannot feasibly be computed when $p > n$. Unlike the work of Chatterjee and Lahiri (2011), our work includes a variable screening procedure in the lasso step (Step 1) so that the OLS estimator can still be used to compute the adaptive weights when $p > n$. Moreover, as mentioned in Algorithm 1, our work focuses on the model averaging estimator based on the OLS post-model selection estimator computed by using the original sample, which is different from the adaptive lasso based residual bootstrap estimator of Chatterjee and Lahiri (2011).

In our simulation study, we will compare these three algorithms for BLA with other model averaging methods in terms of prediction and variable selection accuracy under different simulation settings.

4.3 Simulation studies

In this section, we conduct simulation studies to investigate the finite sample performance of three proposed algorithms for BLA, compared with other widely used model averaging methods. The comparison is conducted by measuring the variable selection and prediction accuracies. A brief description of other model

averaging procedures is listed as follows *.

- *BootAIC*: This procedure is the bootstrap model averaging proposed by Buckland et al. (1997) with a weighting scheme based on AIC. The weight assigned to each model is equal in proportion to the number of times that the model is selected in each bootstrap resample according to AIC. Details of the procedure have been discussed in Section 4.1.

- *BootBIC*: This procedure is identical to BootAIC but with a weighting scheme based on BIC instead of AIC.

- *S-AIC*: This procedure assigns weight or approximate posterior probability to each candidate model M_k in the form:

$$\frac{\exp\left(-0.5\mathrm{AIC}_k\right)}{\sum_{i=1}^{K}\exp\left(-0.5\mathrm{AIC}_i\right)}, \tag{4.3}$$

  where AIC_i is the value of AIC for model M_i. It has been shown that with suitable prior model probabilities , S-AIC can be framed in a Bayesian context (Burnham and Anderson, 2003, 2004; Clyde et al., 2000).

- *S-BIC*: This procedure is identical to S-AIC but with BIC_i replacing AIC_i in equation (4.3). Assuming equal prior model probabilities, such that $P(M_i) = 1/K$, it can be shown that S-BIC provides approximate posterior model probabilities (Raftery, 1995). In many studies the use of S-BIC for assigning weights or posterior probabilities is referred to as Bayesian model averaging (BMA) (Burnham and Anderson, 2003; Claeskens et al., 2008; Clyde et al., 2000).

- *BMA-MC3*: This procedure moves stochastically through model space using an Markov chain Monte Carlo approach (Raftery et al., 1997). Models

*Part of this description is based on an unpublished work "Iterative Frequentist Model Averaging" by Bala Rajaratnm and Steven Roberts

are visited using an Metroplis-Hastings algorithm on the integrated likelihood. Posterior model probabilities are then computed for each model visited. MC3 is implemented using the function *'MC3.REG'* available in the R-package 'BMA'. We implement MC3.REG using 100,000 iterations of the Markov chain sampler and the default hyperparmeter values.

- *BMA-OCC*: As for S-BIC, this approach uses equation (4.3) with BIC_i replacing AIC_i to assign approximate posterior probabilities to each candidate model. The difference compared with S-BIC is that OCC uses Occam's window to exclude models that are far less likely than the most likely model (Raftery, 1995). Our implementation of OCC excludes models that are 20 times less likely than the most likely model. OCC is implemented using the function *'bicreg'* available in the R-package 'BMA'.

4.3.1 The simulation models

Recall the following linear regression model,

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}^* + \boldsymbol{\epsilon},$$

We consider three different examples in our simulation study. Details of these three examples are as follows.

- Example 1. This example is adapted from Zou (2006). There are $p = 8$ variables. The covariates are generated from multivariate normal distribution with a mean zero and

$$Cov(x_{ij}, x_{il}) = \rho^{|j-l|}, \quad 1 \leq j, l \leq 10, \tag{4.4}$$

where $\rho = 0.5$. In other words, the pairwise correlation between the j^{th} and the l^{th} variable is set to be $\rho^{|j-l|}$. In addition, we define the true model as

follows,

$$y_i = 3x_{i1} + 1.5x_{i2} + 2x_{i5} + \epsilon_i, \quad i = 1, \ldots, n. \tag{4.5}$$

Hence, the true model is of the form of equation (4.1), where

$$\boldsymbol{\beta}^* = (0, 3, 1.5, 0, 0, 2, 0, 0, 0)^T.$$

Further, $\epsilon_i \sim N(0, 1)$. In other words, only three regression covariates are significant variables and the rest is noise. In this example, we generate training samples as the above setting with four levels of sample size n=20, 30, 50, and 100.

- Example 2. We use the same model as in Example 1 but with $p = 18$. The true model still follows (4.5) but an additional 10 noise variables are included in the design matrix $\mathbf{X}$. In this example, we consider the sample sizes n=30, 50, 80, and 100.

- Example 3. There are $p = 500$ variables. The first six coefficients are nonzero. The covariates are generated from the standard normal distribution independently. The true model is also of the form of equation (4.1) , where $\boldsymbol{\beta}^* = (0, 3, 3, 1.5, 1.5, 1, 1, 0, \ldots, 0)$ and $\epsilon_i \sim N(0, 1)$. We consider four different levels of sample size n=50, 100, 150, and 200. In this example, the covariates dimension p is larger than n. In this situation, the computations of other model averaging methods are infeasible. Therefore, we compare the BLA with the other two classical variable screening methods instead, namely, SIS (Fan and Lv, 2008) and FR screening (Wang, 2009). We use the notations SIS-Alasso and FR-Alasso to represent the model further selected by the adaptive lasso after a variable screening step by using SIS and FR, respectively. The performances of SIS-Alasso and FR-Alasso are examined for comparison purposes.

We generate 100 simulations for each of the above simulation examples and

investigate the model selection and prediction accuracy of each procedure. Model selection accuracy is measured by a 95 % model confidence set, the coverage rate and the percentage of correctly fitted. For these model averaging procedures, the 95% model confidence set is defined as the smallest set of models required to capture at least 95 % of the posterior model probability or assigned weight. The coverage rate is the proportion of times that the true model fails in the 95% model confidence set over 100 simulations. The percentage of correctly fitted is the proportion of times that the model receiving maximum weight is the true model (for the model averaging procedure) or the selected best model is the true model (for the single best model selection procedure) over 100 simulations. Prediction accuracy is measured by the mean squared prediction error (MSPE) $n^{-1}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2$, computed over a set of independent test samples with the same sample size n as the training sample. We have conducted 100 repeated simulations and have computed the average of the MSPEs and their standard errors.

4.3.2 Simulation results

The simulation results for Example 1 are summarized in Table 4.1 and Figure 4.1. As we can see from Table 4.1, most of the model averaging methods achieve satisfactory coverage rates. Compared with other model averaging procedures, BLA using all three algorithms (BLA1, BLA2, and BLA3) obtains consistently higher percentages of correctly fitted, but exhibits much smaller sizes of confidence sets. Even when the sample size is only 20, bootstrap lasso averaging by using algorithms 2 and 3 (BLA2 and BLA3) can achieve a desirable percentage of correctly fitted at 90%. When the sample size is large, the lasso and AIC-based model selection procedures show an inconsistency in model selection as supported by previous studies (Zou, 2006; Bach, 2008; Zhao and Yu, 2006; Burnham and Anderson, 2004; Claeskens et al., 2008); while the simulation results have numerically confirmed a strong consistency of BLA since both the percentage of

correctly fitted and the size of the confidence set converge to 1. The only comparable method among other model averaging procedures is BMA-MC3, which requires extremely expensive computation. From the prediction point of view, model averaging procedures generally outperform a single best model, indicating a substantial model uncertainty within the simulated data. Among these model averaging methods, BLA demonstrates a competitive performance. When the sample size is 20, BLA2 clearly dominates all other model averaging methods, achieving the lowest average MSPE with the smallest standard error. When the sample size goes large, BLA with all three algorithms still performs no worse than others, which is also shown in Figure 4.1. These simulation results strongly favor the use of BLA.

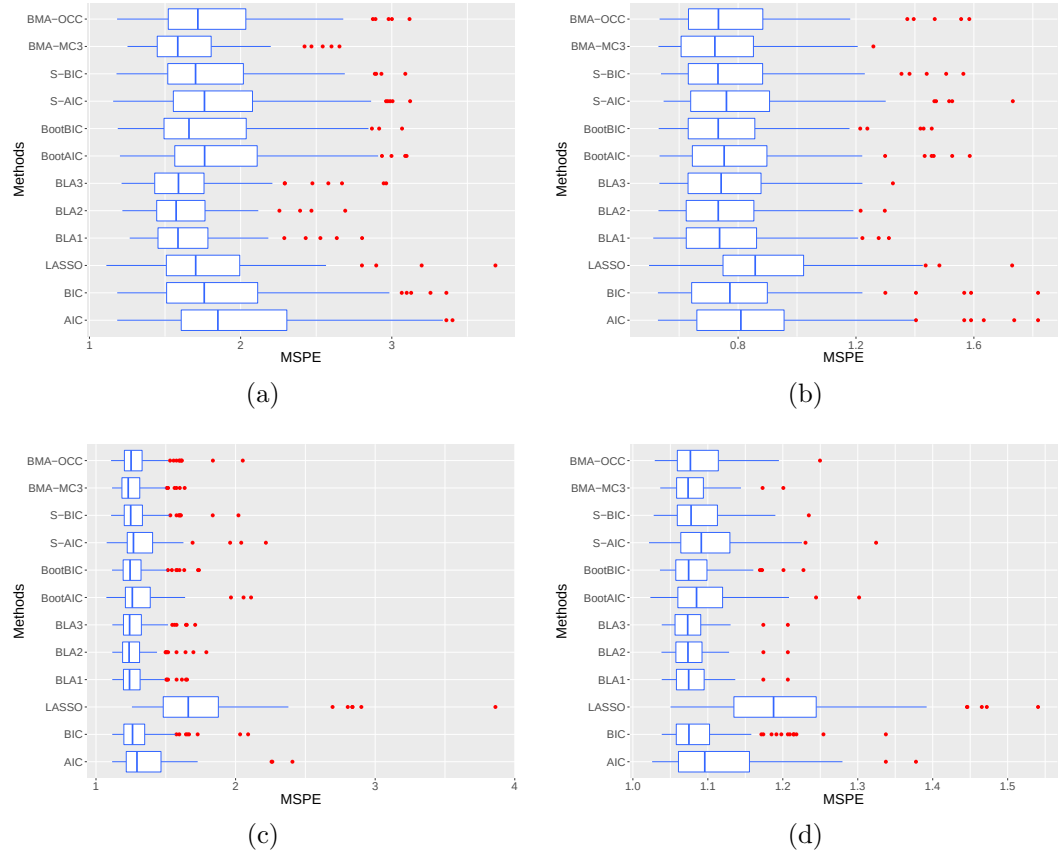


Figure 4.1: Mean squared predication error for Example 1 with (a) $n = 20$, (b) $n = 30$, (c) $n = 50$, and (d) $n = 100$. Note the scales are different in four plots.

Table 4.1: Simulation results for Example 1 ($p = 8$ and $n = 20, 30, 50, 100$)

n	Method	Coverage rate	Confidence set	Correctly fitted	Average MSPE	Std.error MSPE
20	AIC				1.98	0.53
	BIC				1.89	0.50
	LASSO			0.27	1.78	0.44
	BLA1	0.99	2.98	0.79	1.66	0.30
	BLA2	0.99	9.14	0.90	1.63	0.26
	BLA3	0.99	3.18	0.90	1.66	0.34
	BootAIC	0.80	19.16	0.28	1.87	0.43
	BootBIC	0.94	15.63	0.46	1.80	0.40
	S-AIC	0.78	22.92	0.24	1.87	0.43
	S-BIC	0.89	20.63	0.41	1.82	0.40
	BMA-MC3	0.99	6.68	0.86	1.66	0.29
	BMA-OCC	0.82	15.27	0.41	1.83	0.42
30	AIC				0.86	0.27
	BIC				0.81	0.24
	LASSO			0.41	0.90	0.23
	BLA1	1.00	1.94	0.93	0.77	0.19
	BLA2	1.00	3.66	0.96	0.76	0.18
	BLA3	1.00	1.87	0.98	0.77	0.18
	BootAIC	0.95	18.32	0.36	0.81	0.23
	BootBIC	1.00	10.58	0.69	0.78	0.20
	S-AIC	0.95	24.95	0.33	0.82	0.24
	S-BIC	0.98	19.16	0.63	0.79	0.22
	BMA-MC3	1.00	5.11	0.97	0.76	0.18
	BMA-OCC	0.97	12.06	0.63	0.79	0.22
50	AIC				1.36	0.23
	BIC				1.31	0.17
	LASSO			0.36	1.76	0.41
	BLA1	1.00	1.72	0.99	1.27	0.12
	BLA2	1.00	2.61	1.00	1.27	0.12
	BLA3	1.00	1.35	1.00	1.28	0.12
	BootAIC	0.96	15.72	0.34	1.32	0.18
	BootBIC	1.00	6.89	0.69	1.28	0.13
	S-AIC	0.93	24.59	0.31	1.33	0.19
	S-BIC	1.00	16.61	0.68	1.29	0.15
	BMA-MC3	1.00	4.29	0.99	1.26	0.11
	BMA-OCC	0.99	9.78	0.68	1.29	0.15
100	AIC				1.11	0.07
	BIC				1.09	0.05
	LASSO			0.62	1.21	0.10
	BLA1	1.00	1.36	0.99	1.08	0.03
	BLA2	1.00	1.36	1.00	1.08	0.03
	BLA3	1.00	1.00	1.00	1.08	0.03
	BootAIC	0.99	14.36	0.46	1.10	0.05
	BootBIC	1.00	4.58	0.85	1.08	0.04
	S-AIC	0.96	24.64	0.42	1.11	0.05
	S-BIC	0.99	12.47	0.83	1.09	0.04
	BMA-MC3	1.00	3.17	0.99	1.08	0.03
	BMA-OCC	0.99	6.99	0.83	1.09	0.04

Note: Coverage rate and Confidence set are only reported for model averaging methods. The percentages of correctly fitted for AIC and BIC are the same with those of S-AIC and S-BIC, respectively.

Table 4.2: Simulation results for Example 2 ($p = 18$ and $n = 20, 30, 50, 100$)

n	Method	Coverage rate	Confidence set	Correctly fitted	Average MSPE	Std.error MSPE
30	AIC				2.65	1.29
	BIC				2.20	1.01
	LASSO			0.20	1.91	0.42
	BLA1	0.99	4.93	0.87	1.63	0.29
	BLA2	1.00	18.80	0.95	1.58	0.21
	BLA3	1.00	3.60	0.97	1.57	0.18
	BootAIC	0.13	91.02	0.05	2.27	0.95
	BootBIC	0.69	74.88	0.28	1.92	0.63
	S-AIC	0.70	13228.22	0.02	2.22	0.86
	S-BIC	0.96	6966.42	0.22	1.93	0.53
	BMA-MC3	1.00	90.13	0.89	1.59	0.25
	BMA-OCC	0.69	77.23	0.22	2.01	0.68
50	AIC				1.77	0.51
	BIC				1.56	0.39
	LASSO			0.41	1.73	0.25
	BLA1	0.99	3.87	0.89	1.41	0.21
	BLA2	1.00	11.48	0.97	1.40	0.17
	BLA3	1.00	1.98	0.99	1.38	0.13
	BootAIC	0.33	88.63	0.15	1.62	0.39
	BootBIC	0.95	46.12	0.51	1.47	0.27
	S-AIC	0.89	15945.75	0.07	1.64	0.41
	S-BIC	1.00	4171.48	0.45	1.52	0.30
	BMA-MC3	1.00	36.14	0.88	1.39	0.16
	BMA-OCC	0.85	53.48	0.45	1.52	0.33
80	AIC				1.54	0.16
	BIC				1.46	0.13
	LASSO			0.53	1.68	0.16
	BLA1	1.00	3.49	0.96	1.40	0.08
	BLA2	1.00	8.98	0.98	1.40	0.08
	BLA3	1.00	1.29	1.00	1.39	0.08
	BootAIC	0.35	85.71	0.13	1.47	0.12
	BootBIC	0.96	31.00	0.48	1.42	0.09
	S-AIC	0.88	16255.69	0.06	1.49	0.13
	S-BIC	1.00	2669.95	0.45	1.43	0.10
	BMA-MC3	1.00	26.84	0.92	1.40	0.08
	BMA-OCC	0.89	42.16	0.44	1.44	0.11
100	AIC				0.98	0.13
	BIC				0.90	0.10
	LASSO			0.61	1.28	0.21
	BLA1	1.00	1.77	0.98	0.87	0.08
	BLA2	1.00	5.27	0.99	0.87	0.08
	BLA3	1.00	1.04	1.00	0.87	0.08
	BootAIC	0.48	84.45	0.13	0.93	0.11
	BootBIC	0.98	24.41	0.66	0.88	0.09
	S-AIC	0.91	16970.26	0.06	0.95	0.11
	S-BIC	1.00	1917.03	0.61	0.90	0.10
	BMA-MC3	1.00	19.77	0.94	0.87	0.08
	BMA-OCC	0.93	34.18	0.61	0.90	0.10

Note: Coverage rate and Confidence set are only reported for model averaging methods. The percentages of correctly fitted for AIC and BIC are the same with those of S-AIC and S-BIC, respectively.

Table 4.3: Simulation results for Example 3 ($p = 500$ and $n = 50, 100, 150, 200$)

n	Method	Coverage rate	Confidence set	Correctly fitted	Average MSPE	Std.error MSPE
50	LASSO			0.18	1.66	0.40
	SIS-Alasso			0.61	1.21	0.30
	FR-Alasso			0.06	1.53	0.61
	BLA1	0.96	11.01	0.67	1.09	0.27
	BLA2	1.00	47.39	0.96	1.00	0.13
	BLA3	1.00	6.76	0.76	1.04	0.18
100	LASSO			0.47	1.58	0.20
	SIS-Alasso			0.97	1.27	0.16
	FR-Alasso			0.27	1.37	0.22
	BLA1	1.00	8.30	0.96	1.08	0.12
	BLA2	1.00	51.53	1.00	1.09	0.10
	BLA3	1.00	2.04	0.96	1.11	0.15
150	LASSO			0.57	1.54	0.15
	SIS-Alasso			0.99	1.31	0.08
	FR-Alasso			0.69	1.33	0.12
	BLA1	1.00	3.67	0.99	1.17	0.04
	BLA2	1.00	26.65	0.99	1.18	0.04
	BLA3	1.00	1.34	0.99	1.19	0.06
200	LASSO			0.65	1.27	0.10
	SIS-Alasso			1.00	1.15	0.06
	FR-Alasso			0.96	1.14	0.06
	BLA1	1.00	1.75	1.00	1.08	0.03
	BLA2	1.00	6.22	1.00	1.08	0.03
	BLA3	1.00	1.02	1.00	1.08	0.03

Note: Coverage rate and Confidence set are only reported for model averaging methods.

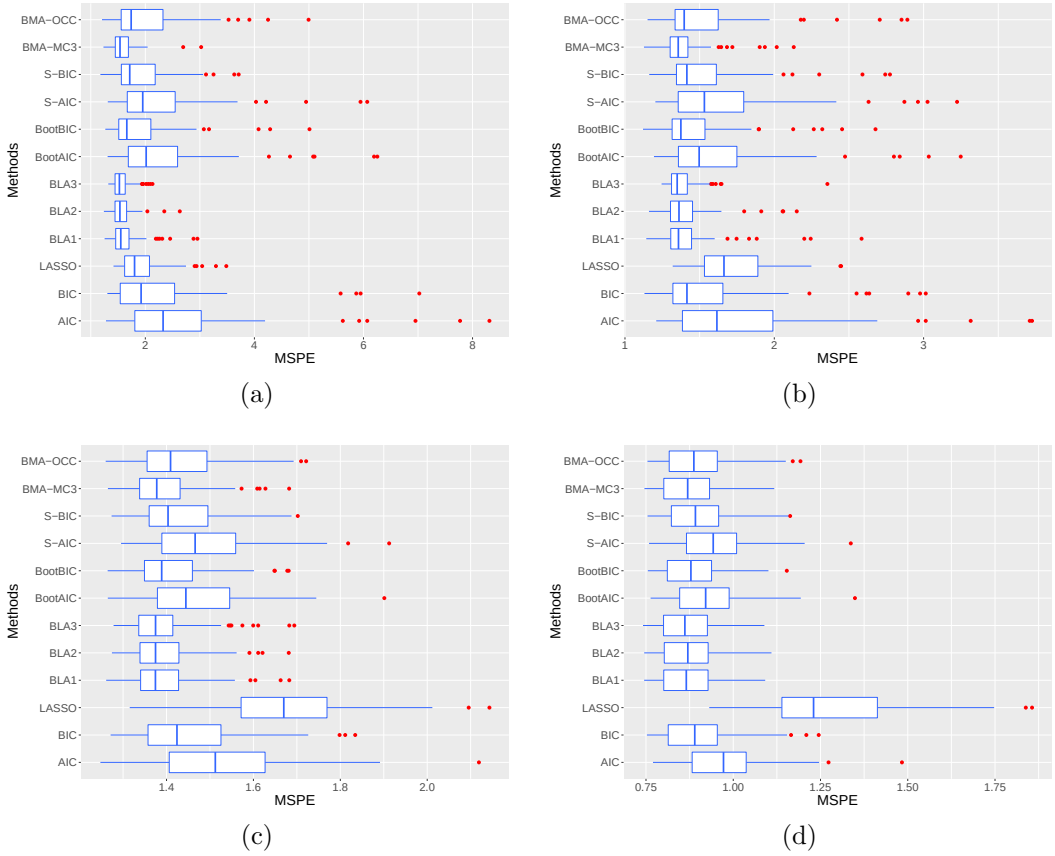


Figure 4.2: Mean squared predication error for Example 2 with (a) $n = 30$, (b) $n = 50$, (c) $n = 80$, and (d) $n = 100$. Note the scales are different in four plots.

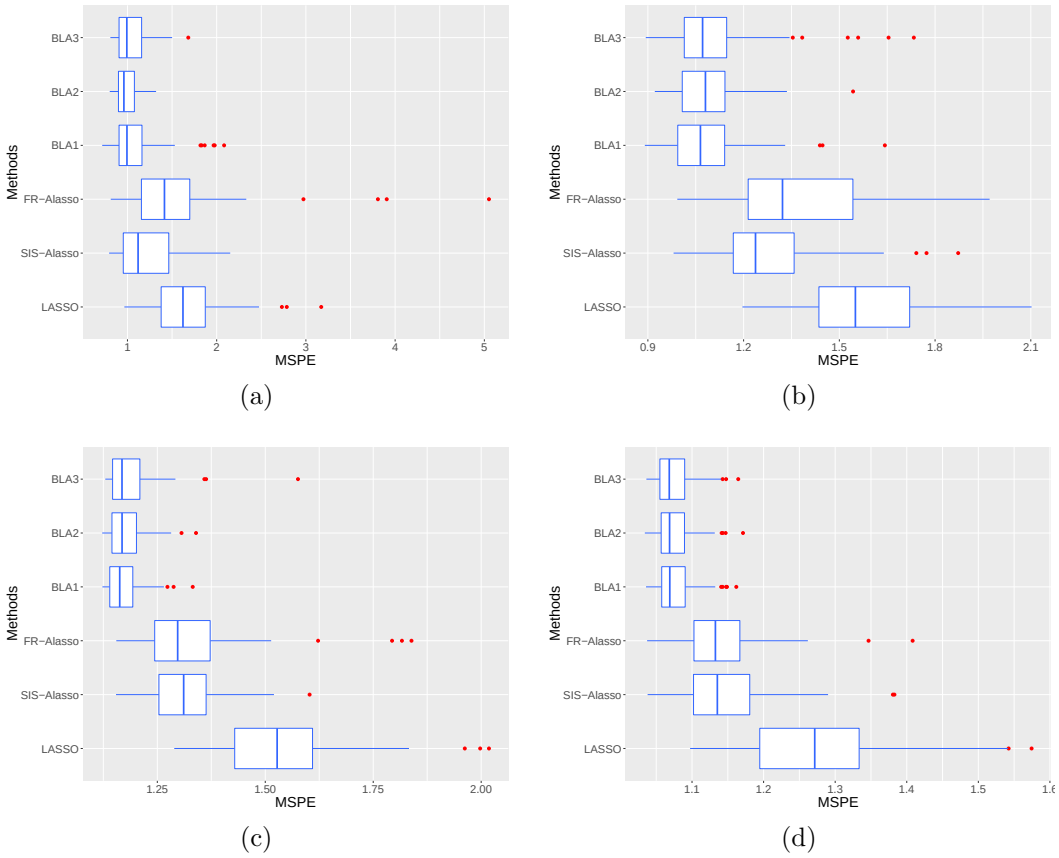


Figure 4.3: Mean squared predication error for Example 3 with (a) $n = 50$, (b) $n = 100$, (c) $n = 150$, and (d) $n = 200$. Note the scales are different in four plots.

The advantage of using BLA is more evident when the predictor dimension becomes large. Table 4.2 and Figure 4.2 present the simulation results for Example 2, where p increases to 18. In this example, it is quite obvious that other model averaging methods generally result in a much larger size of confidence set as a consequence of severe model uncertainly in higher predictor dimensions. For example, the sizes of the confidence sets for S-AIC and S-BIC increase dramatically to thousands, indicating the clear inappropriateness of using traditional smooth AIC and BIC when p is large. Conversely, BLA still demonstrates satisfactory sizes for confidence sets and percentages of the correctly fitted in various sample sizes. When the sample size is small, BLA significantly outperforms the other methods, in terms of both variable selection and prediction accuracies, as shown in Table 4.2 and Figure 4.2. In addition, it is worth noticing that when the sample size increases the percentage of correctly fitted for BLA converges to 1 much faster than in other model selection procedures. All these facts indicate that BLA tends to be a more stable model selection procedure than the other traditional ones.

We finally investigate the simulation results for Example 3, which are displayed in Table 4.3 and Figure 4.3. Since the predictor dimension p is larger than the sample size n, we only present the results of SIS-Alasso and FR-Alasso for comparison purposes as discussed previously. Table 4.3 shows that even when the sample size is only 50, BLA2 achieves a very satisfactory percentage of correctly fitted at 96%, significantly outperforming other methods. As discussed in Section 4.2, this could be explained by the fact that BLA2 tends to be more conservative in variable screening, while the variables excluded from other screening methods (e.g. SIS and FR) will not be reconsidered in further variable selection steps. All three algorithms for BLA perform well as the sample size increases. Among them, BLA3 tends to obtain a smaller confidence set. In addition, FR-Alasso shows a relatively slower convergence rate of model selection consistency as the sample size increases. In terms of prediction accuracy, we notice that all three

algorithms for BLA consistently outperform the lasso, SIS-Alasso and FR-Alasso: they achieve much lower average MSPEs with smaller standard errors than other methods. This is also shown in Figure 4.3.

4.4 Real data examples

For further illustration, we analyze four real data examples by using the proposed BLA and comparing it with the other traditional model selection and model averaging procedures. Since S-AIC and S-BIC consistently perform poorly under all simulation settings, as well as the real data sets we analyze below, we have removed these two methods from the comparisons in this section.

4.4.1 Crime data analysis

The first data set we investigate comprises crime data ($p = 15$; $n = 47$), containing information from 47 states of the United States on crime rates from 1959 to 1960 ($\mathbf{y}$) and 15 socio-economic and demographic variables: percentage of males aged 14-24 (M); indicator variable for southern state (So); mean years of schooling (Ed); police expenditure in 1960 ($Po1$); police expenditure in 1959 ($Po2$); Labor force participation rate (LF); number of males per 1,000 females ($M.F$); state population (Pop); number of nonwhites per 1,000 people (NW); unemployment rate of urban males aged 14-24 ($U1$); unemployment rate of urban males aged 35-39 ($U2$); wealth (GDP); income inequality ($Ineq$); probability of imprisonment ($Prob$); and average time served in state prisons ($Time$). Following Raftery et al. (1997), all the data were transformed logarithmically before analyses. The crime data are available in R-package 'MASS', named by 'UScrime'.

Table 4.4 shows the analysis results of the crime data. The upper panel presents the proportion of times that the variable is included in the model that receives the maximum weight, computed over 100 randomly selected training sets with sample size 20, for each of the variables respectively. It is quite obvious that

each variable has a certain chance of being included in the best model selected by these model selection procedures. This strongly implies the existence of a great amount of model uncertainty within this data set. Among these variables, $Po1$, $M.F$, $Ineq$, and $Prob$ generally have larger frequencies included in the best model than other variables for most of procedures. In addition, AIC-based model selection procedures tend to select more variables, while BLA, especially BLA2, selects substantially fewer variables and identifies $Po1$ as the most outstanding significant variable. As a result of its parsimonious explanation, we may prefer BLA to other model selection or averaging procedures in real data analysis.

The lower panel demonstrates the comparison of predictive performance for various model selection and model averaging procedures at different sizes of training sample (No.train). The remaining data of the size ($n -$ No.train) is treated as test data. The predictive performance is measured by the averaged MSPE, calculated over 100 randomly selected training/test sets of data. As we can see from Table 4.4, when n is only 20, model averaging procedures generally achieve lower average MSPEs than single best model selection methods as a result of great model uncertainty when the size of the training sample is small as previously mentioned. BLA2 obtains the lowest average MSPE with only 0.099, at about half of the averages for AIC and BIC. When the size of the training sample increases, both model selection and model averaging procedures perform comparably well.

4.4.2 Diabetes data analysis

We further investigate a larger data set, the diabetes data ($p = 10$; $n = 442$), described in Efron et al. (2004), containing information on a quantitative measure of disease progression ($\mathbf{y}$) and ten baseline variables: age (AGE); sex (SEX); body mass index (BMI); average blood pressure (BP); and six blood serum measurements ($S1, ..., S6$). The data can be obtained from `http://www-stat.`

Table 4.4: Analysis of the crime data

Variable/ No.train	AIC	BIC	LASSO	BLA1	BLA2	BLA3	Boot AIC	Boot BIC	BMA-mc3	BMA-occ
M			0.34	0.31	0.00	0.15	0.88	0.61	0.21	0.65
So			0.20	0.10	0.00	0.06	0.76	0.43	0.07	0.42
Ed			0.33	0.32	0.02	0.15	0.92	0.79	0.39	0.83
Po1			0.69	0.59	0.40	0.39	0.85	0.64	0.55	0.66
Po2			0.33	0.23	0.09	0.20	0.84	0.59	0.30	0.55
LF			0.34	0.21	0.00	0.14	0.86	0.48	0.08	0.46
M.F			0.43	0.38	0.08	0.25	0.84	0.53	0.11	0.47
Pop			0.20	0.03	0.00	0.03	0.86	0.54	0.10	0.51
NW			0.72	0.39	0.10	0.23	0.94	0.65	0.35	0.67
U1			0.26	0.14	0.00	0.07	0.76	0.49	0.06	0.45
U2			0.35	0.26	0.03	0.14	0.88	0.65	0.16	0.62
GDP			0.14	0.08	0.00	0.04	0.82	0.53	0.23	0.52
Ineq			0.34	0.32	0.04	0.19	0.93	0.83	0.60	0.80
Prob			0.54	0.30	0.03	0.20	0.95	0.69	0.30	0.71
Time			0.24	0.18	0.00	0.10	0.86	0.56	0.12	0.55
20	0.203	0.189	0.118	0.125	0.099	0.128	0.174	0.163	0.102	0.175
30	0.088	0.092	0.091	0.095	0.075	0.100	0.082	0.084	0.078	0.086
40	0.066	0.068	0.079	0.079	0.067	0.088	0.062	0.066	0.068	0.063
2/3	0.086	0.093	0.084	0.094	0.075	0.097	0.079	0.084	0.080	0.081

Note: The upper panel presents the proportion of times that the variable is included in the model that receives the maximum weight, computed over 100 randomly selected training sets with sample size 20. The lower panel shows the averaged MSPE, calculated over 100 randomly selected training/test sets of data. '2/3' corresponds to the case where No.train is equal to two-thirds of the data.

`stanford.edu/~hastie/Papers/LARS/diabetes.data`.

Analytic results of diabetes data are summarized in Table 4.5. We can clearly see that, as in the crime data, when No.train = 20, model uncertainty still exists. However, it is apparent that compared with other traditional methods, BLA still tends to include significant fewer variables in the best model and recognizes BMI and $S5$ as two of the most important variables. This was, also demonstrated in the previous LARS analysis of the diabetes study by Efron et al. (2004). Because of a large sample size with 442 observations in this data, one might be interested in the behavior of these model selection methods when the size of the training sample is large. Table 4.6 demonstrates the frequency of variables selected in the best model when No.train is equal to two-thirds of the data. From Table 4.6, we can see that model uncertainty becomes less obvious since more variables obtain the frequency of selection in the best model, which is close to either 1 or 0. Again, compared with other traditional methods, BLA performs the best in terms of selecting the most parsimonious model. Both BLA2 and BLA3 identify BMI and $S5$ as the only two important variables and the rest as noise variables.

In terms of prediction accuracy, we can see from the lower panel of Table 4.5 that when the size of the training sample is small, model averaging shows an obvious advantage and BLA2 substantially achieves a lower MSPE but with a more parsimonious best model than other traditional methods. As the training sample size increases, the discrepancy between the average MSPE of model averaging and best model selection declines. This is expected because model uncertainty becomes less severe when the sample size is large as we concluded previously. However, even when the training sample size is large, BLA returns a more parsimonious best model but maintains satisfactory prediction accuracy. All of the evidence shows a clear advantage to considering BLA in real data analysis, especially in the case of a small sample size, where great model uncertainty exists.

Table 4.5: Analysis of the diabetes data

Variable/ No.train	AIC	BIC	LASSO	BLA1	BLA2	BLA3	Boot AIC	Boot BIC	BMA-mc3	BMA-occ
AGE			0.08	0.02	0.00	0.00	0.29	0.19	0.05	0.21
SEX			0.07	0.06	0.01	0.00	0.33	0.17	0.11	0.18
BMI			0.48	0.38	0.28	0.17	0.55	0.50	0.48	0.54
BP			0.27	0.06	0.02	0.03	0.41	0.24	0.21	0.27
S1			0.04	0.01	0.00	0.00	0.32	0.18	0.07	0.22
S2			0.08	0.01	0.00	0.00	0.40	0.21	0.11	0.27
S3			0.20	0.07	0.01	0.01	0.36	0.25	0.11	0.24
S4			0.17	0.08	0.07	0.01	0.48	0.32	0.17	0.34
S5			0.43	0.44	0.31	0.19	0.54	0.42	0.40	0.46
S6			0.12	0.02	0.00	0.02	0.35	0.19	0.09	0.20
20	8.01	6.32	5.00	4.96	4.40	5.23	6.62	5.41	4.52	5.67
50	3.82	3.87	3.95	3.71	3.66	4.22	3.68	3.74	3.66	3.67
100	3.30	3.40	3.55	3.34	3.33	3.63	3.28	3.37	3.32	3.30
2/3	3.08	3.10	3.28	3.22	3.20	3.29	3.07	3.11	3.12	3.09

Note: The upper panel presents the proportion of times that the variable is included in the model that receives the maximum weight, computed over 100 randomly selected training sets with sample size 20. The lower panel shows the averaged MSPE, calculated over 100 randomly selected training/test sets of data. '2/3' corresponds to the case where No.train is equal to two-thirds of the data.

Table 4.6: Frequency of variable selected in the best model when No.train $= \frac{2}{3}n$

Variable	LASSO	BLA1	BLA2	BLA3	Boot AIC	Boot BIC	BMA-mc3	BMA-occ
AGE	0.01	0.00	0.00	0.00	0.02	0.00	0.00	0.00
SEX	0.31	0.32	0.12	0.00	1.00	0.87	0.64	0.89
BMI	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
BP	0.99	0.23	0.13	0.01	1.00	0.99	0.93	0.99
S1	0.02	0.00	0.00	0.00	0.82	0.30	0.32	0.39
S2	0.06	0.00	0.00	0.00	0.62	0.13	0.09	0.19
S3	0.96	0.17	0.08	0.00	0.20	0.69	0.57	0.61
S4	0.01	0.00	0.00	0.00	0.23	0.04	0.03	0.07
S5	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
S6	0.26	0.00	0.00	0.00	0.20	0.00	0.00	0.00

4.4.3 Glioblastoma gene expression data analysis

Compared with other traditional model averaging methods, one main advantage of BLA is its computational feasibility in a high dimensional case where $p > n$. For a better illustration of the use of BLA proposed in this paper, we now investigate real datasets where $p > n$. We analyze the glioblastoma gene expression data originally studied by Horvath et al. (2006), and further investigated in Wang et al. (2011) and Roberts and Nowak (2014). The glioblastoma data from two independent sets of clinical tumor samples of $n = 55$ and $n = 65$ with expression values of $p = 3600$ genes are available from `https://labs.genetics.ucla.edu/horvath/CoexpressionNetwork/ASPMgene/`. Details on these data can be found in Horvath et al. (2006). Following Wang et al. (2011), before analyzing these data we excluded nine censored subjects, five from the first set of 55 patients and four from the second set, and used the logarithm of time to death as the response. As suggested by Roberts and Nowak (2014), we further removed patient 29 from the first set, which was identified as an outlier since he or she had a much smaller survival time than other patients. Then, the first data set was served as the training set where $n = 49, p = 3600$ and the second set was designated as the test set where $n = 61, p = 3600$.

Following Wang et al. (2011), we assessed each of the 3600 genes by running simple linear regression on the training set and select 1000 genes with the smallest p-values. Starting with these 1000 genes, Table 4.7 presents the glioblastoma gene expression data analysis and compares the performance of BLA with the lasso and other variable screening methods. The lasso selects 30 variables and achieves a MSPE of 1.54, while BLA2 selects a more parsimonious model with a lower MSPE of 1.16. We can see that BLA2 clearly outperforms all other methods for this dataset. Table 4.8 further shows the estimated coefficients of the best model selected by BLA2. However, similar to the findings of Roberts and Nowak (2014), we also found that the results for the glioblastoma data displayed in Table 4.7 are extremely volatile because of the random fold assignment of cross-validation when

we determined the tuning parameter for the lasso problems. Therefore, the results presented in Table 4.7 are contingent on the particular fold assignment. To take into account this instability, we have performed the lasso, SIS-Alasso, FR-Alasso, and three versions of BLA 100 times on the glioblastoma data and we present the results in Table 4.9. It is worth noting that BLA2 significantly outperforms other methods. It achieves the lowest average MSPE with a small standard error, while identifying the least number of variables in the best model. This implies that BLA2 has potential to be less sensitive to random fold assignments in cross-validation. Overall, BLA2 has obvious advantages over the other methods in the analysis of the glioblastoma data.

Table 4.7: Glioblastoma gene expression data analysis

Method	Size of the best model	MSPE
LASSO	30	1.54
SIS-Alasso	23	1.42
FR-Alasso	22	2.23
BLA1	27	2.03
BLA2	19	1.16
BLA3	17	1.76

Table 4.8: Estimated coefficients of the best model selected by BLA2 for Glioblastoma gene expression data

Gene.Symbol	Coefficient	Gene.Symbol	Coefficient
UROD	0.44	HSD17B4	0.35
FBL	1.50	IGHG1	0.03
CSN3	-0.18	GTSE1	-1.09
WTAP	0.28	DNASE1L1	-1.15
ANKRD25	0.45	NBL1	-0.77
CA9	0.13	GEMIN6	0.47
TCF20	-0.01	LYPLA1	-1.61
CORO1A	0.22	FLJ14281	0.35
ELTD1	0.42	TRPM2	1.54
OSTF1	-1.38		

Table 4.9: Glioblastoma gene expression data analysis averaged over 100 runs

Method	Average size of the best model	Average MSPE	Std.error MSPE
LASSO	29.31	1.45	0.17
SIS-Alasso	15.83	1.30	0.16
FR-Alasso	22.27	2.24	0.02
BLA1	22.13	1.84	0.24
BLA2	11.25	1.13	0.03
BLA3	18.34	1.69	0.21

4.4.4 Near-Infrared (NIR) spectroscopy of biscuit doughs data

To further illustrate the performance of BLA in a high dimensional case, we now analyze the Near-Infrared (NIR) spectroscopy of biscuit doughs data as discussed in Brown et al. (2001). This data set contains measurements from quantitative NIR spectroscopy. For more details, see Brown et al. (2001). Briefly, two similar sample sets were made up, with the standard recipe varied to provide a large range for each of the four constituents under investigation: fat, sucrose, dry flour, and water, which are presented as percentages in the dataset. The spectral data consist of 700 points measured from 1100 to 2498 nanometers (nm) in steps of 2 nm. In our work, we have limited our investigation to the part of the spectral data that is most important for predicting one of the four constituents, fat. There are 40 samples in the original training set but with sample 23 identified as an outlier; there are a further 32 samples in the validation set with example 21 considered an outlier. Following Brown et al. (2001), we have removed these two outliers to consider a training set where $n = 39, p = 700$ and a test set where $n = 31, p = 700$.

Table 4.10 displays the performances of various methods for addressing biscuit doughs data. It shows that all three algorithms for BLA outperform the lasso by achieving lower MSPEs and selecting more parsimonious models but without missing important variables (as shown in Table 4.11). Among them,

BLA2 performs best with the lowest MSPE of 0.052. Conversely, SIS-Alasso and FR-Alasso demonstrate very unsatisfactory predictive performance in terms of biscuit doughs data. As we can see from Table 4.11, the variables (spectrum) selected by SIS-Alasso and FR-Alasso are quite different from those of the lasso and BLA. It implies that the screening methods of SIS and FR may have excluded some important variables and produced inaccurate predictions. However, the three different versions of BLA tend to discard unnecessary details and select similar ranges of spectrum as having predictive potential.

Table 4.10: Near-Infrared (NIR) spectroscopy of biscuit doughs data

Method	Size of the best model	MSPE
LASSO	14	0.082
SIS-Alasso	1	1.932
FR-Alasso	5	0.416
BLA1	4	0.070
BLA2	3	0.052
BLA3	5	0.070

4.5 Conclusion

To conclude, we propose the extension of a bootstrap model averaging approach, called BLA. The simulation results and real data examples demonstrate the following advantages of BLA: (i) compared with traditional model averaging procedures, BLA shows a comparable performance but at far less computational cost. When n is small and p is large, BLA strongly outperforms most of other model averaging methods; (ii) BLA shows a superior performance in terms of both variable selection and prediction accuracies. If we treat BLA as a single model selection method (e.g. for selecting the model that receives the maximum weight), it can be viewed as a more stabilized version of the lasso; (iii) in contrast to classical variable screening procedures like SIS and FR, BLA tends to accurately identify most significant variables with fewer restrictions on the size

Table 4.11: Estimated coefficients of the best model selected for biscuit doughs data

Method	Variable	Coefficient
LASSO	54	215.73
	55	-160.28
	206	290.23
	207	490.33
	208	-833.03
	209	23.43
	247	217.96
	252	-337.91
	253	35.79
	312	-515.04
	313	596.46
	486	-150.17
	487	140.35
	685	-0.66
BLA1	55	45.27
	208	-52.01
	252	-59.76
	313	83.80
BLA2	259	-97.39
	312	101.94
	488	-17.13
BLA3	55	27.93
	207	-48.63
	252	-56.44
	313	91.62
	486	-6.11
SIS-Alasso	241	-20.92
FR-Alasso	57	218.42
	108	-114.07
	245	1248.31
	246	-1324.13
	632	6.19

of the sample and finally achieves a better prediction. Our numerical simulations and real data examples suggest that BLA should receive more attention in the application of the model averaging or the lasso to prediction problems.

Chapter 5

Conclusion and Future Work

The three independent essays contained in this thesis analyse various topics in model selection, robust statistics and model averaging. In Chapter 2, we propose a robust AIC for MM-estimation and an adjusted robust scale based AIC for M and MM-estimation. We compare our proposed criteria with other robust model selection criteria discussed in previous literature. Our simulation studies demonstrate a significant outperformance of robust AIC based on MM-estimation in the presence of outliers in the covariates and show a better performance by the adjusted robust scale based AIC for MM-estimation when the proportion of outliers in the response is relatively high. Real data examples also show a better performance of robust AIC based on MM-estimation. In Chapter 3, we propose the Tukey-lasso method, which combines Tukey's biweight loss and the adaptive lasso penalty. Using the APG method, the Tukey-lasso can be computed very efficiently and rapidly. Our simulation studies demonstrate that the Tukey-lasso compares favorably with the adaptive lasso and other robust implementations of the lasso. Real data examples also support the use of the Tukey-lasso in variable selection and prediction problems. In Chapter 4, we propose an extension of a bootstrap model averaging approach, called BLA. Our numerical simulations and real data examples show a superior performance of the BLA and suggest that the BLA should receive more attention in the application of the model averaging or

the lasso to prediction problems.

The work presented in this thesis leaves several directions open for future research. First, we consider only the AIC penalty in Chapter 2, while more robust model selection criteria with different penalty terms (e.g. robust BIC) can be further investigated. Second, we have not examined and discussed the inference based on the model selected by the Tukey-lasso. Asymptotic standard errors can be obtained from Theorem 1. Conversely, finite sample standard errors for the Tukey-lasso could be estimated by ensuring that the local quadratic approximation (LQA) referred to Fan and Li (2001) and Zou (2006) is more robust, or by further developing the bootstrapping adaptive lasso estimators discussed in Chatterjee and Lahiri (2011). Such examinations may be included in future work. Finally, we implement the traditional lasso and the adaptive lasso to perform variable selection in BLA. However, these methods are not robust for outliers. To cope with outliers, we can utilize a robust version of the lasso (e.g. the Tukey-lasso as proposed in Chapter 3) to perform the bootstrap model averaging in future work. Additionally, we only provide some asymptotic properties of the bootstrap model averaging estimators for fixed covariate dimension p in Chapter 4). In future work, we can extend these asymptotic properties in situations where the number of parameters p diverges with the sample size n.

Appendix A

Proof of Theorem 1:

To prove asymptotic normality, we let $\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta} + \frac{\mathbf{u}}{\sqrt{n}}$ and $\widehat{\sigma} = \sigma + \frac{\delta}{\sqrt{n}}$, where $\boldsymbol{\beta}$ and σ are the true location and scale parameters respectively, and define,

$$\Psi_n(\mathbf{u}, \delta) = 2\sum_{i=1}^{n} \rho_d \left(\frac{y_i - X_i^T \left(\boldsymbol{\beta} + \frac{\mathbf{u}}{\sqrt{n}} \right)}{\sigma + \frac{\delta}{\sqrt{n}}} \right) + \lambda_n \sum_{j=1}^{p} \widehat{w}_j |\beta_j + \frac{u_j}{\sqrt{n}}|.$$

Let $\widehat{\mathbf{u}}^{(n)} = \arg\min \Psi_n(\mathbf{u}, \delta)$. Then, $\widehat{\boldsymbol{\beta}}^{(n)} = \boldsymbol{\beta} + \frac{\widehat{\mathbf{u}}^{(n)}}{\sqrt{n}}$, or equivalently, $\widehat{\mathbf{u}}^{(n)} = \sqrt{n}(\widehat{\boldsymbol{\beta}}^{(n)} - \boldsymbol{\beta})$. Therefore, we prove the asymptotic normality of $\widehat{\mathbf{u}}^{(n)}$. Further let $\Psi_n(\mathbf{u}, \delta) - \Psi_n(\mathbf{0}, \delta) = V^{(n)}(\mathbf{u}, \delta)$, where

$$\begin{aligned} V^{(n)}(\mathbf{u}, \delta) =& 2\sum_{i=1}^{n} \left\{ \rho_d \left(\frac{y_i - X_i^T \left(\boldsymbol{\beta} + \frac{\mathbf{u}}{\sqrt{n}} \right)}{\sigma + \frac{\delta}{\sqrt{n}}} \right) - \rho_d \left(\frac{y_i - X_i^T \boldsymbol{\beta}}{\sigma + \frac{\delta}{\sqrt{n}}} \right) \right\} \\ &+ \frac{\lambda_n}{\sqrt{n}} \sum_{j=1}^{p} \widehat{w}_j \sqrt{n} \left(|\beta_j + \frac{u_j}{\sqrt{n}}| - |\beta_j| \right) \\ =& A^{(n)}(\mathbf{u}, \delta) + B^{(n)}(\mathbf{u}, \delta) \end{aligned}$$

We defined the first summation term by $A^{(n)}(\mathbf{u}, \delta)$ and the second summation term by $B^{(n)}(\mathbf{u}, \delta)$. Note that $r_i = y_i - X_i^T \boldsymbol{\beta}$ has mean 0 and variance σ^2. By a

Taylor expansion of degree 2, we have

$$\begin{aligned}\sum_{i=1}^{n}\rho_d\left(\frac{y_i - X_i^T\left(\boldsymbol{\beta}+\frac{\mathbf{u}}{\sqrt{n}}\right)}{\sigma+\frac{\delta}{\sqrt{n}}}\right) = \sum_{i=1}^{n}\Bigg\{&\rho_d\left(\frac{r_i}{\sigma}\right) - \frac{1}{\sigma}\psi_d\left(\frac{r_i}{\sigma}\right)\frac{X_i^T\mathbf{u}}{\sqrt{n}} \\ &- \frac{r_i}{\sigma^2}\psi_d\left(\frac{r_i}{\sigma}\right)\frac{\delta}{\sqrt{n}} + \frac{1}{2\sigma^2}\psi_d'(\frac{r_i}{\sigma})\left(\frac{X_i^T\mathbf{u}}{\sqrt{n}}\right)^2 \\ &+ \left(\frac{1}{\sigma^2}\psi_d\left(\frac{r_i}{\sigma}\right) + \frac{r_i}{\sigma^3}\psi_d'\left(\frac{r_i}{\sigma}\right)\right)\frac{X_i^T\mathbf{u}\delta}{n} \\ &+ \frac{1}{2}\left(\frac{2r_i}{\sigma^3}\psi_d\left(\frac{r_i}{\sigma}\right) + \frac{r_i^2}{\sigma^4}\psi_d'\left(\frac{r_i}{\sigma}\right)\right)\left(\frac{\delta}{\sqrt{n}}\right)^2 \\ &+ \Delta\left(\frac{X_i^T\mathbf{u}}{\sqrt{n}}, \frac{\delta}{\sqrt{n}}\right)\Bigg\}\end{aligned}$$

with $\Delta(\frac{X_i^T\mathbf{u}}{\sqrt{n}}, \frac{\delta}{\sqrt{n}})/\|\frac{X_i^T\mathbf{u}}{\sqrt{n}}, \frac{\delta}{\sqrt{n}}\|^2 \to 0$, and,

$$\begin{aligned}\sum_{i=1}^{n}\rho_d\left(\frac{y_i - X_i^T\boldsymbol{\beta}}{\sigma+\frac{\delta}{\sqrt{n}}}\right) = \sum_{i=1}^{n}\Bigg\{&\rho_d\left(\frac{r_i}{\sigma}\right) - \frac{r_i}{\sigma^2}\psi_d\left(\frac{r_i}{\sigma}\right)\frac{\delta}{\sqrt{n}} \\ &+ \frac{1}{2}\left(\frac{2r_i}{\sigma^3}\psi_d\left(\frac{r_i}{\sigma}\right) + \frac{r_i^2}{\sigma^4}\psi_d'\left(\frac{r_i}{\sigma}\right)\right)\left(\frac{\delta}{\sqrt{n}}\right)^2 + \Delta\left(\frac{\delta}{\sqrt{n}}\right)\Bigg\}\end{aligned}$$

with $\Delta(\frac{\delta}{\sqrt{n}})/(\frac{\delta}{\sqrt{n}})^2 \to 0$. Therefore, after taking the difference, we can write $A^{(n)}(\mathbf{u}, \delta)$ as,

$$\begin{aligned}A^{(n)}(\mathbf{u},\delta) = &-\frac{2}{\sigma}\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}\psi_d\left(\frac{r_i}{\sigma}\right)X_i^T\right)\mathbf{u} + \frac{1}{\sigma^2}\mathbf{u}^T\left(\frac{1}{n}\sum_{i=1}^{n}\psi_d'\left(\frac{r_i}{\sigma}\right)X_iX_i^T\right)\mathbf{u} \\ &+ \frac{2}{\sigma^2}\left(\frac{1}{n}\sum_{i=1}^{n}\psi_d\left(\frac{r_i}{\sigma}\right)X_i^T + \frac{1}{n}\sum_{i=1}^{n}\frac{r_i}{\sigma}\psi_d'\left(\frac{r_i}{\sigma}\right)X_i^T\right)\mathbf{u}\delta \\ &+ 2\sum_{i=1}^{n}\Delta\left(\frac{X_i^T\mathbf{u}}{\sqrt{n}}, \frac{\delta}{\sqrt{n}}\right) - 2\sum_{i=1}^{n}\Delta\left(\frac{\delta}{\sqrt{n}}\right).\end{aligned}$$

We now analyse the asymptotic behaviour of each term in $A^{(n)}(\mathbf{u}, \delta)$. Since $E\psi_d = 0$ and $Var(\psi_d) = E\psi_d^2$, the multidimensional central limit theorem yields,

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}\psi_d\left(\frac{r_i}{\sigma}\right)X_i^T\right) \xrightarrow{d} \mathbf{K} = \mathcal{N}(\mathbf{0}, \frac{E\psi_d^2}{n}\mathbf{X}^T\mathbf{X}) \to \mathcal{N}(\mathbf{0}, E\psi_d^2\mathbf{C}).$$

Since $Var(\psi_d')$ is finite, by assumption A2, $Var\left(\frac{1}{n}\sum_{i=1}^{n}\psi_d'\left(\frac{r_i}{\sigma}\right)X_iX_i^T\right) \to \mathbf{0}$. Thus, by law of large numbers, we have

$$\frac{1}{n}\sum_{i=1}^{n}\psi_d'\left(\frac{r_i}{\sigma}\right)X_iX_i^T \xrightarrow{p} E\psi_d'\mathbf{C}$$

Again, since $E\psi_d = 0$ and $Var(\psi_d)$ is finite, $Var\left(\frac{1}{n}\sum_{i=1}^{n}\psi_d\left(\frac{r_i}{\sigma}\right)X_i^T\right) \to \mathbf{0}$ and we have,

$$\frac{1}{n}\sum_{i=1}^{n}\psi_d\left(\frac{r_i}{\sigma}\right)X_i^T \xrightarrow{p} \mathbf{0}.$$

It is easy to show that ψ_d' is an even function so $E\left[\frac{r_i}{\sigma}\psi_d'\left(\frac{r_i}{\sigma}\right)\right] = 0$ and

$$Var\left(\frac{r_i}{\sigma}\psi_d'\left(\frac{r_i}{\sigma}\right)\right) = E\left(\frac{r_i^2}{\sigma^2}\psi_d'^2\left(\frac{r_i}{\sigma}\right)\right) \leq M,$$

where M is a finite number since ψ_d' is bounded. Therefore, we have

$$Var\left(\frac{1}{n}\sum_{i=1}^{n}\frac{r_i}{\sigma}\psi_d'\left(\frac{r_i}{\sigma}\right)X_i^T\right) \to \mathbf{0}$$

and,

$$\frac{1}{n}\sum_{i=1}^{n}\frac{r_i}{\sigma}\psi_d'\left(\frac{r_i}{\sigma}\right)X_i^T \xrightarrow{p} E\left[\frac{r_i}{\sigma}\psi_d'\left(\frac{r_i}{\sigma}\right)\right]\frac{\sum_{i=1}^{n}X_i^T}{n} = \mathbf{0}$$

In other words, the interaction term with $\mathbf{u}\delta$ goes to 0 in probability as $n \to \infty$. Now we consider the remainder terms. Recall that $\Delta(\frac{X_i^T\mathbf{u}}{\sqrt{n}}, \frac{\delta}{\sqrt{n}})/\|\frac{X_i^T\mathbf{u}}{\sqrt{n}}, \frac{\delta}{\sqrt{n}}\|^2 \to 0$.

Gathering this property with assumption A2, we have $\forall \xi > 0$, $\exists N_\xi$, $\forall n \geq N_\xi$,

$$\sum_{i=1}^{n} \left| \Delta \left(\frac{X_i^T \mathbf{u}}{\sqrt{n}}, \frac{\delta}{\sqrt{n}} \right) \right| \leq \sum_{i=1}^{n} \xi \left(\left(\frac{X_i^T \mathbf{u}}{\sqrt{n}} \right)^2 + \frac{\delta^2}{n} \right) = \xi \left(\sum_{i=1}^{n} \left(\frac{X_i^T \mathbf{u}}{\sqrt{n}} \right)^2 + \delta^2 \right)$$

since $\sum_{i=1}^{n} \left(\frac{X_i^T \mathbf{u}}{\sqrt{n}} \right)^2$ is bounded, it ensures that $\sum_{i=1}^{n} \Delta \left(\frac{X_i^T \mathbf{u}}{\sqrt{n}}, \frac{\delta}{\sqrt{n}} \right)$ tends to 0 as $n \to \infty$. Similar proof applies to $\sum_{i=1}^{n} \Delta \left(\frac{\delta}{\sqrt{n}} \right)$.
Therefore, to conclude, the asymptotic behaviour of $A^{(n)}(\mathbf{u}, \delta)$ is,

$$A^{(n)}(\mathbf{u}, \delta) \xrightarrow{d} -\frac{2}{\sigma} \mathbf{u}^T \mathbf{K} + \frac{E\psi_d'}{\sigma^2} \mathbf{u}^T \mathbf{C} \mathbf{u}$$

Now we consider the limiting behaviour of $B^{(n)}(\mathbf{u}, \delta)$; the argument is similar to that in Zou(2006). Recall that $B^{(n)}(\mathbf{u}, \delta) = \frac{\lambda_n}{\sqrt{n}} \sum_{j=1}^{p} \widehat{w}_j \sqrt{n} \left(|\beta_j + \frac{u_j}{\sqrt{n}}| - |\beta_j| \right)$. If $\beta_j \neq 0$ then, $\widehat{w}_j = 1/|\widehat{\beta}_j^{MM}| \xrightarrow{p} 1/|\beta_j|$, due to the consistency of MM-estimates as discussed in Yohai (1987). Futher, $\sqrt{n} \left(|\beta_j + \frac{u_j}{\sqrt{n}}| - |\beta_j| \right) \to u_j \text{sign}(\beta_j)$. Therefore, by Slutsky's theorem and the assumption that $\lambda_n / \sqrt{n} \to 0$, when $\beta_j \neq 0$,

$$\frac{\lambda_n}{\sqrt{n}} \widehat{w}_j \sqrt{n} \left(|\beta_j + \frac{u_j}{\sqrt{n}}| - |\beta_j| \right) \xrightarrow{p} 0$$

If $\beta_j = 0$, then $\sqrt{n} \left(|\beta_j + \frac{u_j}{\sqrt{n}}| - |\beta_j| \right) = |u_j|$ and hence,

$$\frac{\lambda_n}{\sqrt{n}} \widehat{w}_j \sqrt{n} \left(|\beta_j + \frac{u_j}{\sqrt{n}}| - |\beta_j| \right) = \lambda_n (\sqrt{n} \widehat{\beta}_j^{MM})^{-1} |u_j| \xrightarrow{p} \infty$$

since $\sqrt{n} \widehat{\beta}_j^{MM} = O_p(1)$ and $\lambda_n \to \infty$. Combining these results with the result for $A^{(n)}(\mathbf{u}, \delta)$, we have $V^{(n)}(\mathbf{u}, \delta) \xrightarrow{d} V(\mathbf{u})$ for every $\mathbf{u}$, where

$$V(\mathbf{u}) = \begin{cases} -\frac{2}{\sigma} \mathbf{u}_{\mathcal{A}}^T \mathbf{K}_{\mathcal{A}} + \frac{E\psi_d'}{\sigma^2} \mathbf{u}_{\mathcal{A}}^T \mathbf{C}_{11} \mathbf{u}_{\mathcal{A}} & \text{if } u_j = 0 \quad \forall j \notin \mathcal{A} \\ \infty & \text{otherwise.} \end{cases}$$

Further note that $V^{(n)}(\mathbf{u}, \delta)$ is convex and the unique minimum of $V(\mathbf{u})$ is

$$\left(\frac{\sigma}{E\psi_d'}\mathbf{C}_{11}^{-1}\mathbf{K}_{\mathcal{A}}, 0\right)^T.$$

Therefore, as mentioned in Zou (2006), from the epi-convergence results of Geyer (1994) and Knight and Fu (2000), we obtain,

$$\begin{aligned}\widehat{\mathbf{u}}_{\mathcal{A}}^{(n)} \xrightarrow{d} \frac{\sigma}{E\psi_d'}\mathbf{C}_{11}^{-1}\mathbf{K}_{\mathcal{A}} =& \frac{\sigma}{E\psi_d'}\mathbf{C}_{11}^{-1}\mathcal{N}(\mathbf{0}, E\psi_d^2\mathbf{C}_{11}) \\ =& \mathcal{N}(\mathbf{0}, \sigma^2\frac{E\psi_d^2}{(E\psi_d')^2}\mathbf{C}_{11}^{-1}) \quad \text{and} \quad \widehat{\mathbf{u}}_{\mathcal{A}^c}^{(n)} \xrightarrow{d} \mathbf{0}.\end{aligned}$$

Thus we have proved the asymptotic normality of the adaptive robust lasso estimates with Tukey's biweight loss. It is worth noticing that when the tuning constant in the Tukey's biweight $d = 4.685$ and the underlying distribution is normal, we have $\frac{E\psi_d^2}{(E\psi_d')^2} = 1/0.95$. In other words, this adaptive robust lasso estimates achieve 95 % asymptotic efficiency.

Finally, we prove the consistency in variable selection. For all $j \in \mathcal{A}$, we see that $\widehat{\boldsymbol{\beta}}^{(n)} \xrightarrow{p} \boldsymbol{\beta}$ from the asymptotic normality established above. Therefore, $P(j \in \mathcal{A}_n) \to 1$. If we can show that for all $j' \notin \mathcal{A}, P(j' \in \mathcal{A}_n) \to 0$, then consistency in variable selection holds. Consider the event $\forall j' \in \mathcal{A}_n$. By the KKT optimality conditions, we know that

$$\frac{1}{\widehat{\sigma}_n}\sum_{i=1}^{n}\psi_d\left(\frac{y_i - X_i^T\widehat{\boldsymbol{\beta}}^{(n)}}{\widehat{\sigma}_n}\right)x_{ij'} = \lambda_n\widehat{w}_{j'}$$

where the left hand side is the derivative of the loss function with respect to β. Dividing both sides by $\sqrt{n}$ and now note that the right hand side becomes

$$\frac{\lambda_n\widehat{w}_{j'}}{\sqrt{n}} = \lambda_n(\sqrt{n}\widehat{\beta}_{j'}^{MM}) \xrightarrow{p} \infty.$$

The left hand side is

$$\begin{aligned}\frac{1}{\widehat{\sigma}_n\sqrt{n}}\sum_{i=1}^{n}\psi_d\left(\frac{y_i-X_i^T\widehat{\boldsymbol{\beta}}^{(n)}}{\widehat{\sigma}_n}\right)x_{ij'} &= \frac{1}{\widehat{\sigma}_n\sqrt{n}}\sum_{i=1}^{n}\psi_d\left(\frac{y_i-X_i^T\left(\boldsymbol{\beta}+\frac{\widehat{\mathbf{u}}^{(n)}}{\sqrt{n}}\right)}{\sigma+\frac{\widehat{\delta}^{(n)}}{\sqrt{n}}}\right)x_{ij'} \\ &= \frac{1}{\widehat{\sigma}_n\sqrt{n}}\sum_{i=1}^{n}\left\{\psi_d\left(\frac{r_i}{\sigma}\right)-\frac{1}{\sigma}\psi_d'\left(\frac{r_i}{\sigma}\right)\frac{X_i^T\widehat{\mathbf{u}}^{(n)}}{\sqrt{n}}\right. \\ &\quad \left.-\frac{r_i}{\sigma^2}\psi_d'\left(\frac{r_i}{\sigma}\right)\frac{\widehat{\delta}^{(n)}}{\sqrt{n}}+\Delta\left(\frac{X_i^T\widehat{\mathbf{u}}^{(n)}}{\sqrt{n}},\frac{\widehat{\delta}^{(n)}}{\sqrt{n}}\right)\right\}x_{ij'},\end{aligned}$$

by a first-order Taylor expansion. Further note that

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}\psi_d\left(\frac{r_i}{\sigma}\right)x_{ij'}\right)\xrightarrow{d}\mathcal{N}(\mathbf{0},\frac{E\psi_d^2}{n}\sum_{i=1}^{n}x_{ij'}^2)$$

and $\frac{1}{n}\sum_{i=1}^{n}\psi_d'\left(\frac{r_i}{\sigma}\right)X_i^T\widehat{\mathbf{u}}^{(n)}x_{ij'}$ converges in distribution to a normal distribution with bounded variance as $\widehat{\mathbf{u}}^{(n)}=\sqrt{n}(\widehat{\boldsymbol{\beta}}^{(n)}-\boldsymbol{\beta})$. Moreover, similarly to the proof of asymptotic normality, $\frac{1}{n}\sum_{i=1}^{n}\frac{r_i}{\sigma}\psi_d'\left(\frac{r_i}{\sigma}\right)x_{ij'}\to\mathbf{0}$ and $\sum_{i=1}^{n}\Delta\left(\frac{X_i^T\widehat{\mathbf{u}}^{(n)}}{\sqrt{n}},\frac{\widehat{\delta}^{(n)}}{\sqrt{n}}\right)\to\mathbf{0}$. Therefore, we conclude that

$$P(j'\in\mathcal{A}_n)\leq P\left(\frac{1}{\widehat{\sigma}_n}\sum_{i=1}^{n}\psi_d\left(\frac{y_i-X_i^T\widehat{\boldsymbol{\beta}}^{(n)}}{\widehat{\sigma}_n}\right)x_{ij'}=\lambda_n\widehat{w}_{j'}\right)\to 0$$

and hence the adaptive robust lasso estimates with Tukey's biweight are variable selection consistent.

Bibliography

C. Agostinelli. Robust model selection in regression via weighted likelihood methodology. *Statistics & probability letters*, 56(3):289–300, 2002.

H. Akaike. A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723, 1974.

A. Alfons. robusthd: Robust methods for high-dimensional data. *R package version 0.5. 0*, 2014.

A. Alfons, C. Croux, S. Gelper, et al. Sparse least trimmed squares regression for analyzing high-dimensional large data sets. *The Annals of Applied Statistics*, 7(1):226–248, 2013.

N. Augustin, W. Sauerbrei, and M. Schumacher. The practical utility of incorporating model selection uncertainty into prognostic models for survival data. *Statistical Modelling*, 5(2):95–118, 2005.

F. R. Bach. Bolasso: model consistent lasso estimation through the bootstrap. In *Proceedings of the 25th international conference on Machine learning*, pages 33–40. ACM, 2008.

A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

A. Belloni, V. Chernozhukov, et al. Least squares after model selection in high-dimensional sparse models. *Bernoulli*, 19(2):521–547, 2013.

R. J. Bhansali and D. Y. Downham. Some properties of the order of an autoregressive model selected by a generalization of akaike s epf criterion. *Biometrika*, 64(3):547–551, 1977.

L. Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

P. J. Brown, T. Fearn, and M. Vannucci. Bayesian wavelet regression on curves with application to a spectroscopic calibration problem. *Journal of the American Statistical Association*, 96(454):398–408, 2001.

A. Buchholz, N. Holländer, and W. Sauerbrei. On properties of predictors derived with a two-step bootstrap model averaging approacha simulation study in the linear regression model. *Computational Statistics & Data Analysis*, 52(5):2778–2793, 2008.

S. T. Buckland, K. P. Burnham, and N. H. Augustin. Model selection: an integral part of inference. *Biometrics*, pages 603–618, 1997.

K. P. Burnham and D. R. Anderson. *Model selection and multimodel inference: a practical information-theoretic approach*. Springer Science & Business Media, 2003.

K. P. Burnham and D. R. Anderson. Multimodel inference understanding aic and bic in model selection. *Sociological methods & research*, 33(2):261–304, 2004.

A. Chatterjee and S. Lahiri. Asymptotic properties of the residual bootstrap for lasso estimators. *Proceedings of the American Mathematical Society*, 138(12): 4497–4509, 2010.

A. Chatterjee and S. N. Lahiri. Bootstrapping lasso estimators. *Journal of the American Statistical Association*, 106(494):608–625, 2011.

G. Claeskens, N. L. Hjort, et al. *Model selection and model averaging*, volume 330. Cambridge University Press Cambridge, 2008.

M. Clyde and E. I. George. Model uncertainty. *Statistical science*, pages 81–94, 2004.

M. Clyde et al. Model uncertainty and health effect studies for particulate matter. *Environmetrics*, 11(6):745–763, 2000.

I. Daubechies, M. Defrise, and C. De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on Pure and Applied Mathematics*, 57(11):1413–1457, 2004.

D. L. Donoho and J. M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455, 1994.

N. R. Draper and H. Smith. Fitting a straight line by least squares. *Applied Regression Analysis, Third Edition*, pages 15–46, 1981.

B. Efron. *Bootstrap methods: another look at the jackknife*. Springer, 1992.

B. Efron, T. Hastie, I. Johnstone, R. Tibshirani, et al. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004.

J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456): 1348–1360, 2001.

J. Fan and J. Lv. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911, 2008.

J. Friedman, T. Hastie, and R. Tibshirani. glmnet: Lasso and elastic-net regularized generalized linear models. *R package version*, 1, 2009.

J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1, 2010.

C. J. Geyer. On the asymptotics of constrained m-estimation. *The Annals of Statistics*, pages 1993–2010, 1994.

M. Grant, S. Boyd, and Y. Ye. Cvx: Matlab software for disciplined convex programming, 2008.

P. Hall, E. R. Lee, and B. U. Park. Bootstrap-based penalty choice for the lasso, achieving oracle performance. *Statistica Sinica*, pages 449–471, 2009.

F. R. Hampel. A general qualitative definition of robustness. *The Annals of Mathematical Statistics*, pages 1887–1896, 1971.

N. L. Hjort and G. Claeskens. Frequentist model average estimators. *Journal of the American Statistical Association*, 98(464):879–899, 2003.

J. A. Hoeting, D. Madigan, A. E. Raftery, and C. T. Volinsky. Bayesian model averaging: a tutorial. *Statistical science*, pages 382–401, 1999.

S. Horvath, B. Zhang, M. Carlson, K. Lu, S. Zhu, R. Felciano, M. Laurance, W. Zhao, S. Qi, Z. Chen, et al. Analysis of oncogenic signaling networks in glioblastoma identifies aspm as a molecular target. *Proceedings of the National Academy of Sciences*, 103(46):17402–17407, 2006.

P. J. Huber. *Robust statistics*. Springer, 2011.

P. J. Huber et al. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, 1964.

J. A. Khan, S. Van Aelst, and R. H. Zamar. Robust linear model selection based on least angle regression. *Journal of the American Statistical Association*, 102 (480):1289–1299, 2007.

K. Knight and W. Fu. Asymptotics for lasso-type estimators. *Annals of statistics*, pages 1356–1378, 2000.

S. Konishi and G. Kitagawa. Generalised information criteria in model selection. *Biometrika*, 83(4):875–890, 1996.

S. Lambert-Lacroix and L. Zwald. Robust regression through the hubers criterion and adaptive lasso penalty. *Electronic Journal of Statistics*, 5:1015–1053, 2011.

G. Li, H. Peng, and L. Zhu. Nonconcave penalized m-estimation with a diverging number of parameters. *Statistica Sinica*, 21(1):391, 2011.

J. A. Machado. Robust model selection and m-estimation. *Econometric Theory*, 9(03):478–493, 1993.

C. L. Mallows. Some comments on c p. *Technometrics*, 15(4):661–675, 1973.

R. A. Maronna. Robust ridge regression for high-dimensional data. *Technometrics*, 53(1):44–53, 2011.

R. D. Martin, R. H. Zamar, et al. Bias robust estimation of scale. *The Annals of Statistics*, 21(2):991–1017, 1993.

F. Mosteller and J. W. Tukey. Data analysis, including statistics. *Handbook of social psychology*, 2:80–203, 1968.

S. Müller and A. Welsh. Outlier robust model selection in linear regression. *Journal of the American Statistical Association*, 100(472):1297–1310, 2005.

A. B. Owen. A robust hybrid of lasso and ridge regression. *Contemporary Mathematics*, 443:59–72, 2007.

A. E. Raftery. Bayesian model selection in social research. *Sociological methodology*, 25:111–164, 1995.

A. E. Raftery, D. Madigan, and J. A. Hoeting. Bayesian model averaging for linear regression models. *Journal of the American Statistical Association*, 92 (437):179–191, 1997.

J. S. Rao and R. Tibshirani. The out-of-bootstrap method for model averaging and selection. *University of Toronto*, 1997.

S. Roberts and G. Nowak. Stabilizing the lasso against cross-validation variability. *Computational Statistics & Data Analysis*, 70:198–211, 2014.

E. Ronchetti. Robust model selection in regression. *Statistics & Probability Letters*, 3(1):21–23, 1985.

E. Ronchetti and R. G. Staudte. A robust version of mallows's cp. *Journal of the American Statistical Association*, 89(426):550–559, 1994.

E. Ronchetti, C. Field, and W. Blanchard. Robust linear model selection by cross-validation. *Journal of the American Statistical Association*, 92(439):1017–1023, 1997.

P. Rousseeuw and V. Yohai. Robust regression by means of s-estimators. In *Robust and nonlinear time series analysis*, pages 256–272. Springer, 1984.

P. J. Rousseeuw. Least median of squares regression. *Journal of the American statistical association*, 79(388):871–880, 1984.

P. J. Rousseeuw and C. Croux. Alternatives to the median absolute deviation. *Journal of the American Statistical association*, 88(424):1273–1283, 1993.

M. Salibian-Barrera and V. J. Yohai. A fast algorithm for s-regression estimates. *Journal of Computational and Graphical Statistics*, 15(2):414–427, 2006.

G. Schwarz et al. Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464, 1978.

J. Shao. Linear model selection by cross-validation. *Journal of the American statistical Association*, 88(422):486–494, 1993.

E. Smucler and V. J. Yohai. Robust and sparse estimators for linear regression models. *arXiv preprint arXiv:1508.01967*, 2015.

E. Smucler and V. J. Yohai. Robust and sparse estimators for linear regression models. *Computational Statistics & Data Analysis*, 111:116–130, 2017.

M. Stone. An asymptotic equivalence of choice of model by cross-validation and akaike's criterion. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 44–47, 1977.

K. Tharmaratnam and G. Claeskens. A comparison of robust versions of the aic based on m-, s-and mm-estimators. *Statistics*, 47(1):216–235, 2013.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

J. W. Tukey. A survey of sampling from contaminated distributions. *Contributions to probability and statistics*, 2:448–485, 1960.

H. Wang. Forward regression for ultra-high dimensional variable screening. *Journal of the American Statistical Association*, 104(488):1512–1524, 2009.

H. Wang, G. Li, and G. Jiang. Robust regression shrinkage and consistent variable selection through the lad-lasso. *Journal of Business & Economic Statistics*, 25 (3):347–355, 2007.

S. Wang, B. Nan, S. Rosset, and J. Zhu. Random lasso. *The annals of applied statistics*, 5(1):468, 2011.

V. J. Yohai. High breakdown-point and high efficiency robust estimates for regression. *The Annals of Statistics*, pages 642–656, 1987.

V. J. Yohai and R. H. Zamar. High breakdown-point estimates of regression by means of the minimization of an efficient scale. *Journal of the American statistical association*, 83(402):406–413, 1988.

Z. Yuan and Y. Yang. Combining linear regression models. *Journal of the American Statistical Association*, 2012.

P. Zhao and B. Yu. On model selection consistency of lasso. *The Journal of Machine Learning Research*, 7:2541–2563, 2006.

H. Zou. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429, 2006.

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67 (2):301–320, 2005.

H. Zou, T. Hastie, R. Tibshirani, et al. On the degrees of freedom of the lasso. *The Annals of Statistics*, 35(5):2173–2192, 2007.